Unsupervised Adverse Drug Event related document detection with Bert-based model

Xuchun Zhang, Michel Riveill

Unsupervised Adverse Drug Event related document detection with Bert-based model

Xuchun ZHANG, Michel RIVEILL

Université Côte d'Azur, CNRS, INRIA, I3S, France

1. Motivation

- The post-marketing pharmacovigilance (PhV) practice aims at detecting, monitoring, characterizing and preventing adverse drug events (ADEs) in the medical reports.
- Identification of ADEs 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.

2. Problems

We defined firstly "block" (noted as b), as the basic unit of textual content to analyse, which can be either a whole document, a paragraph, a sentence or some pieces of sentences, etc. An ADE involves a drug d, a symptom s in a kind of textual block b with the description containing the two entities, which can be represented as:

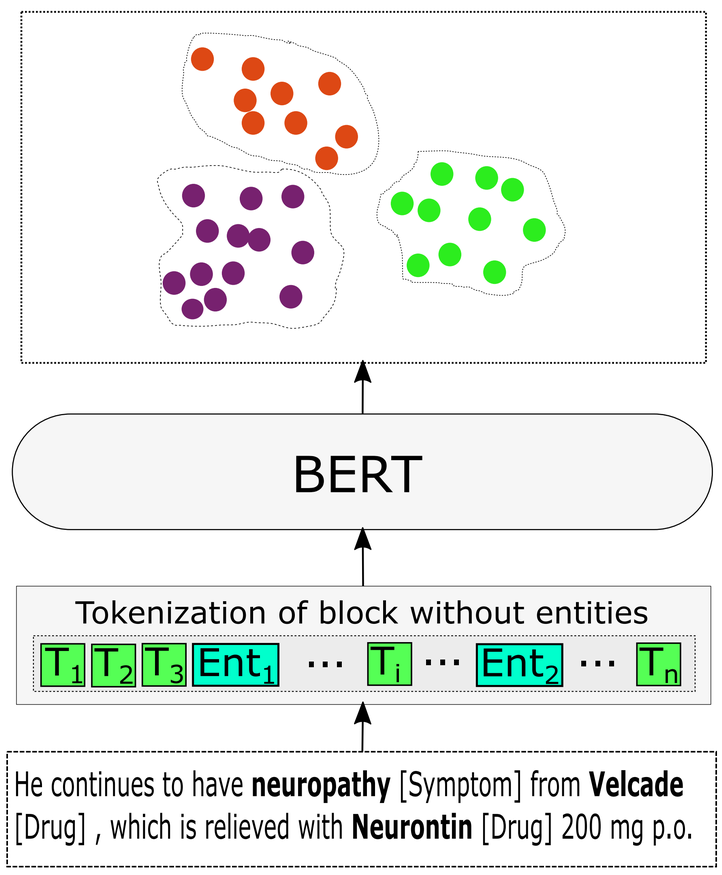
$$(relation)\ r(d, s, b) = \mathbf{ADE}(d, s, b)$$

Our basic assumption is that "the ADE relation lies in the contexts of the target entities". More specifically, we assume there exists a short text z in b around the entity pair (d,s) which is sufficient for the representation of ADE.

$$\mathbf{ADE}(d, s, b) \simeq \mathbf{ADE}(d, s, z)$$

3. Unsupervised BERT based ADE dectection

We chose sentence as our "block" b together with its entity annotations as basic element to analyse and used Bidirectional Encoder Representation from Transformers (BERT) as well as its two modifications: Sentence-BERT and BioBERT, to encode each given textual context z into vectors of the same size. Finally A clustering algorithm is used for creating clusters of similar sentences.



Modified BERT model

- **Sentence-BERT**, applying a siamese fine-tuning structure on basic BERT model to capture features in semantic space.
- **BioBERT**, pre-trained and fine-tuned in a traditional way but on the medical corpora.

4. The data

We used data in MADE challenge, whose corpora are only electronic health records with annotations of their entities. The granularity of corpus was set to sentence.

- **made_sent_all** All examples that contains only drugs or only symptoms were removed and thus we got a dataset where each block has at least one drug and one symptoms.
- **made_sent_1d1s** For the dataset above, we extract then those who has exactly one drug and one symptom, which called "1d1s" as "perfect situation".

5. What we found firstly

We have chosen a fully supervised approach (Bag of Words + Logistic Regression Classifier) as the baseline, which provided us a vision about the upper bound of our method. While the Bag of words with dummy classifier clarifies the lower bound.

made_sent_1d1s

Category	Exp	Precision	Recall	F1 Score	F2 Score
Supervised	BOW+LR	0.7862	0.7064	0.7438	0.7685
Supervised	BOW+Dummy	0.5294	0.5143	0.5217	0.5263
Unsupervised	BERT+Clustering	0.5200	0.6393	0.5735	0.5402
Unsupervised	BioBERT+Clustering	0.7467	0.6829	0.7134	0.7330
Unsupervised	S-BERT+Clustering	0.5200	0.7358	0.6094	0.5524

made_sent_all

Category	Exp	Precision	Recall	F1 Score	F2 Score
Supervised	BOW+LR	0.8372	0.8105	0.8234	0.8316
Supervised	BOW+Dummy	0.4946	0.5236	0.5087	0.5002
Unsupervised	BERT+Clustering	0.4538	0.5315	0.4896	0.4675
Unsupervised	BioBERT+Clustering	0.6923	0.7087	0.7004	0.6955
Unsupervised	S-BERT+Clustering	0.6154	0.6349	0.6250	0.6192

- Comparing to the basic BERT-based method, the Sentence-Bert-based method has an advantage in semantic meaning representation.
- Comparing to the other two BERT-based method, the BioBERT-based one shows a relatively good performance in both f1 and f2 score, especially for the "made_sent_1d1s" data, which is quite close to the supervised results. This may due to the domain specific vocabulary dictionary from the BioBERT model.

6. Perspectives

- Sentence-BioBERT may be more useful in this scenario
- cluster topic analysis
- improve for inner-block relations
- deal with inter-block problem

References

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." ArXiv:1810.04805 [Cs], May 24, 2019. http://arxiv.org/abs/1810.04805.
[2] Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." Bioinformatics, September 10, 2019, btz682. https://doi.org/10.1093/bioinformatics/btz682.
[3] Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." ArXiv:1908.10084 [Cs], August 27, 2019. http://arxiv.org/abs/1908.10084.
[4] Eugene Agichtein and Luis Gravano. "Snowball: extracting relations fromlarge plain-text collections". In:Proceedings of the fifth ACM conference onDigital libraries - DL '00. the fifth ACM conference. San Antonio, Texas,United States: ACM Press, 2000, pp. 85–94.ISBN: 978-1-58113-231-1.DOI:10.1145/336597.336644.URL:http://portal.acm.org/citation.cfm?doid=336597.336644.
[5] Jagannatha, Abhyuday, Feifan Liu, Weisong Liu, and Hong Yu. "Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes (MADE 1.0)." Drug Safety 42, no. 1 (2019): 99–111. https://doi.org/10.1007/s40264-018-0762-z.