Correcting Crowdsourced annotations to Improve Detection of Outcome types in Evidence Based Medicine

Micheal Abaho[1*], **Danushka Bollegala**[1], **Paula Williamson**[4,3] and **Susanna Dodd**[4]
[1]University of Liverpool, Computer Science Department
[3]MRC Hub for Trials Methodology Research Network, Department of Biostatistics
[4]University of Liverpool, Biostatistics Department
{m.abaho,danushka,prw,shinds}@liverpool.ac.uk

Abstract

The validity and authenticity of annotations in datasets massively influences the performance of Natural Language Processing (NLP) systems. In other words, poorly annotated datasets are likely to produce fatal results in at-least most NLP problems hence misinforming consumers of these models, systems or applications. This is a bottleneck in most domains, especially in healthcare where crowdsourcing is a popular strategy in obtaining annotations. In this paper, we present a framework that automatically corrects incorrectly captured annotations of outcomes, thereby improving the quality of the crowdsourced annotations. We investigate a publicly available dataset called EBM-NLP, built to power NLP tasks in support of Evidence based Medicine (EBM) primarily focusing on health outcomes.

1 Background

Evidence Based Medicine (EBM) is a popular health research paradigm that enforces healthcare decision making through the explicit and judicious use of current best evidence [Sackett *et al.*, 1996]. In practice, researchers in EBM widely use a framework entitled PICO, representing a collection of elements that form the basis of clinical questions, i.e. **P**atients, **I**nterventions, **C**omparators and **O**utcomes [Huang *et al.*, 2006]. This framework has significantly contributed towards various key health-care delivery indicators such as identification of evidence of the effectiveness of a certain treatment or diagnosis, strategies to evaluate quality of studies and mechanisms implemented in healthcare [Santos *et al.*, 2007].

EBM supported by NLP involves extraction of evidence from biomedical literature powered by several opensource tools such as BioNLP[1]MetaMap tools[2]. This extraction largely entails extraction of PICO framework elements. This paper focuses on the extraction of outcomes, emphasizing the flaws/faults discovered in crowdsourced annotations of health outcomes within medical research abstracts.

[*]Contact Author
[1]http://biocreative.sourceforge.net/bionlp_tools_links.html
[2]https://metamap.nlm.nih.gov/

1.1 Outcome detection in EBM

An outcome is a measurement or an observation used to capture and assess the effect of treatment such as assessment of side effects (risk) or effectiveness (benefits) [Williamson *et al.*, 2017]. Examples of outcomes are blood pressure, anxiety, stress, fatigue and quality of life. In this paper, we attempt to review and rectify flaws in outcome annotations utilizing NLP methods. Flaws examined here are ideally perceived as errors made while manually annotating outcomes within medical research abstracts. These may vary from, capturing non-outcomes as outcomes to capturing unnecessary text such as context as part of an outcome to compressing multiple outcomes into a single outcome and many others, as Section 3 discusses. Flaws such as these constrain the ability to build systems resilient enough to detect outcomes. From an NLP perspective, the more fragile the quality of annotations is, the less accurate the prediction models would be. Ultimately this hampers the overall objective of building systems that enhance the effective search for evidence within published literature hence impeding the aims of EBM [Nye *et al.*, 2018].

1.2 How and what did we do to achieve the goals in the study?

We investigate a recently published corpus EBM-NLP [Nye *et al.*, 2018], comprising abstracts annotated with outcome types using a highly scrutinised crowdsource labelling strategy. An outcome type is a classification or category that collectively embodies a group of outcomes measured during Randomised Control Trials (RCTs). This investigation begins with an assessment of whether the annotations retain the true identity of an outcome as defined in previous paragraph, and if not, what flaws recur across these annotations. Flaws are carefully identified with the supervision of domain experts inorder to eliminate any non-medical judgment or analysis that would bias our approach. This is followed by formulating constraints to examine the syntactic and semantic structure of the annotations to correct the identified flaws.
In summary, this paper reveals the noise (flaws) discovered in crowdsourced outcome annotations, It proposes an approach to mitigate the downsides of noisy data. It concludes with NLP tasks performed using SOTA approaches (biLSTM, CNN and SVM) [Young *et al.*, 2018] to evaluate the impact made by the adopted corrective techniques.

2 Related Work

With the increasing trend in applying NLP and machine learning to healthcare, a tremendous amount of effort from researchers in this space has been directed towards building tools to create high-quality datasets. This has birthed a host of web-based text annotation tools such as APLenty [Nghiem and Ananiadou, 2018], BRAT[3] and Prodigy[4]. Whilst such tools enhance text annotation through their rich features, they still require people to manually annotate data, which often is a costly and tedious task. Yang *et al.* (2019) empirically prove that annotation tasks can be difficult, however they discover, that despite the noise in crowdsourced annotations, reasonable models can possibly achieve similar performance on these annotations as they would when using less expert data. Vittayakorn and Hays (2011) embarked on work closely related ours, however focusing on a computer vision dataset. They assessed the quality of user-annotations by defining annotation quality functions that calculated scores representative of the ground-truth of an annotation. Our proposed approach explores the syntactic and semantic structure of annotation spans to automatically filter out errors in a pre-annotated medical dataset, hence improve the quality of the annotations.

3 Flaws discovered in crowdsourced annotations of outcomes in healthcare

Faulting during manual annotation is almost inevitable, because of the complexity, ambiguity and variation in how heathcare terms are described across different studies [Xu *et al.*, 2016] and [Dodd *et al.*, 2018]. The huge wage expectations from domain expert annotators makes it even worse [Bruno, 2018]. This section breaks down the different flaws observed in outcome annotations,

Flaw 1: Inclusion of unnecessary text that is either supportive of the actual outcome or an elaborated context of an outcome. Two kinds of unnecessary text identified and presented in Table 1 are,

1. Statistical metrics.
   Statistical terms such as *mean*, *median*, *standard deviation* are relevant in reporting results but are not considered as outcomes themselves.
2. Modification or descriptive Part-Of-Speech (POS).
   Comparative POS such as adjectives, conjunctions and adverbs were captured as part of the sequence of words in outcome-spans. E.g. *Lower* in *Lower maternal attachment* can also be *higher* which are comparative adjectives describing the change as applied to an outcome *maternal attachment*.

Flaw 2: Failure to identify independent or rather granular out- comes. This was observed across the following,

1. Multiple outcomes annotated as a single outcome.
   Some outcome-spans were captured as a sequence of distinct outcomes syntactically separated by either logical conjunctions (and/or) or punctuation characters such as commas, full and semi-colons.
2. Outcomes co-joined by a dependency term.
   These included outcome-spans that depicted two or more distinct but related outcomes. e.g. *Systolic and Diastolic blood pressure* represents two different but related outcomes and POSas Table 2 below indicates,

Incorrectly captured Outcome	Correct Outcome
1. mean arterial blood pressure 2. median Survival	arterial blood pressure Survival
1. Improved ADHD symptoms 2. Lower maternal attachment	ADHD symptoms maternal attachment

Table 1: Examples of unnecessary text such as statistical terms.

Incorrectly captured Outcome	Correct Outcome
cardiovascular events- (myocardial infarction, stroke and- cardiovascular death)	1. myocardial infarction 2. stroke 3. cardiovascular death
Systolic and Diastolic blood- pressure	1. Systolic blood pressure 2. Diastolic blood pressure

Table 2: Examples of multiple distinct outcomes compressed into one outcome.

Flaw 3: Capturing Measurement tools, metrics and results as outcomes.
scores within *Work-related stress scores* is metric result reported during RCTs, but the outcome itself is *Work-related stress*. Other examples may include tools such as questionnaires and tests used in RCTs. Examples are shown in Table 3,

Incorrectly captured Outcome	Correct Outcome
1. Quality of life Questionnaire 2. Work-related stress scores 3. Weight-test	Quality of life Work-related stress Weight

Table 3: Examples of measurement tools and scores captured as outcomes

Flaw 4: Imprecise outcome annotations resulting from inadequate domain knowledge of annotators, Examples in Table 4.

1. Non-outcomes incorrectly captured e.g. *Severity, Effect sizes, significant improvement.*
2. Misrepresented outcome types, especially in the Mortality outcome type.

Outcome-span	Incorrect Type	Correct Type
Nauseas and Vomiting	Mortality	Physical
suicidal ideations	Mortality	Mental

Table 4: Examples of outcomes labeled with incorrect types.

[3]https://brat.nlplab.org/installation.html
[4]https://prodi.gy/

Flaw 5: Combining annotations of outcomes in non-human studies together with those in human studies.
Despite the validity of outcomes in non-human species, they ought to be separately annotated. Example, *time needed to treat commercial beef cattle* is an outcome extracted from non-human medical abstracts included in outcome annotations for human medical abstracts.

4 Proposed Hybrid approach to correct outcome annotations

4.1 Part-Of-Speech Tagging

Biomedical NLP is supported by a number of POS taggers. These include, MedPost/SKR Tagger, which was trained on 5,700 manually tagged Medline sentences achieving 97.43% accuracy on a test set of 1000 sentences [Smith *et al.*, 2004]. GENIA tagger also reported accuracies higher than 97% in POS tagging sentences from various combination of Genia corpus[5], Wall Street Journal and PennBioIE datasets [Tsuruoka *et al.*, 2005]. In our approach, we use a POS tagger available in spaCY[6], a SOTA NLP industry-scale library [Honnibal and Montani, 2017] for advanced NLP. We train the tagger on the Medpost corpus, publicly available corpus containing 6,700 Medline sentences annotated with 60 POS tags [Smith *et al.*, 2004]. The trained tagger subsequently assigned POS tags to every individual word in our dataset. The model conforms to Penn Treebank POS tagging guidelines, with a few adjustments that include,

- All words that ended with '+' such as CIN2+ were assigned noun tags, 'NN'. This catered for some medical compounds/substances with a similar syntax that could have not appeared in the training set.
- Punctuation symbols such as period (.), single quotation (') and semi-colon (;) were eliminated because the EBM-NLP dataset had several of these as redundant punctuation tokens.
- Square brackets retained their syntax as a POS tag i.e. '[' and ']' were tagged as '[' and ']' respectively.

4.2 Dealing with Statistical Terms

Statistical terms within outcomes were eliminated irrespective of their position in the outcome-spans. These terms were referenced from a couple of sources including the international institute of statistics glossary[7] and another in a book for medical device clinical trials [Abdel-Aleem and Abdel-aleem, 2009].

4.3 Rule-based Chunking

The chunking algorithm (chunker) relies on a set of rules to determine where the chunk of interest (correct outcome-span) begins and ends. These rules are handcrafted linguistic constraints created to influence the capturing of sequences of words relevant to an outcome within the incorrect crowd-sourced outcome-spans. Exposed to outcome-spans from the previous step, this chunker uses underlying syntactical patterns known as regular expressions to programmatically extract one or more sub text-spans that constitute of the actual outcome-span of interest. For example, given an incorrect outcome-span such as *"lower_JJR maternal_JJ attachment_NN"*. Based on one of the predefined constraints below that suggests removal of comparative POS such as comparative adjectives tagged 'JJR', the chunker uses the positional information of word tagged with the unwanted POS i.e. *"lower_JJR"* to strip it off and retain *"maternal attachment"* as the outcome. Below is a list of chunking constraints used,

- Penalizing POS tags including **TO** (infinitive marker), **II** (Preposition), **CC** (coordinating conjunction) and **DD** (determiner).
  Words tagged with these POS tags were deemed irrelevant and therefore removed when they were located at,
  - Start or end of outcome-spans. e.g **the_DD** memory_NN loss_NN.
  - First or last within a two-worded outcome-span.e.g. **and_CC** fatigue_NN.
  - Every position in an outcome-span, i.e. all words tagged with a mixture of only these.
- Eliminating contextually comparative or quantification terms from start or ending positions of an outcome-span sequence. Comparative terms included comparative adjectives and adverbs with tags JJR and RRR such as *longer* and *better* resp, then superlative adjectives and adverbs with tags JJT and RRT such as *highest* and *most*. We additionally considered a set of terms depicting quantity and their synonyms extracted from WordNet [Miller, 1995]. These included total, average, increase and decrease.
- Removing unnecessary word sequences at the start of outcome-spans. Unwanted starting sequence included (NNS II) or (NNS DD) or (NNS TO) e.g. *predictors of* is unnecessary in *predictors_NNS of_II sex_NN risk_NN behavior_NN*, and so is *changes in changes_NNS in_II BNP_NN*
- Splitting long outcomes via 'CC'(coordinating conjunction) and ','(comma) POS tags. e.g. *Serum_NN folate_NN and_CC vitamin_NN B12_NN* is split at *and_CC*.
- Stripping off square, curved or curly brackets wrapped around outcome-spans. The content would then be subjected to processing outlined by all the above constraints.
- Outcome-spans with a sequence of words tagged as nouns were preserved. e.g. *platelet_NN thromboxane_NN formation_NN.*

5 Dataset and Experiments

5.1 POS tagging and Rule-based Chunking

Initially, experiments for POS tagging and Chunking are performed on ca.70,000 outcome-spans extracted from the EBM-NLP corpus comprising of ca.5,000 abstracts describing RCTs annotated in detail with PICO elements[8]. The

[5]https://www.nlm.nih.gov/bsd/medline.html
[6]https://spacy.io/usage/training
[7]http://isi.cbs.nl/glossary/bloken00.htm
[8]https://ebm-nlp.herokuapp.com/index

	Model	Adverse-effects [4489/1593]	Mental [8596/3875]	Mortality [1715/1176]	Pain [1649/839]	Physical [34997/18287]	Other [17996/5499]
	Baseline (SVM)	0.55/0.65	0.63/0.72	0.62/0.89	0.70/0.77	**0.67/0.85**	0.68/0.77
	CNN	0.31/0.44	0.58/0.69	0.49/0.61	0.52/0.70	**0.55/0.71**	0.57/0.59
	LSTM	0.39/0.45	0.54/0.68	0.63/0.85	0.61/0.75	**0.72/0.86**	0.42/0.64
	MNB	0.26/0.35	0.49/0.57	0.20/0.79	0.36/0.50	**0.74/0.81**	0.46/0.49
	bi-LSTM (BM)	**0.59/0.66**	**0.71/0.80**	**0.77/0.90**	**0.74/0.81**	**0.90/0.90**	**0.62/0.75**
bi-LSTM	BM - Flaw 1	0.37	0.69	0.83	0.65	0.78	0.58
	BM - Flaw 2	**0.65**	**0.70**	**0.90**	**0.76**	0.85	**0.60**
	BM - Flaw 3	0.56	0.70	0.72	0.66	**0.88**	0.59
	BM - Flaw 4	0.51	0.63	0.50	0.70	**0.88**	0.57

Table 5: Average F1-score for each class before/after (before and after correcting outcome-spans). Additional scores reported for the Best Model (BM) when subjected to data with flaws independently corrected. Enclosed in the brackets at the top is the instance count per class before/after, (Results rounded off to two decimal places).

outcome-spans are annotated with six outcome types namely **Adverse-effects**, **Mental**, **Mortality**, **Pain**, **Physical** and **Other**. Applying techniques and constraints narrated from section 4.1 to 4.3 narrows down the dataset to ca. 32,000[9].

5.2 Classification Model

Three different neural network architectures: Long Short-Term Memory (LSTM) [Sundermeyer *et al.*, 2012], Convolutional Neural Network (CNN) [Kim, 2014], Bidirectional Long Short-Term Memory (bi-LSTM) [Zhang *et al.*, 2015] as well as two bag-of-word models: Support Vector Machines (SVM) [Wang *et al.*, 2006] and Multinomial Naive Bayes (MNB) [Frank and Bouckaert, 2006] are adopted to perform classification both on the initial extract of outcome-spans ca. 70,000 and the corrected outcome-spans ca.32,000.

Given a training-set, $O = \{(X_t, y_t)\}_{t=1}^{T}$, where X_t is an instance of an outcome-span defined as sequence of words i.e. $x = (o_1, o_2, \ldots, o_N)$ where N is the length of the outcome-span sentence to be classified and each o_n is a 50-dimensional word embedding for the respective word in the outcome-span, Embeddings are obtained using pre-trained 840B 300d GloVe word vectors [Pennington *et al.*, 2014]. y_t is a one-hot vector for the corresponding label. The goal is to learn a classifier $f : X \rightarrow Y$.

In all experiments, five-fold cross validation is used for evaluation, with a batch-size of 500, trained for 100 epochs and a drop-out of 0.2 for each single fold. **Note**: The bag-of-words models take as input, a tf-idf vector [Yun-tao *et al.*, 2005] representation of the sequence of words[9].

5.3 Evaluating the experiment results

Results presented in Table 6 indicate that the accuracies increased after correcting the errors in the outcome-spans. Moreover, the increase was not only consistent across the five different models used, but even across prediction of the six classes in the dataset. Notably, the bi-LSTM outperforms all the other models, however, the bag-of-words SVM model seems to achieve the second-highest scores. This suggests that neural networks are most effective in learning representations of medical literature such as outcomes in this study.

[9]https://github.com/MichealAbaho/pico-outcome-prediction

5.4 Flaw Analysis

In-order to examine the impact the flaws individually had on the classification performance, the flaw correction process was broken down to independently cater for the different flaws one by one. The best performing model (bi-LSTM) would then be tested on input data where only annotations with flaw 1 had been corrected and the rest ignored. This was repeatedly done for flaws 2, 3 and 4 as reported in the bottom half of Table 5. Flaw 5 was not considered in this additional analysis because of the extremely few cases it was responsible for. Despite the largely analogous results, We observed that corrections targeted to fix Flaw 2 alone, had a significantly higher impact on the performance, scoring higher F1-scores for the six classes with the exception of the Physical class. This implied that, granularity and distinctness is vitally important when detecting not just outcomes but any relevant clinical entities in biomedical literature. Nonetheless, neither of the F1-scores in this analysis would match up to the originally obtained F1-scores with all flaws corrected (line 5 - Table 5).

5.5 Conclusion

Manually annotating medical data is a challenging and costly process. As a result, crowdsourced annotations are often noisy and inconsistent. This work performs a sanity check on crowdsourced annotations in a public corpus EBM-NLP revealing various flaws in the annotations. We train a spaCY POS tagging model on Medline articles and use a rule based chunking algorithm to fix these errors/flaws. Classification experiments at the end justify the positive impact our corrective approach has on the dataset.

As part of future work, we aim to explore dependency graphs to capture disjoint or entities to achieve required granularity in outcome reporting. For instance, the outcome, *chest and abdominal pain* is best detected as two independent outcomes,*chest pain* and *abdominal pain* where pain is simply a disjoint entity. We shall further on adopt expertly annotated data to maximize precision, recall and quality of ground-truth annotations and thereby, utilize transfer learning to automatically detect outcomes. Upon satisfactorily achieving quality annotations, we shall utilize semi-supervised learning to build a corpus of outcomes ready to support NLP tasks in EBM.

References

[Abdel-Aleem and Abdel-aleem, 2009] Salah Abdel-Aleem and Salah Abdel-aleem. *Design, execution, and management of medical device clinical trials*. Wiley Online Library, 2009.

[Bruno, 2018] Godefroy Bruno. On the viability of crowdsourcing nlp annotations in healthcare, July 2018.

[Dodd *et al.*, 2018] Susanna Dodd, Mike Clarke, Lorne Becker, Chris Mavergames, Rebecca Fish, and Paula R. Williamson. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *Journal of Clinical Epidemiology*, 96:84–92, 4 2018.

[Frank and Bouckaert, 2006] Eibe Frank and Remco R Bouckaert. Naive bayes for text classification with unbalanced classes. In *European Conference on PKDD*, pages 503–510. Springer, 2006.

[Honnibal and Montani, 2017] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[Huang *et al.*, 2006] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of PICO as a knowledge representation for clinical questions. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 359–63, 2006.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Nghiem and Ananiadou, 2018] Minh-Quoc Nghiem and Sophia Ananiadou. Aplenty: annotation tool for creating high-quality datasets using active and proactive learning. In *Proc. of the EMNLP'14: System Demonstrations*, pages 108–113, 2018.

[Nye *et al.*, 2018] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J. Marshall, Ani Nenkova, and Byron C. Wallace. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. 2018.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of the EMNLP'14*, pages 1532–1543, 2014.

[Sackett *et al.*, 1996] David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

[Santos *et al.*, 2007] Cristina Mamédio da Costa Santos, Cibele Andrucioli de Mattos Pimenta, and Moacyr Roberto Cuce Nobre. The pico strategy for the research question construction and evidence search. *Revista latino-americana de enfermagem*, 15(3):508–511, 2007.

[Smith *et al.*, 2004] L Smith, Thomas Rindflesch, and W John Wilbur. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004.

[Sundermeyer *et al.*, 2012] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[Tsuruoka *et al.*, 2005] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*, pages 382–392. Springer, 2005.

[Vittayakorn and Hays, 2011] Sirion Vittayakorn and James Hays. Quality assessment for crowdsourced object annotations. In *BMVC*, pages 1–11, 2011.

[Wang *et al.*, 2006] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, and Xin Li. An optimal svm-based text classification algorithm. In *2006 International Conference on Machine Learning and Cybernetics*, pages 1378–1381. IEEE, 2006.

[Williamson *et al.*, 2017] Paula R Williamson, Douglas G Altman, Heather Bagley, Karen L Barnes, Jane M Blazeby, Sara T Brookes, Mike Clarke, Elizabeth Gargon, Sarah Gorst, Nicola Harman, et al. The comet handbook: version 1.0. *Trials*, 18(3):280, 2017.

[Xu *et al.*, 2016] Boyi Xu, Ke Xu, LiuLiu Fu, Ling Li, Weiwei Xin, and Hongming Cai. Healthcare data analytics: using a metadata annotation approach for integrating electronic hospital records. *Journal of Management Analytics*, 3(2):136–151, 2016.

[Yang *et al.*, 2019] Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C Wallace, and Ani Nenkova. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. *arXiv preprint arXiv:1905.07791*, 2019.

[Young *et al.*, 2018] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligence magazine*, 13(3):55–75, 2018.

[Yun-tao *et al.*, 2005] Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, 6(1):49–55, 2005.

[Zhang *et al.*, 2015] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proc of the 29th PACLIC'2015*, pages 73–78, 2015.