Chemical – Protein Interaction via Machine Learning and Deep Learning Approaches

Angel Paul, Jithin Antony Joseph, Reethu Biju, Uditi Bajaj

Abstract

Relation extraction is a process that helps extract meaningful associations from unstructured data by finding connections between entities in text. It finds many uses in natural language processing, such as knowledge graph construction, information retrieval, and analysis of biomedical literature. Relation extraction is utilised in biomedical corpora such as CHEMPROT to reveal complex relationships between chemicals and proteins, which helps to uncover important information about drug interactions, molecular pathways, and biomedical research. In this paper, two new approaches that contribute differently to relation extraction are presented. The first approach demonstrates the effectiveness of Support Vector Machines (SVM) using a vectorised dataset with computed relative positions of entities. Combining Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory Networks (Bi-LSTM), and BioBERT, the second approach is a weighted ensemble that employs a sophisticated voting mechanism. The weighting of the ensemble model enhances the interpretation of each component, highlighting the synergistic power of deep learning architectures in illuminating intricate relationships found in biomedical texts. To evaluate the performance of the proposed models, precision, recall, and F1 score metrics were employed. The comprehensive evaluation highlighted the ensemble model's enhanced ability to discern complex relationships, surpassing the individual SVM model in precision, recall, and overall F1 score, thus affirming its superior performance in relation extraction from biomedical data.

Introduction

Over the past few decades, accurately identifying interactions between chemical and biological entities has become increasingly critical, particularly in the biomedical domain where relation extraction is essential. It is within this context that BioCreative has introduced tasks focused on extracting relationships between biological entities, such as protein-protein, chemical-induced diseases, and notably, chemical-protein interactions (CPI) [1]. These CPI tasks are crucial for drug discovery and ensuring the safe co-administration of medications, as they allow for the detailed depiction of interactions and organize this information in a structured format.

Building on the foundation of accurately identifying these interactions, our work specifically utilizes the CHEMPROT dataset. This dataset is obtained by applying text mining techniques to a vast array of scientific articles, with a significant focus on PubMed abstracts [1]. The CPI data is also collected from open sources like ChEMBL, BindingDB, and DrugBank, alongside commercial databases WOMBAT and WOMBAT-PK, focusing on active compounds from validated experiments for high-confidence annotations. Additionally, drug-target information and chemical-protein effects are incorporated from CTD and STITCH, enriching the dataset with both binding activities and modulation effects of chemicals on proteins [2]. Enhanced with gold-standard entities [3], the dataset encompasses 3 columns which are, i) text- the sentences extracted from the articles ii) label- the relationship between the entities, and iii) metadata. To ensure the integrity of the dataset, a thorough examination of duplicate records- instances where both entities and their corresponding labels are identical was conducted. This review confirmed the absence of any repeated records, affirming the dataset's uniqueness and reliability for subsequent analyses. Unlike typical tasks that may require explicit named entity recognition, this work emphasizes the extraction of relationships between two target entities, streamlining the process by focusing solely on the connections rather than the entities themselves [4]. The dataset is divided into training, development, and test partitions, with the training and development segments merged for model training purposes. The training, development, and test have 4168, 2426, and 3469 respectively. Checks were performed to identify duplicate records, where the entities and the labels are the same, but it was identified that there were no repeated records. Upon examining the distribution of records across various relation classes, it became evident that there is an uneven distribution, as detailed in Table 1.

Relation	Record counts
INHIBITOR	2454
SUBSTRATE	802
INDIRECT-DOWNREGULATOR	757
INDIRECT-UPREGULATOR	665
ACTIVATOR	580
ANTAGONIST	434
PRODUCT-OF	363
AGONIST	267
DOWNREGULATOR	152
UPREGULATOR	84
SUBSTRATE PRODUCT-OF	19
AGONIST-ACTIVATOR	15
AGONIST-INHIBITOR	4

Table 1. Distribution of labels across records in the merged train and dev partitions

In biomedical relation extraction, the context in which words are used is crucial, particularly when identical entities have different relational labels. Traditional embeddings give each word a fixed vector, ignoring context. In contrast, advanced embeddings like ELMo consider the full sentence, capturing the nuanced meanings essential for precise tasks like CPI extraction [5].

Our paper examines two relation extraction techniques: a selective hybrid weighted voting method combining BiLSTM,

BioBert, and CNN models, and an SVM-based standalone method. BiLSTM, CNN, and SVM undergo the following preprocessing steps: stopword removal, POS tagging, lemmatization, and ELMo embedding application. BioBert, conversely, processes context natively, without such steps. To tackle class imbalance, BiLSTM uses Gaussian Probability Distribution-based class weights. The first technique employs a weighted voting mechanism on BiLSTM, BioBert, and CNN outputs, pooling insights to determine the predominant relations. The second technique supplements SVM with relative position features and ELMo embeddings to discern intricate details about spacial context and directionality, hence enhancing extraction accuracy.

Our paper's structure is as follows: it starts with a concise overview of existing research work. The subsequent section provides details on the preprocessing performed and models used for relation extraction. We then demonstrate our findings using various evaluation metrics. The paper concludes with a discussion of the results, limitations of our approaches and offers suggestions for potential areas of exploration for future research.

Related Works

With the introduction of novel neural network-based techniques, the field of Chemical-Protein Interaction (CPI) extraction from biomedical literature has experienced a notable transformation. Analysing important research in this field, with a focus on the CHEMPROT challenge, reveals a range of approaches and ideas that work together to improve the precision of CPI extraction for a range of biomedical uses.

An innovative method for CPI extraction achieves an excellent F1-score of 76.56% on the CHEMPROT corpus by incorporating external biomedical knowledge and the Gaussian probability distribution into a neural network-based model [6]. By adding a probabilistic component, this innovative approach improves the model's comprehension of chemical-protein interactions and advances accuracy overall.

Another study looks at recursive neural networks in the context of the CHEMPROT challenge and presents three different methods, including a tree-LSTM model with additional features like position and subtree containment features, and a Stack-augmented Parser Interpreter Neural Network (SPINN) model [4]. Improvements made after the challenge yield higher F1-scores (63.7% and 64.1%), highlighting the importance of model architecture, preprocessing stages in the extraction of CPI and importance of iterative development of models.

An alternative viewpoint is provided by the creation of an ensemble system that combines SVM, CNN, and RNN for the extraction of chemical-protein relations [7]. By combining the advantages of several models, ensemble systems offer a comprehensive method for CPI extraction, help to provide a deeper comprehension of the interactions between chemicals and proteins and provides a nuanced solution to the difficulties presented by CPI extraction.

Deep learning models like CNN and Bi-LSTM networks are used to improve information retrieval and automatic relation extraction [3]. Through the use of CNN and Bi-LSTM recurrent networks, deep learning models demonstrate their effectiveness in producing precise CPI extraction, opening the door to more extensive uses in relation extraction and information retrieval.

Attention-based neural networks are used for chemical-protein relationship extraction, such as CNN and attention-based RNN [8]. The research identifies areas for additional attention-based model development by recommending future work on examining external knowledge bases and applying token-level weights. The extraction process is given a more nuanced focus by attention-based models, which highlights the significance of particular tokens and their connections to chemical-protein interactions.

A study that integrates deep contextualised word representations such as ELMo and multihead self-attention mechanisms highlights the importance of accurately detecting chemical-protein interactions [5]. With a high F1-score of 65.9% on the ChemProt corpus, the suggested neural model combines ELMo, Bi-LSTMs, and multihead attention. In terms of encapsulating the complex subtleties of chemical-protein interactions, deep contextualised word representations and multihead self-attention mechanisms mark a significant advancement.

By using various neural network-based techniques, these studies collectively advance the continuous progress in CPI extraction. The development of techniques is evident in these developments, which lay the groundwork for further research and investigation in the field. These developments range from ensemble systems and attention-based models to Gaussian probability distribution. An all-encompassing understanding of chemical-protein interactions is made possible by the combination of external knowledge integration, model architecture exploration, simplicity, and attention mechanisms. The field of CPI extraction is constantly changing as a result of these studies, opening up new avenues for application in biomedical research and other fields.

Methodology

Data Pre-Processing. Prior to any computational analysis, we have made sure to perform a rigorous data pre-processing process. Subsequently, we have executed an intricate sequence of computational linguistic techniques designed to dissect and understand the textual data comprehensively. Starting with merging train and dev datasets to diversify our corpus to cover all possible entity combinations and relations, we then split the two entities into individual columns E1 and E2, as well as dropped the metadata column. Further, after removing stop words using NLTK library to reduce noise and filler words, we have exploited features such as tokenization, part-of-speech (POS) tagging, lemmatization, deep contextualized word representations via ELMo, and dependency parsing: **Tokenization** served as the foundational step, segmenting the corpus into discrete tokens. This phase is critical as it transforms raw text into a structured format that is amenable to computational processing. **Part-of-Speech (POS) Tagging** was applied, where each token was annotated with its corresponding grammatical category, depending on the word and its context. This step is requisite for uncovering the grammatical backbone of the sentences, thereby aiding in clarifying the meaning of words that may have multiple grammatical functions. **Lemmatisation** ensued, wherein tokens were normalized to their lemma or dictionary form. This normalization is important in natural language processing for reducing the complexity of the linguistic data, allowing the model to effectively generalize across different morphological forms of a word. To capture the nuanced meanings of

words in context, we utilized Deep Contextualized Word Representations via ELMo. ELMo embeddings are derived through a sophisticated model that accounts for the complex characteristics of word use and how these uses vary across linguistic contexts [5]. The formula for generating these representations is as follows:

$$ELMo_{\text{word}} = \gamma^{\text{task}} \sum_{j=0}^{L} s_j^{\text{task}} h_j^{\text{word}}$$

where h_j^{word} denotes the layered representation from the bi-directional Language model (biLM), s_j^{task} are softmax-normalised mixture model weights, and γ^{task} is a scaling factor for the overall usefulness of the ELMo task.

Further, the preprocessing methodology of our research is characterised by its novel approach to **Relative Position Features**, which offer directionality and spatial context for entities within sentences. Especially for machine learning models, these valuable features enhance understanding and identification of potential relationships, and bolster contextual representations of entities. This concept allows for the nuanced capture of relationships between entities, improving the accuracy and depth of relation extraction tasks by providing a comprehensive view of entity interactions. Following is the simplified mathematical formula which quantifies the relative positions of words in a text by measuring their average distance to specific entities (e1pos and e2pos)

$$\text{avg}\left(\left|\frac{i}{\text{lenS}} - \text{e1pos}\right| + \left|\frac{i}{\text{lenS}} - \text{e2pos}\right|\right)$$

for i = 1, 2... no. of words

where $e1pos$ and $e2pos$ are calculated as the entity's position divided by the total number of words in the sentence, providing a relative measure of location.

Approach	Model	PreProcessing Steps
Standalone approach	SVM	Input 1: Stop words removal, POS Tagging, Lemmatization, ELMo Embedding Input 2: Computed Relative position feature
Selective Hybrid Weighted Voting approach	CNN	Stop words removal, POS Tagging, Lemmatization, ELMo Embedding
	BiLSTM with Class weights	
	BioBERT	Replace Entity 1 with Entity and Entity 2 with Entity2, Tokenization

Table 2. Preprocessing steps performed for each model

Support Vector Machines. SVM, capable of producing state-of-the-art results, maps the input vectors x into a high-dimensional feature space Z to find an optimal separating hyperplane that uses hinge loss as its objective function. Given a set of training samples, a linear SVM's task is to effectively separate positive from negative instances using a hyperplane ensuring the widest margin possible. In scenarios, where the samples are not linearly separable, the SVM employs a kernel function to implicitly map inputs into high-dimensional feature spaces where the data might be separable. For relation extraction tasks, SVMs are highly valuable due to their ability to discern intricate boundaries in complex feature spaces, crucial for identifying semantic relationships in text. In our approach, upon performing data preprocessing steps, we have then used ELMo embeddings for deep contextualisation and transformed these features into high dimensional vector space. We passed this vectorised input, along with relative position features, into the SVM model, enabling it to effectively learn and distinguish the complex relation patterns in our corpus.

Selective Hybrid Weighted Voting Approach. In our second approach, we have used CNNs, Bi-LSTM and Bio-BERT [12] that run independently whose output is then passed into the voting process [7]. The CNN was implemented using tensor flow with 1D convolutional layer with 64 filters and a kernel size of 3, using ReLU activation, followed by a max pooling layer with pool size 2. It then flattens the output and connects to a dense layer with 128 neurons (ReLU activation), and a softmax output layer matching the number of relations we have which is 13. The model is compiled with the Adam optimizer and categorical crossentropy loss, measuring accuracy and loss as metrics with a train-test split of 80-20 respectively. CNNs are chosen for relation extraction because of their exceptional ability to automatically extract local, relevant features from text, identifying patterns like phrases that indicate specific relationships. Additionally, their computational efficiency makes them suitable for processing large datasets quickly. The Bi-LSTM model holds a bidirectional LSTM layer with 64 units, a specified vector dimension, followed by another bidirectional LSTM layer with 64 units. A dense output layer with 13 neurons uses softmax activation for multi-class classification. Compiled with Adam optimizer and categorical cross entropy loss, it measures accuracy. Class weights determined using Gaussian probability distribution were added as a feature to address the problem of class imbalance [6]. This helps in improving the model's ability to generalize across all classes whereby reducing any chances of majority classes taking control. The class weights passed into the model have been modified using Gaussian probability to assign greater weight to less frequent classes [6]. In entirety, Training set involves encoded class labels with label encoding, class weights for imbalance handling, and 80-20 validation split for performance evaluation. Predictions are made on test data, converting probabilities to class labels, then decoding to original labels for comparison with true labels. BiLSTMs excel in understanding the full context of a sentence by analysing text from both directions, which is crucial for accurately capturing relationships between entities. They also effectively manage long-term dependencies, allowing them to recognize connections between entities that are far apart in the text. To construct our third model, we performed tokenisation which was then passed onto Bio-BERT. The setup fine-tunes a pre-trained BioBERT model for sequence classification, using a DataLoader in batches of 8 and shuffling. It utilizes AdamW op-

timizer with a learning rate of 5e-5, runs on GPU if available, and iterates over 3 epochs, displaying loss updates in a tqdm loop for monitoring prior to which the dataset was label encoded and then tokenised. Bio-BERT by itself is an extremely powerful model tailor made for this particular use case as it by itself being trained on BioMedical texts through Tokenisation, Embedding Conversion, Contextual Embedding, Layer Wise Transformation, Pooling, Fine-Tuning, Output Generation, Loss Calculation and then finally Backpropagation for Optimisation. The model was tuned on the full dataset without employing a train-validation split. Having developed three independent models, each demonstrating satisfactory performance, we are now prepared to pass them onto the selective weighted hybrid voting process as it helps in reduction of overfitting, increased overall stability and improved accuracy by leveraging model diversity. Each model independently casts a vote for the relation type it identifies, contributing its own output to the decision-making process. In case all the three models have different outputs then it goes into a weighted output where the models have been given weights of 0.5,0.3 and 0.2 for Bio-BERT,Bi-LSTM and CNN respectively. For a given sentence, the relation receiving the majority of votes from the models is selected as the final relation label. This is again evaluated on top of the test dataset that has been pre-processed similar to the train data.

Results and Discussions

To evaluate the predictive power of our models, we have selected the F1-score as our primary metric due to its balanced consideration of precision and recall. It offers a comprehensive measure of a model's accuracy, particularly in relation extraction tasks, where the ability to correctly identify relationships amid complex interactions without excessively misclassifying irrelevant instances is crucial for maintaining the integrity of the extracted information.
By integrating relative position features and ELMo embeddings into the preprocessing pipeline, our SVM model attained an F1 score of 0.70 in training and 0.49 in testing, illustrating its proficiency in identifying relational patterns during training, with a notable performance variance when applied to unseen test data.
Table 3 exhibits the independent performance of the models on the validation split. CNN has a right balance between the precision and recall which is evident in the F1-Score. Bi-LSTM scores lower on the metrics against CNN suggesting that it might be less accurate and inconsistent in predicting the true positive values. Table 3 displays the standalone performance metrics of the models evaluated on the test subset. Unexpectedly, the CNN exhibits a pronounced decline in its performance metrics on the test set relative to the validation metrics wherein the F1 Score dropped from 0.74 to 0.545, indicating its inability to adequately generalize its predictions. This trend points towards a certain degree of overfitting during the training process. Bio-BERT significantly surpasses the other models, delivering exceptionally strong performance metrics, which states its effectiveness in accurately identifying the majority of positive cases. This superior performance aligns with expectations, considering that Bio-BERT is specifically designed for this domain.
Table 4 shows the final results from the Weighted Hybrid Voting where the combined approach yields a precision of 0.689, a recall of 0.68, and an F1-Score of 0.68 on the test set. This indicates a substantial improvement in model generalization over the individual CNN and Bi-LSTM models. The ensemble method effectively integrates the strengths of individual models, thereby enhancing prediction reliability and reducing the likelihood of overfitting seen in single model approaches.

	Train-Dev Split		**Test Split**		
	CNN	**Bi-LSTM**	**CNN**	**Bi-LSTM**	**BioBERT**
P	0.74	0.652	0.545	0.554	0.808
R	0.718	0.54	0.521	0.445	0.807
F1	0.722	0.568	0.526	0.478	0.8

Table 3. Individual Model Performances

	SVM		**Weighted Hybrid Voting**
	Train	**Test**	
Precision	0.70	0.51	0.689
Recall	0.71	0.53	0.68
F1-score	0.70	0.49	0.68

Table 4. Final Approach Metrics

Conclusion

In this research, we investigate two distinct approaches, the selective hybrid weighted voting approach and a standalone SVM, for CPI extraction on the CHEMPROT dataset. Our findings reveal that the BioBert model alone sets a benchmark with an F1-score of 0.80, demonstrating unparalleled performance in comparison to other models tested individually. Building on this, our selective hybrid weighted voting strategy combines multiple models, achieving an F1-score of 0.68. The utilization of ELMo embeddings significantly contributes to capturing semantic context, thus enhancing the efficacy of models not specifically tailored to domain-specific tasks. Our approach innovatively tackles class imbalance by incorporating class weights into the BiLSTM model, leveraging Gaussian Probability distribution. This approach improves the model's performance by fine-tuning class weights, ensuring that the dominant class does not overshadow the others. These findings indicate that our ensemble system is effective at identifying chemical-protein interactions within biomedical texts. Incorporating a relative position feature as an additional input into our SVM model has resulted in an F1-score of 0.49. A notable limitation encountered is the substantial amount of computational resources required by the BioBert model, necessitating extensive time for execution. The SVM results were adequate, considering computational constraints. Efforts to add dependency parsing features, known for enhancing model quality, were hindered by memory and processing limitations, necessitating their exclusion despite their potential to substantially boost performance. Future endeavors will focus on advancing relation extraction techniques through the integration of external biomedical information to enrich analysis and a thorough investigation into the role of cross-sentence dynamics in understanding complex interactions, setting the stage for more sophisticated and accurate models.

References

[1] Warikoo, N., Chang, Y.-C. and Hsu, W.-L. (2018). LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task. Database, 2018. doi:https://doi.org/10.1093/database/bay108.

[2] Olivier Taboureau, Sonny Kim Nielsen, Audouze, K., Weinhold, N., Edsgärd, D., Roque, F.S., Kouskoumvekaki, I., Bora, A., Curpan, R., Thomas Skøt Jensen, Søren Brunak and Oprea, T.I. (2010). ChemProt: a disease chemical biology database. Nucleic Acids Research, 39(Database), pp.D367–D372. doi:https://doi.org/10.1093/nar/gkq906.

[3] Extraction of chemical–protein interactions from the literature using neural networks and narrow instance representation. (2019). Database. doi:https://doi.org/10.1093/database/baz095.

[4] Lim, S. and Kang, J. (2018). Chemical–gene relation extraction using recursive neural network. 2018. doi:https://doi.org/10.1093/database/bay060.

[5] Zhang, Y., Lin, H., Yang, Z., Wang, J. and Sun, Y. (2019). Chemical–protein interaction extraction via contextualized word representations and multihead attention. Database, 2019. doi:https://doi.org/10.1093/database/baz054.

[6] Sun, C., Yang, Z., Su, L., Wang, L., Zhang, Y., Lin, H. and Wang, J. (2020). Chemical–protein interaction extraction via Gaussian probability distribution and external biomedical knowledge. Bioinformatics, 36(15), pp.4323–4330. doi:https://doi.org/10.1093/bioinformatics/btaa491.

[7] Peng, Y., Rios, A., Kavuluru, R. and Lu, Z., 2018. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models. arXiv preprint arXiv:1802.01255.ting

[8] Liu, S., Shen, F., Ravikumar Komandur Elayavilli, Wang, Y., Majid Rastegar-Mojarad, Chaudhary, V. and Liu, H. (2018). Extracting chemical–protein relations using attention-based neural networks. Database, 2018. doi:https://doi.org/10.1093/database/bay102.

[9]allenai.org. (n.d.). AllenNLP - ELMo — Allen Institute for AI. [online] Available at: https://allenai.org/allennlp/software/elmo.

[10] Fundel, K., Kuffner, R. and Zimmer, R. (2006). RelEx–Relation extraction using dependency parse trees. Bioinformatics, 23(3), pp.365–371. doi:https://doi.org/10.1093/bioinformatics/btl616.

[11] Handmark, O. (2020). NLP: Deep learning for relation extraction. [online] Medium. Available at: https://towardsdatascience.com/nlp-deep-learning-for-relation-extraction-9c5d13110afa [Accessed 8 Mar. 2024].

[12] Satoshi Hiai, Kazutaka Shimada, Taiki Watanabe, Akiva Miura, and Tomoya Iwakura. 2021. Relation Extraction Using Multiple Pre-Training Models in Biomedical Domain. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 530–537, Held Online. INCOMA Ltd..