CS410 – Technology Review

Sembian2@illinois.edu

Introduction

One of the standard analysis tasks for pharmaceuticals is the medical education for healthcare professionals to understand better the differences on various clinical aspects defined by FDA involves reviewing and comparing multiple drug labels of the same class in multiple topics including analysis of clinical study results, efficacy, adverse events, dosing to label positive, neutral and negative labeling on various topics. The task involves analyzing the text-based drug label documents to identify Adverse Events using Named Entity Recognition. In this review, Regulatory guidelines and Deep learning approaches and standards in extracting Adverse Events from clinical text labels using TensorFlow NLP

The review attempts to evaluate TensorFlow API libraries and deep learning architectures that could be implemented and using statistical models to do text mining, topic modeling and classification of drug labels and provide possible methodologies in an approach to design a system that could take FDA drug label extracting adverse events between two product drug labels. **Food and Drug Administration (FDA)in the United states has established guidelines and standards on Labelling and Annotation standards**, this standard guidelines help in defining Adverse Drug Reactions, the accuracy of the detection is very critical and requires careful inference of the reactions as it may be used for medical decisions. The Adverse Reactions is of three broad categories Adverse Event, Adverse Reaction and Severe Adverse Reaction

Adverse Event: refers to medical event associated with use of drug in humans, and consideration of drug-related

Adverse reaction: an undesirable effect reasonably associated with the use of a drug indicating a casual relationship between drug and occurrence of adverse event.

Serious Adverse reaction: refers to a reaction occurring at any dose of drug leading to death, life-threatening adverse experience, hospitalization significant disability or birth defects.

The identification of adverse drug reactions relies on annotated samples of drug labels on the above specified AdverseReaction(AR) mentions.

As per the FDA guidelines the Adverse Reaction mentions and relations including negation and negated relations. An example of an annotated sample is given below

EXAMPLE PARTIAL DOCUMENT: VIBATIV (BOXED WARNING)

- Patients with pre-existing moderate/severe renal impairment (CrCl ≤50 mL/min) who were treated with VIBATIV for hospital-acquired bacterial pneumonia/ventilator-associated bacterial pneumonia had increased **mortality** observed versus vancomycin. Use of VIBATIV in patients with pre-existing moderate/severe renal impairment (CrCl ≤50 mL/min) should be considered only when the anticipated benefit to the patient outweighs the potential risk. (5.1)
- Nephrotoxicity: **New onset** or **worsening renal impairment** has occurred. Monitor renal function in all patients. (5.3)
- Women of childbearing potential should have a serum pregnancy test prior to administration of VIBATIV. (5.4, 8.1)
- Avoid use of VIBATIV during pregnancy unless potential benefit to the patient outweighs potential risk to the fetus. (8.1)
- **Adverse developmental outcomes** observed in 3 **animal** species at clinically relevant doses raise concerns about **potential adverse developmental outcomes** in humans. (8.1)

Note the following annotations (or lack thereof):

- *renal impairment*: this text in the first sentence would not be annotated because it is a pre-existing condition.
- *CrCl ≤50 mL/min*: would not be annotated because it is a pre-existing condition.
- *hospital-acquired bacterial pneumonia/ventilator-associated bacterial pneumonia*: this would not be annotated because these are indications.
- *mortality:* this would be annotated because this is an **AR**.
- *nephrotoxicity:* this would not be annotated because it is a header.
- *new onset renal impairment:* This disconnected span is an **AR**.
- *worsening renal impairment*: This is an **AR**.
- *monitor renal function*: this would not be annotated because it is a monitoring recommendation.
- *pregnancy test*: this would not be annotated because it is a therapeutic recommendation/diagnostic work-up.
- *pregnancy*: this would not be annotated because it is a contraindication/therapeutic recommendation (avoid use during pregnancy).
- *adverse developmental outcomes* (first instance): this would be annotated because it is an **AR** and would also be connected to *animal* (type **Animal**) using a **Hypothetical** relation.
- *adverse developmental outcomes* (second instance): this would be annotated because it is an AR and would also be connected to *potential* (type **Factor**) using a **Hypothetical** relation.

The above figure is an example of some of atypical choices made in terms of what and how certain annotations are created.

The final goal is to identify ADR in a given drug label following considerations are needed:

1. The ADR identification may not be precise offsets or relations the ADR be linked to MeDRA knowledge source based on the MeDRA Hierarchy[16]
2. ADR mentioned multiple times should not carry more weight
3. Enabling further annotations using unique ADRs aggregated at the document level and further annotations using MeDRA Low Level Terms(LLT) and the corresponding Preferred Term(PT)

Adverse Reaction Extraction - Key Tasks and Natural Language Processing Methods

Steps	Task	NLP Methods
Step 1	Extract AdverseReactions and related mentions like (Severity, Factor, DrugClass, Negation, Animal)	Named Entity Recognition
Step 2	Identifying relations between AdverseReactions and related Mentions (Negated, Hypothetical and Effect)	Relation Identification
Step 3	Identify the sentiment(positive) AdverseReactions mention names in the label 1. Uncased strings of all ADR that have not been negated 2. Not related by a hypothetical relationto a DrugClass or Animal	Sentiment Analysis

Step 4	Provide the MeDRA Preferred Term(PT) and Low Level Terms(LLT)	Normalization of terms from Step 3

Drug Label Reference Standards

The MeDRA (Medical Dictionary for Regulatory Activities) standards are followed for identification of Medical terms and is used to normalize the inference or output of the entities to its proper Medical terminology.

Overview of TensorFlow

TensorFlow originally developed by researchers at Google is an open sourced machine learning platform, the library helps in development of machine learning models for various platforms and has flexible ecosystem of libraries, tools and resources allowing state-of-the-art and the best documented open source available in Machine learning developers build and deployment of Machine learning powered applications.

TensorFlow Natural Language Processing (NLP) Capabilities

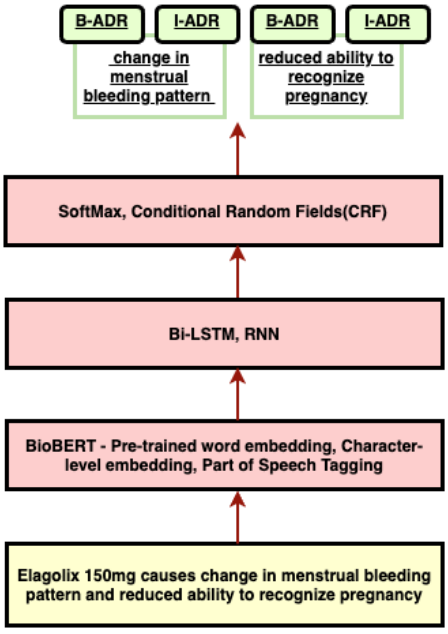
Tensorflow provides numerous modular architecture in terms of front-end architecture and supports Convolutional and recurrent networks, deep auto encoders, Named Entity Relation(NER) and LSTM (Long Short-Term Models and Natural Language processing capabilities which can be applies for topic modelling and classification use cases.

Proposed Approach for Named Entity Recognition in Identifying Adverse Drug Reactions using Transfer Learning and BioBERT[9] embeddings

While there are multiple approaches to the task of extracting ADR from drug labels the key task would include extraction of the Adverse event using Named Entity Recognition and Normalization to standard terms. This paper aims to propose an approach using TensorFlow deep learning model for NER extraction using Bi-directional Long Short-Term Memory which uses prior knowledge about the rest of the sentence combined with BioBERT character embeddings to incorporate semantic meanings to improve performance as BioBERT is generated from a corpus that is more related to Bio Terms.

TensorFlow NLP for NER Tasks

The Drug label sentences are first processed using BIO sequence tagging indicating the Beginning Inside and Outside (B_ADR / I-ADR / O-ADR) with every token having an offset indicating the character count before the first character of the given token. Text processing pipeline methods can be implemented to streamline the loading of Drug Label document and extracting sentences and tokenizing.



Deep Learning Techniques for NER

Named Entity Recognition is am information extraction technique for identification and classification of named entities in text, based on the annotated drug labels the entities can be pre-defined and generic locations like Beginning Inside and Outside of a named entity.

The success of NER system relies heavily on its input representation. Integrating or fine-tuning pre-trained language models like the BioBERT embeddings is new paradigm for Neural Named Entity Recognition techniques. Leveraging the language model provides significant improvements in F-score on analysing drug labels which are more structured, the challenge remains in other unstructured data like Social media, Patient Health Records which may have noisy data.

Though there are traditional approaches which are mostly rule based using NLTK, Machine learning approaches have improved accuracy the two approaches that can be explored are:

Approach A – Multi-Class Classification: The ADR identification as a multi-class classification where the named entities are labels and apply classification algorithms, this approach would required a detailed understanding od context of the sentence and sequence of word labels in it.

Approach B Conditional Random Field (CRF) Model: Using Conditional Random Field model using Tensorflow and Keras, a probabilistic model to model sequential data such as label of words in a sentence. This approach is also has its downsides of identifying current and previous labels but falls short of forward labels as it is a important in a drug label to identify both previous current and forward contexts

Deep learning approaches have evolved to be a better approach for NER, In tensorflow we could use the accuracy while training using different iterations(epochs), due the critical nature of medical documents it is very important to have a greater accuracy of identification of Adverse Reactions on drug labels. Some of the main metrics could be the F1 Score to get a balanced Precision and Recall scores. Another important approach is using LSTM specifically Bi-directional LSTM using Tensorflow and keras layers to tackle NER Problem considering that the drug label is sequential data format, for NER since the adverse events covers pas nd future labels in a sequence we need consider both past and future information into account.

Deep Learning Architectures for detecting Adverse Events using Drug labels

Bi-LSTM-CRF – Bi Directional Long Short-Term Memory + Conditional Random Fields would be an Hybrid approach. CRFs have been widely used in the feature-based supervised learning approaches and using CRF as the tag decoder on top of the bidirectional LSTM layer in this proposed approach. While CRF is a common choice for tag decoder it still cannot make full use of segment-level as the inner properties of segments cannot be fully encoded with word-level representations, gated recursive semi-markov CRFs could be used to automatically extract segment-level through gated recursive convolutional neural network leveraging both word and segment-level information for segment calculation.

Bi-LSTM-CNN – This approach uses character embeddings and a LST to encode every word to an vector for a sequence of characters that represent a word, max-pooling architecture or a Convolutional Neural network could also be used, then the vectors a fed into another LSTM with learned word embedding.

Bi-LSTM-CNNs-CRF – This approach users word embeddings and character embeddings and fed into a Forward LSTM and the learned embeddings are the passed to a Conditional Random Filed layer for tag decodings. The challenge is the word embeddings may not have the full representation of its weights and would require a pre-trained language model to get its weighted scores.

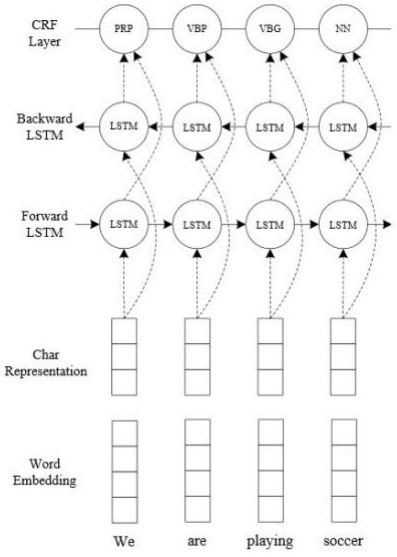


Figure 1 Bi-LSTM-CNN-CRF Architecture – Source Named Entity Recognition with Keras and Tensorflow

BioBERT + Bi-LSTM + CNN + CRF – This approach uses transfer learning process using BioBERT(BiDirectional Encoder Representations from Transformers for BioMedical Text Mining) alleviates the issue of out-of-vocabulary issue thus new words could be represented by frequent sub-words this is achieved by the large collection of word embeddings of 18B Bio Terms using PubMed and other large medical corpus. This also allows improves the quality of biomedical NER in drug safety text as compared to traditional methods and allows the approach go beyond the structured drug labels and identify ADR's from Social media.

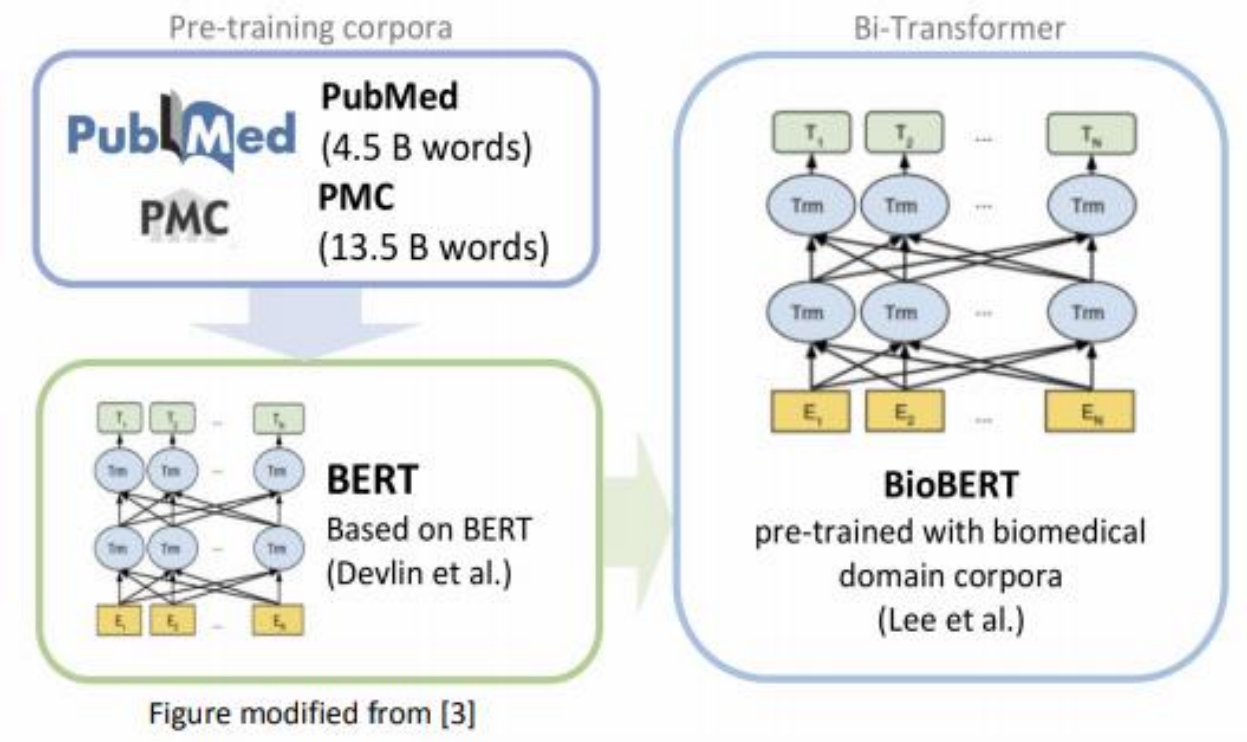


Figure 2 BioBERT Medical Word Embeddings source: http://web.stanford.edu/

This method leverages a pre-trained model with word embeddings to better detect the biological terms and achieves state-of-the-art performance, while the challenge remains of

language models while the lose the current context like BioBERT may not contain the current COVID situations and may require other

Conclusion:

The review on Deep learning based NER solutions using Tensorflow is to have a comprehensive understanding of extraction of FDA regulated Adverse Drug Reactions from drug labels. Including some traditional and current state-of the-art approaches including challenges, future research directions. The understanding of the definition of ADR is critical to modeling the architecture and taxonomy of the drug labels and lack of training data on annotated drug labels is a key data pipeline requirements in building NER models for ADRs. While Tensorflow and keras libraries are able to achieve promising accuracy there are also other abstractions like the transformers, Amazon Comprehend evaluation monitoring to design train models on large scale datasets. The drug label is a structured document and a different model has to be considered for extraction of ADR from patient community interactions like blogs, articles and other social media platform and continues to be a area of further research in a goal to save Patients lives. Adverse Event Reactions is a critical component in managing Patient health and help Medical professionals take necessary actions in saving lives, Machine Learning and Deep Learning plays a key role along with human involvement will help early detection of ADR and Helps Regulators design better Clinical studies and Drug formulations and have Madical professional educated on ADRs for a specific drug. The Deep learning approaches such as Convolutional Network and Neural Networks have been emerged and show a high potential on accomplishing the NER tasks using the TensorFLow and Keras NLP libraries.

References:

1. Food and Drug Administration. (2006). Guidance for Industry-Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products—Content and Format. *Rockville, MD: US Department of Health and Human Services*.
2. El-allaly, E. D., Sarrouti, M., En-Nahnahi, N., & El Alaoui, S. O. (2018, July). Adverse drug reaction mentions extraction from drug labels: an experimental study. In *International Conference on Advanced Intelligent Systems for Sustainable Development* (pp. 216-231). Springer, Cham.
3. Bisgin, H., Liu, Z., Fang, H., Xu, X., & Tong, W. (2011, December). Mining FDA drug labels using an unsupervised learning technique-topic modeling. In *BMC bioinformatics* (Vol. 12, No. S10, p. S11). BioMed Central.
4. Mehta, D., Uber, R., Ingle, T., Li, C., Liu, Z., Thakkar, S., ... & Zhou, G. (2020). Study of pharmacogenomic information in FDA-approved drug labeling to facilitate application of precision medicine. *Drug Discovery Today*.
5. Wu, Y., Yang, X., Bian, J., Guo, Y., Xu, H., & Hogan, W. (2018). Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. In *AMIA Annual Symposium Proceedings* (Vol. 2018, p. 1110). American Medical Informatics Association.
6. Gurulingappa H, Rajput A, Roberts A, et. al. Development of a benchmark corpus to support automatic extraction of drug-related adverse effects from medical case reports, J Biomed Inform 2012 Oct; 45(5):885-892. PubMed PMID: 22554702.
7. Sorbello, A., Ripple, A., Tonning, J., Munoz, M., Hasan, R., Ly, T., ... & Bodenreider, O. (2017). Harnessing scientific literature reports for pharmacovigilance: prototype software analytical tool development and usability testing. *Applied clinical informatics*, *8*(1), 291.
8. Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.
9. https://github.com/dmis-lab/biobert BioBERT: a pre-trained biomedical language representation model

10. FDA Guidelines Sources:

http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm075057.pdf

https://www.fda.gov/downloads/Drugs/Guidances/ucm075096.pdf

11. https://bionlp.nlm.nih.gov/tac2017adversereactions/
12. https://bionlp.nlm.nih.gov/tac2017adversereactions/AnnotationGuidelines_TAC2017ADR.pdf
13. https://www.meddra.org/
14. Demner-Fushman, D., Shooshan, S. E., Rodriguez, L., Aronson, A. R., Lang, F., Rogers, W., ... & Tonning, J. (2018). A dataset of 200 structured product labels annotated for adverse drug reactions. *Scientific data*, *5*, 180001.
15. https://github.com/lhncbc/fda-ars/tree/master/transform - Code for converting FDA Structured Product Labels (SPL) into formats suitable for use by the BRAT Annotation Tool and the MetaMap named-entity recognition tool.
16. https://www.meddra.org/how-to-use/basics/hierarchy
17. Tiftikci, M., Özgür, A., He, Y. et al. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. BMC Bioinformatics 20, 707 (2019). https://doi.org/10.1186/s12859-019-3195-5
18. Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.
19. Named Entity Recognition (NER) with keras and tensorflow