Detecting health outcomes from medical text records

Micheal Abaho, Danushka Bolleagala, Paula R Williamson & Susanna Dodd

University of Liverpool

Introduction

An Outcome is a measurement or an observation used to capture and assess the effect of a treatment [1]. Automating Outcome Detection (OD) could **speed up** access to evidence necessary in health care decision making. Given a sentence, "There was no significance between group difference in the incidence of **wheezing** or **shortness of breath**", OD extracts outcomes such as those underlined and in bold font.

OD has however previously been hindered by an abscence of a consensus on how outcomes should be reported and classified. Moreover, datasets like EBM-NLP [2] supporting OD have been found erroneous [3] with flaws like,

- Flaw 1: Inclusion of unnecessary text.
  - statistical metrics e.g. "~~mean~~ arterial pressure".
  - Clinical measurement tools e.g. "Quality of life ~~Questionnaire~~".
- Flaw 2: Failure to identify independent and granular outcomes.
  e.g. "cardiac arrest ~~and~~ heart failure"
- Flaw 3: Imprecise outcome annotations.
  e.g. "Suicidal Ideations" annotated as a ~~Mortality~~ outcome rather than Mental outcome.

Correcting flawed outcome annotations

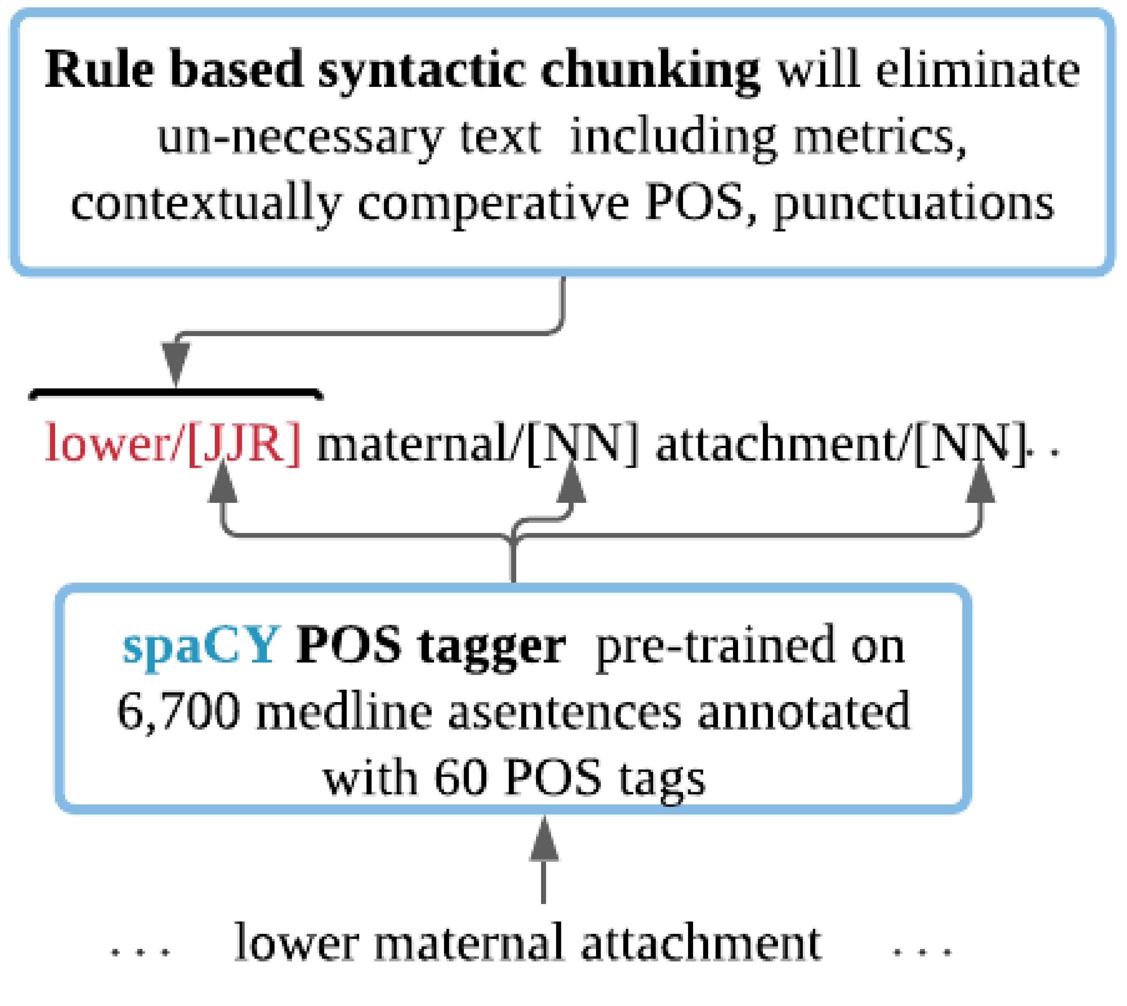

Figure 1: Part-Of-Speech (POS) Tagging and Rule-based Chunking to build EBM-NLP$_{rev}$

Model	EBM-NLP	EBM-NLP$_{rev}$
biLSTM	72.2	**80.3**
biLSTM - Flaw 2	_	**74.3**

Table 1: F1 (%) for OD on original and revised version of EBM-NLP, including when only Flaw 2 is corrected.

Joint Outcome Span Detection (OSD) & Outcome classification (OC)

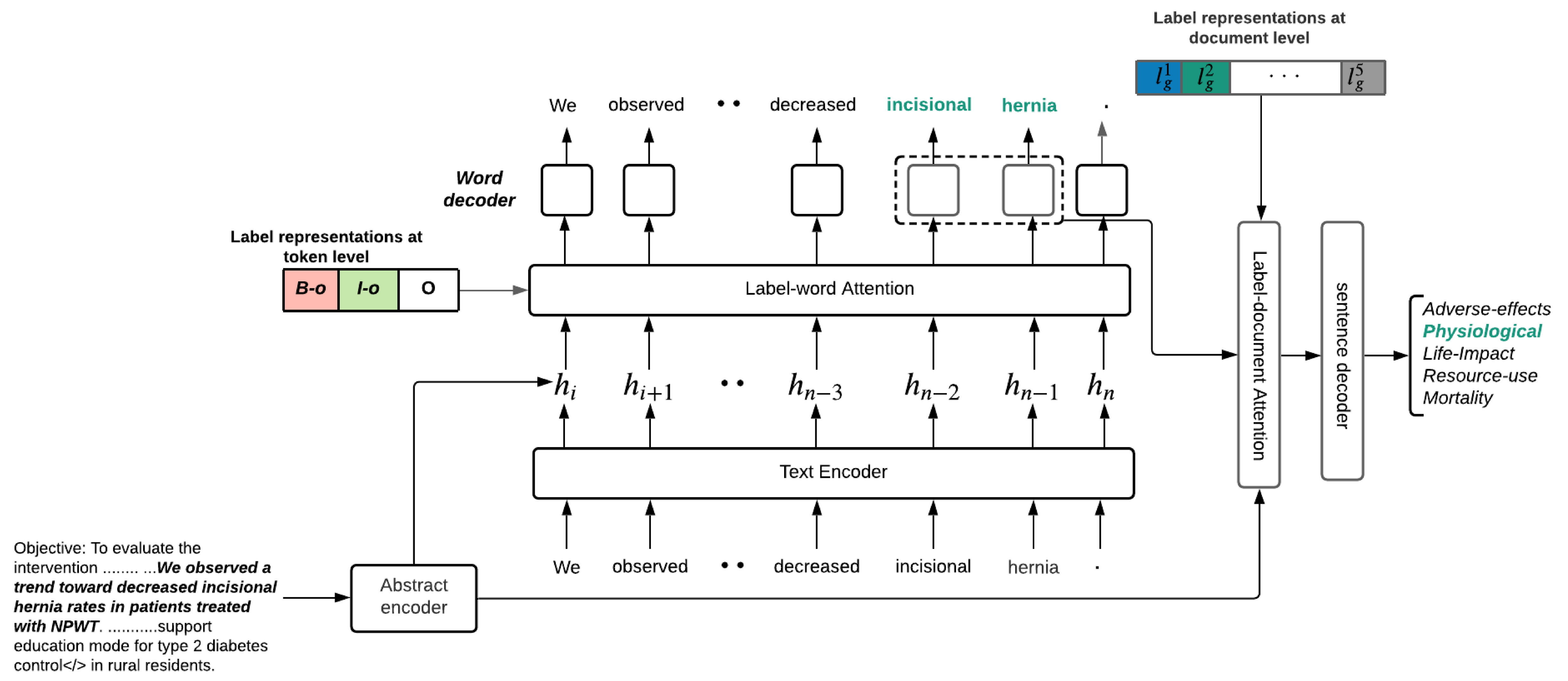

Figure 2: Label-word context aware attention framework (LCAM) for joint OSD and OC [9]

EBM-COMET (Annotation & Evaluation)

Annotation category	Annotated text
Simple	...Tai Chi may alleviate **<P 0, 28>**depression</>of the elderly through modulating autonomous nervous system or **<P 0>**heart rate variability</>...
	depression - [0:*Physiological*, 28:*Emotional Functioning*] **heart rate variability** - [*Physiological*]
Complex	...The objective of this study was to evaluate **<P 0>(S2)**right heart size **<P 0>**and function</>assessed by echocardiography during long term treatment
	right heart size - [0:*Physiological*] **right heart function** - [0:*Physiological*]

Data set statistics: 300 RCT PubMed Abstracts, 5193 sentences, an average of 0-4 outcome phrases/sentence.

Full Outcome phrase detection

Ground truth:- Systolic blood pressure
Predicted:- Systolic blood pressure

	P	R	F
Traditional NER evaluation:-	100	66.7	80.2
Full outcome phrase evaluation:-	0	0	0

Model	EBM-NLP$_{rev}$	EBM-COMET
BioBERT [4]	**53.1**	**81.3**
BioELMo [5]	52.0	75.0
SciBERT [6]	52.8	77.6
ClinicalBERT [7]	51.0	68.5
BioFLAIR [8]	51.4	68.5

Table 2: F1 (%) for OD using in-domain CLMs

OSD & OC

Given a sentence $s = \{w_i\}_{i=1}^{M}$, **OSD identifies an outcome span** $o_d = \{b_i\}_{i\geq 1}^{N}$, **and OC predicts an outcome type** $t(o_d) \in \mathcal{Y}$ **for** o_d, where, $N \leq M$

$$\boldsymbol{h}_n^c = \text{BioBERT}(w_n) + \frac{1}{|a|}\sum_{n=1}^{|a|}(\text{BioBERT}(w_n)) \quad (1)$$

Label-word attention representation (OSD)

$$\boldsymbol{A}_n^{(1)} = \text{softmax}(\mathbf{W}\tanh(\mathbf{V}\boldsymbol{h}_n)) \ \&\ \boldsymbol{A}_n^{(2)} = \mathbf{U}\boldsymbol{h}_n \quad (2)$$

$$\boldsymbol{E}^{t_l} = \mathbf{A}^{(1)}\boldsymbol{h}_n{}^{\top} + A^{(2)}\boldsymbol{h}_n^{\top} \quad (3)$$

$$L_{osd} = -\sum_{n=1}^{N}\sum_{i=1}^{|l_w|} y_{n,i}\log(\hat{y}_{n,i}). \quad (4)$$

Label-word attention representation (OC)

$$L_{oc} = -\sum_{i=1}^{|L_s|} y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \quad (5)$$

$$\text{Combined loss} L = L_{osd} + L_{oc} \quad (6)$$

Model	OSD	OC
LCAM	68.0	83.0
LCAM - Abstract	65.0	78.0
LCAM - Attention	58.0	71.0

Table 3: OSD and OC performance F1 (%) on EBM-COMET

Label-alignment for data augmentation

Algorithm 1 Label Alignment

Input: comparable datasets $\mathcal{S}$ & $\mathcal{T}$
for each label l in $\mathcal{S}$:
  Create an **embedding** $\boldsymbol{l}_s$ by $\boldsymbol{l}_s = \frac{1}{|l_s|}\sum_i^{|l_s|}\boldsymbol{O}_{l_s}$
  where $\boldsymbol{O}_{l_s} = \frac{1}{d}\sum_i^{i+(d-1)}$ **BioBERT**(w_i)
  and i & $i+(d-1)$ are the first and last words
  of an outcome span labelled l_s i.e. $\boldsymbol{O}_{\boldsymbol{l}_s}$
 for each label l in $\mathcal{T}$
   Compute cosine_similarity (cos) of $\boldsymbol{l}_s$ & $\boldsymbol{l}_t$
 Reannotate l_t outcomes with most similar l_s.

References

[1] Williamson et al. Comet handbook: version 1.0. 2017.
[2] Nye et al. EBM-NLP corpus. ACL, 2018.
[3] Abaho et al. Correcting crowdsourced outcome annotations. In CEUR Workshop Proceedings, 2019.
[4] Lee et al. Biobert. Bioinformatics, 36(4):1234–1240, 2020.
[5] Jin et al. Probing biomedical embeddings. NAACL, 2019.
[6] Beltagy et al. Scibert. EMNLP, 2019.
[7] Alsentzer et al. Clinicalbert. NAACL, June 2019.
[8] Sharma and Daniel. Bioflair. arXiv:1908.05760, 2019.
[9] Abaho et al. Joint span detection and classification for health outcomes. arXiv preprint arXiv:2104.07789, 2021.
[10] Abaho et al. Assessment of contextualised representations in outcome detection. Manuscript under review, 2021.