Biomedical Text Mining for Drug Repurposing Using a Word Embedding Implementation With SVM and Deep Neural Networks

Sruthi P. G. B150254CS
Shravani Sridhar B150062CS
Aditya Kumar B150476CS
Ahmed Bilal B150363CS
R. Swathy B140394CS

Department of Computer Science and Engineering
NIT Calicut

28th March, 2019

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
    - Computational Method for Drug Repurposing
    - Network-based Approach for Drug Repurposing
    - Biomedical Text Mining Using a Full Parser
    - Text Mining using Bag-of-Words (BoW) Model for Prediction
    - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
    - Dataset
    - Structure
    - Method
5. Pseudocode
6. Results
7. References

Abstract

- Drug repurposing is the process of applying an existing drug to treat a different disease than the one it was originally used for.
- One major challenge in drug repurposing is finding new drug- disease relationships.
- One approach for this is text mining.
- To counter this, recently a new approach has been tested, namely, word embedding, which generates relatively short numerical vectors as representations of word sense.
- The final classification model, which was based on SVM (Support Vector Machine).

Problem Definition

- We aim to repurpose existing and approved drugs using biomedical text mining of a large cancer-related corpus, specifically using a word embedding implementation to generate semantically equivalent representations of words in the text and then a deep neural network for predicting drug-disease relationships from the representations.
- We will also use an SVM for the task of prediction and then compare the results with that from using a DNN.

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Computational Method for Drug Repurposing

- A data mining process using publicly available gene expression datasets associated with a few diseases and drugs was carried out to identify existing drugs that could be used to treat genes causing lung cancer and breast cancer.
- A more recent, alternative approach to drug repurposing is text mining. We will use biomedical text mining, specifically.

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
    - Computational Method for Drug Repurposing
    - Network-based Approach for Drug Repurposing
    - Biomedical Text Mining Using a Full Parser
    - Text Mining using Bag-of-Words (BoW) Model for Prediction
    - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
    - Dataset
    - Structure
    - Method
5. Pseudocode
6. Results
7. References

Network-based Approach for Drug Repurposing

- A network-based approach for drug repurposing was utilized that takes into account the human interactome network, proximity measures between drug targets and disease-associated genes, potential side-effects, genome-wide gene expression, and disease modules that emerge through pertinent analysis.
- The network-based approach was found to provide a fast and efficient way to determine likely candidates for drug repurposing and understand their underlying mechanisms

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Biomedical Text Mining Using a Full Parser

- A system which uses a full parser for analyzing biomedical text was developed. A pre-processor was used to partially overcome the shortcomings of full parsing.
- A full parser requires an oversized memory size, and it is slower in execution.

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
    - Computational Method for Drug Repurposing
    - Network-based Approach for Drug Repurposing
    - Biomedical Text Mining Using a Full Parser
    - Text Mining using Bag-of-Words (BoW) Model for Prediction
    - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
    - Dataset
    - Structure
    - Method
5. Pseudocode
6. Results
7. References

Text Mining using Bag-of-Words (BoW) Model for Prediction

- It is a technique for computationally representing linguistic units in text like sentences, words, and documents.
- An easy-to-use framework was devel- oped for accelerated usage of the BoW model in text mining and processing.
- Due to the large dimensionality and high sparseness of the vectors that the BoW model tends to generate, other approaches such as word embedding have been proposed.

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing

- The study overall showed that deep learning can achieve reasonable performance in drug repurposing (DNN had achieved the best ROC-AUC for depression/anxiety disorders in the weighted analysis.).
- We plan to use a larger dataset to test the effectiveness of our deep neural network in predicting drug-disease relationships from the word sense representations. We will also use an SVM approach for the same. Then we will compare the performances of both and see if the former's is better.

Project Design

1. POS tagging for sentence parsing.
2. word2vec algorithm for word embedding.
3. Comparison of the performances of Continuous Bag-of-Words (CBOW) and skip-gram training algorithms in word2vec.
4. Comparison of the performances of DNN and SVM in the prediction of drug-disease relationships using our large dataset.
5. Python 3.x as our programming language.
6. TensorFlow library for numerical computation that makes deep learning easier.

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Dataset

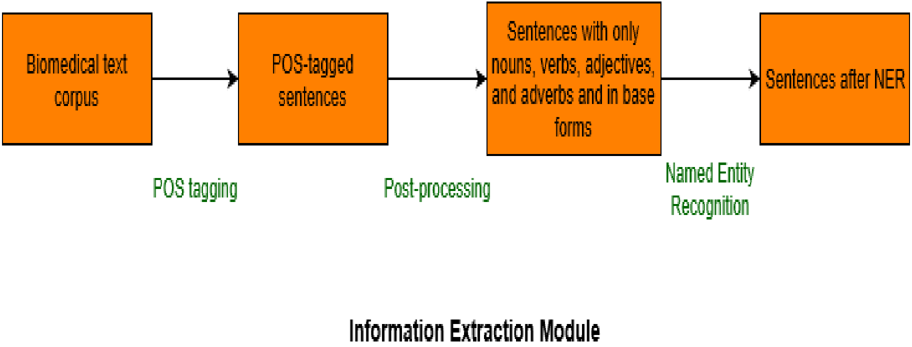
- We will use a set of cancer-related reports downloaded from PubMed as our raw corpus.

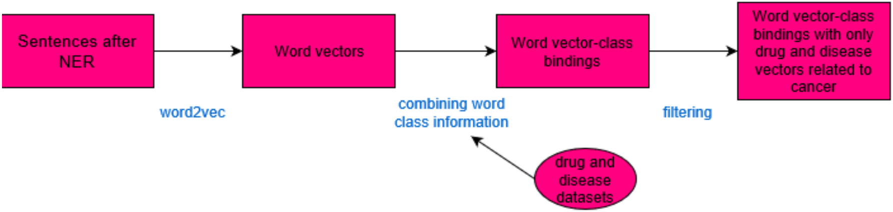
Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Structure

- Our project will be divided into three modules as follows





Sentences after NER

word2vec

Word vectors

combining word class information

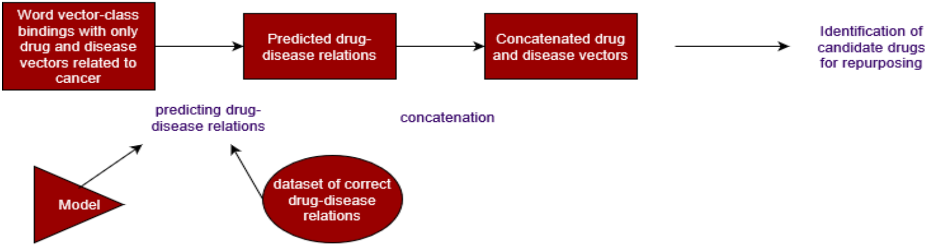
drug and disease datasets

Word vector-class bindings

filtering

Word vector-class bindings with only drug and disease vectors related to cancer

Word Embedding Module



Word vector-class bindings with only drug and disease vectors related to cancer

predicting drug-disease relations

Model

dataset of correct drug-disease relations

Predicted drug-disease relations

concatenation

Concatenated drug and disease vectors

Identification of candidate drugs for repurposing

Classification and Prediction Module

Outline

1. Abstract
2. Problem Definition
3. Literature Survey
   - Computational Method for Drug Repurposing
   - Network-based Approach for Drug Repurposing
   - Biomedical Text Mining Using a Full Parser
   - Text Mining using Bag-of-Words (BoW) Model for Prediction
   - Comparing Machine Learning and Deep Learning Approaches for Drug Repurposing
4. Project Design
   - Dataset
   - Structure
   - Method
5. Pseudocode
6. Results
7. References

Method

After executing the first two modules on the training set, we will train our models separately on the output of the second module and obtain the results.

- Training: After executing the first two modules on the training set, we will train our models separately on the output of the second module and obtain the results.
- Testing: We will compare the results of the performances of using SVM and DNN and conclude.

Pseudocode

The following is the general algorithm for classifying data using SVM:

```
1. Import the dataset
2. Encode the target features as vectors
3. Split the dataset into training and testing
4. Perform feature scaling
5. Fit SVM to the training set
6. Predict the test set results
7. Visualize the training set results
8. Visualize the test set results
```

We will use the word2vec software for word embedding implementation.

Results

- The DNN had a higher classification accuracy than SVM and hence better performance in predicting new drug-disease relationships.
- With a smaller dataset, both word embedding training algorithms (CBOW and Skip-gram) gave equivalent outputs, i.e, equivalent word vector representations of input words.
- Skip-gram was better able to understand words that were infrequent (with respect to the corpus) than CBOW was; hence the accuracy of the Skip-gram is higher than the CBOW model for larger dataset.

Conclusion

- In this presentation, we explained the abstract and problem definition of our project.
- We discussed the literature survey we conducted, and then we described our project design.
- The design elaborated on the features we are going to include in the project, the dataset we plan to use, the overall structure of our project, and lastly our method.
- Finally, we introduced pseudocode important for our project.

References

- K. M. Shabana, K. A. A. Nazeer, M. Pradhan, and M. Palakal. A computational method for drug repositioning using publicly available gene expression data. BMC Bioinformatics, 2015, vol. 16, https://doi.org/10.1186/1471-2105-16- S17-S5.
- K. Kim, N. Rai, M. Kim, and I. Tagkopou- los. A network-based model for drug re- purposing in Rheumatoid Arthritis. Univer- sity of California, California, US, 2018, https://doi.org/10.1101/335679.
- P. Govindarajan, K. S. Ravichandran. Text mining from biomedical domain using a full parser. 2016 International Conference on Inventive Computation Technologies, 2016, vol. 3, https://doi.org/10.1109/INVENTIVE.2016.7824887.
- D. S, P. Raj, and S. Rajaraajeswari. A Framework for Text Analytics using the Bag of Words (BoW) Model for Predic- tion. International Journal of Advanced Networking Applications, Raja Rajeswari College of Engineering, Bangalore, India.

References

- D. L. Ngo, N. Yamamoto, V. A. Tran, N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. Kubo, K. Satou. Application of Word Embed- ding to Drug Repositioning. J Biomedical Sci- ence and Engineering, 2016, vol. 9, pp. 7-16, https://doi.org/10.4236/jbise.2016.91002.

- Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa. Electron spectroscopy studies on magneto- optical media and plastic substrate interface. IEEE Transl. J. Magn. Japan, vol. 2, pp. 740- 741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

- M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Thank You