Contrastive Clustering of medical reports to characterise Adverse Drug Reactions

Xuchun Zhang, Michel Riveill, Milou-Daniel Drici

Contrastive Clustering of medical reports to characterise Adverse Drug Reactions

Xuchun ZHANG[1], Michel RIVEILL,[1] Milou-Daniel. DRICI,[2]

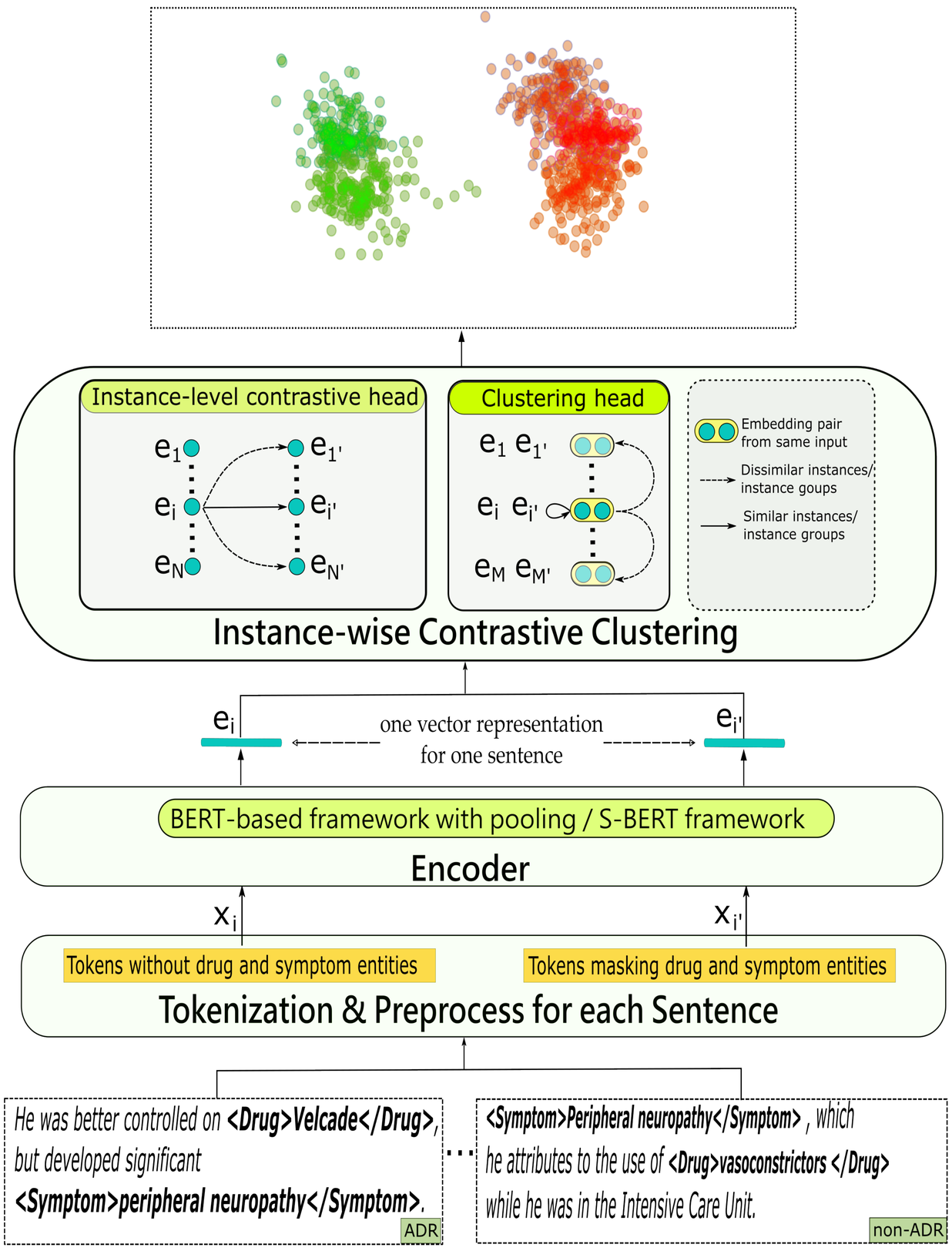
1: Université Côte d'Azur, CNRS, INRIA, I3S Laboraptory, France {xuchun.zhang, michel.riveill} @inria.fr
2: Université Côte d'Azur, CNRS, CHU Nice, France drici.md@chu-nice.fr

1. Motivation

- Adverse Drug Reactions (ADRs) being mandatory for healthcare professionals, pharmacovigilance still suffers from a significant under-reporting, accounting for only 5-10% of all ADRs.
- Identification of ADRs relies on the well-trained health professionals, while there are still an enormous amount of documents waiting to be reviewed.
- Annotating such electronic health records (EHRs) is very expensive, yet supervised approaches need a large amount of annotated data.

2. Structure of Model Pipeline

- Assumption : "the ADR relation lies in the contexts of the drug-symptom target entities". More specifically, we assume there exists a short text around the entities pair which is sufficient for the representation of ADR.
- Our goal is to separate the ADR-related text blocks (noted as positive blocks β^+) from non-related ones (noted as negative blocks β^-).



- **BioBERT** model is pre-trained and fine-tuned on biomedical corpora instead of employing general domain text corpora in BERT.
  Both BERT and BioBERT are token-wise language model.
- **Sentence-BERT** applys a siamese fine-tuning structure on basic BERT model to capture features in semantic space, which transform a whole sentence direclty into vector representation.

4. Contrastive Learning Enhanced Clustering (CLEC)

- For the Instance-level Contrastive head, the the model is trained to distinguish between similar and dissimilar data by evaluating the loss[1]
  $\ell_i^{IC} = -log\frac{exp\left(sim(x_i,x_{i'})/\tau\right)}{\sum_{j=1}^{M}\mathbb{1}_{j\neq i'}exp(sim(x_i,x_j)/\tau)}$ for each instance,

  The Instence Contrastive Loss is $\mathcal{L}_{IC} = \sum_{i=1}^{M}\ell_i^{IC}/M$
- For the clustering head, we use the Student's t-distribution[2] to measure the similarity between mapped point z_i and centroid μ_k:
  $q_{ik} = \frac{\left(1+\|z_i-\mu_k\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'}\left(1+\|z_i-\mu_{k'}\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}$, and refine the clusters with the help of an auxiliary target distribution[3] $p_{ik} = \frac{q_{ik}^2/f_k}{\sum_{k'}q_{ik'}^2/f_{k'}}$ with $f_k = \sum_i q_{ik}$ as soft cluster frequencies,

  The objective is a KL divergence loss between the soft assignments q_i and the auxiliary distribution for both instance pairs:
  $\mathcal{L}_C = \sum_{(i,i')}\ell_i^C/M + \ell_{i'}^C/M$, where $\ell_i^C = \sum_i\sum_j p_{ik}\log\frac{p_{ik}}{q_{ik}}$
- Total objective of CLEC: $\mathcal{L}_{CLEC} = \mathcal{L}_{IC} + \eta\mathcal{L}_C$.

5. Results and Discussion

Experimental Data

- Data source from MADE challenge 2018, whose corpora are only electronic health records.
- The information about medications and symptoms are given.
- **MADE 1d1s** where we extract those who has exactly one drug and one symptom as the "perfect situation", a nearly balanced dataset with short textual blocks.
- **MADE multi-d-s** has longer blocks of the EHR corpus with an almost balanced distribution between negative and positive blocks. Each block contains at least one drug and one symptom.
- In BioBERT based model, we use average pooling for short text input in **MADE 1d1s** and took *[CLS]* representation for long inputs in **MADE multi-d-s**

We have chosen a fully supervised approach (Bag of Words + Logistic Regression Classifier) as the upper bound baseline, and a BoW with random guessing as the lower bound.
The results for unsupervised approaches (*) are always followed by a KMedoids clustering

Category	Exp	MADE 1d1s		
		Prec	Recall	F1
Supervised	BOW+LR	0.702	0.797	0.746
Unsupervised	BioBERT*[4]	0.651	0.663	0.650
	SBERT*[4]	0.733	0.615	0.665
Unsupervised	BioBERT+CLEL	**0.768**	**0.669**	**0.715**
	SBERT+CLEL	0.646	0.711	0.677
Supervised	BOW+Random	0.514	0.529	0.522

Category	Exp	MADE multi-d-s		
		Prec	Recall	F1
Supervised	BOW+LR	0.809	0.847	0.827
Unsupervised	BioBERT*[4]	0.604	0.634	0.619
	SBERT*[4]	0.653	0.673	0.663
Unsupervised	BioBERT+CLEL	0.582	**0.827**	0.673
	SBERT+CLEL	0.660	0.773	**0.702**
Supervised	BOW+Random	0.509	0.440	0.472

- Both BioBERT-based model and SBert-based Model benefit the contrastive clustering comparing to our previous work.
- For short EHR data, BioBERT based model benefits more from its biomedical domain specified dictionary comparing to

the SBERT-based model,

- SBert model has more adventage for long textual EHR data,

References

1. He, Kaiming *et al.* "Momentum Contrast for Unsupervised Visual Representation Learning." 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 9726-9735.
2. Maaten, Laurens van der and Geoffrey E. Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research 9 (2008): 2579-2605.
3. Xie, Junyuan *et al.* "Unsupervised Deep Embedding for Clustering Analysis." ArXiv abs/1511.06335 (2015): n. pag.
4. X. Zhang, M.-D. Drici, M. Riveill. "Unsupervised Text Clusterisation to characterize Adverse Drug Reactions from hospitalization reports". ASPAI 2022