informatics

MDPI

Article

Knowledge-Based Intelligent Text Simplification for Biological Relation Extraction

Jaskaran Gill [1,*], Madhu Chetty [1,*], Suryani Lim [1] and Jennifer Hallinan [1,2]

[1] Health Innovation and Transformation Centre, Federation University, Ballarat, VIC 3842, Australia; suryani.lim@federation.edu.au (S.L.); j.hallinan@biothink.net (J.H.)
[2] BioThink Pty Ltd., Brisbane, QLD 4020, Australia
* Correspondence: jaskarankaurgill@students.federation.edu.au (J.G.); madhu.chetty@federation.edu.au (M.C.)

Abstract: Relation extraction from biological publications plays a pivotal role in accelerating scientific discovery and advancing medical research. While vast amounts of this knowledge is stored within the published literature, extracting it manually from this continually growing volume of documents is becoming increasingly arduous. Recently, attention has been focused towards automatically extracting such knowledge using pre-trained Large Language Models (LLM) and deep-learning algorithms for automated relation extraction. However, the complex syntactic structure of biological sentences, with nested entities and domain-specific terminology, and insufficient annotated training corpora, poses major challenges in accurately capturing entity relationships from the unstructured data. To address these issues, in this paper, we propose a **K**nowledge-based **I**ntelligent **T**ext **S**implification (KITS) approach focused on the accurate extraction of biological relations. KITS is able to precisely and accurately capture the relational context among various binary relations within the sentence, alongside preventing any potential changes in meaning for those sentences being simplified by KITS. The experiments show that the proposed technique, using well-known performance metrics, resulted in a 21% increase in precision, with only 25% of sentences simplified in the Learning Language in Logic (LLL) dataset. Combining the proposed method with BioBERT, the popular pre-trained LLM was able to outperform other state-of-the-art methods.

Keywords: sentence simplification; named entity recognition; relation extraction; BioBERT; BERN2

Citation: Gill, J.; Chetty, M.; Lim, S.; Hallinan, J. Knowledge-Based Intelligent Text Simplification for Biological Relation Extraction. *Informatics* **2023**, *10*, 89. https://doi.org/10.3390/informatics10040089

Academic Editor: Zhiwen Yu

Received: 16 October 2023
Revised: 27 November 2023
Accepted: 5 December 2023
Published: 11 December 2023

1. Introduction

Biological relationships refer to the connections, interactions, or associations between different biological entities within living organisms or in biological systems [1,2]. These biological relationships are crucial for understanding the complexity of living organisms and their interactions with the environment. In recent times, the discovery of biological relationships has progressed significantly due to advancements in cutting-edge technologies, such as high-throughput sequencing, advanced imaging techniques, and computational tools, and these findings are published in research papers [3–5]. The observed biological interactions buried in the published literature hold significant value, contributing to drug discovery, treatment development, and the comprehension of disease progression. With the explosive growth in the published literature in the biological domain, biological relation extraction (RE) is becoming increasingly challenging. Figure 1 depicts the exponential growth in the published literature related to gene regulation [6,7]. Despite the significant progress in Natural Language Processing (NLP), manual extraction remains the primary method for RE in most public repositories [8]. While the approach produces high accuracy and precision, the time-consuming nature of manual curation acts as a hindrance to scalability and efficiency. Given the exponential growth in the literature, keeping up with new discoveries can be challenging and may lead to delays in implementing applications.

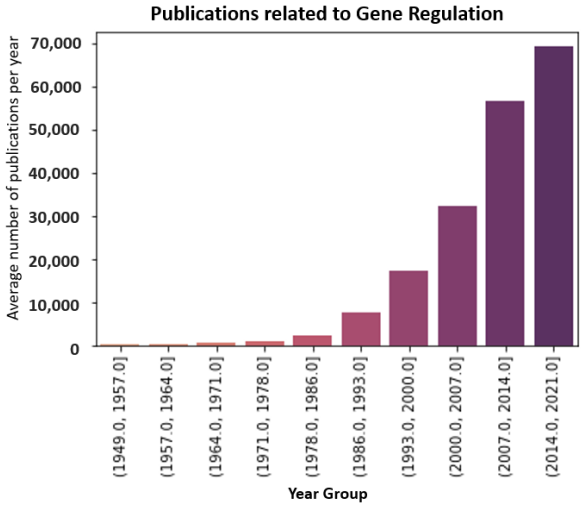

Figure 1. The biological literature related to gene regulation published over the years according to Medline [7].

Despite the substantial efforts devoted to implementing machine- and deep-learning NLP techniques in biological RE, the accuracies of existing approaches continue to be relatively low [9]. Community challenges and workshops in information extraction and related fields serve as collaborative platforms for researchers from diverse backgrounds to address shared problems and benchmark state-of-the-art methods. The BioCreative VI ChemProt challenge [10], held in 2017, comprised two tasks: the extraction of relational pairs and the identification of interaction types, such as inhibition, binding, or induction associations. Peng et al. emerged with the top-performing method in the BioCreative VI ChemProt challenge, utilising an ensemble model that incorporated various machine-learning algorithms, including support vector machines and deep-learning models [11]. Despite employing machine learning and deep-learning architectures, the ensemble method produced subpar results, with a precision of 0.7266 and a recall of 0.5735, resulting in an F-score of 0.6410. Their paper emphasised that the proposed ensemble method encountered challenges in relation extraction, particularly with longer sentences in which the distance between entities was substantial.

Several factors exist which limit the performance of RE methodologies in the context of biological text. These include, for example, the use of domain-specific terminology, the structural complexity of biological sentences, the limited availability of annotated data, and the model's lack of domain-specific knowledge and biological context [9,12,13]. Among the factors, complex sentence structure poses a significant challenge in RE because the biological literature frequently features complex sentence structures, including nested clauses and long sentences [14,15]. Figure 2 depicts the distribution of the length of sentences (the number of words in each sentence) for the LLL dataset [16]. The length of the longest sentence in the LLL dataset is 89 words containing six phrases and 10 gene/protein entities, some of which may be repeated. The presence of multiple entities and phrases affects the ability of a model to accurately predict true regulatory interactions. Further, the presence of highly specialised and domain-specific terms that may not be well represented in general language models further amplifies the challenge in extracting biological relationships accurately [9,17,18].

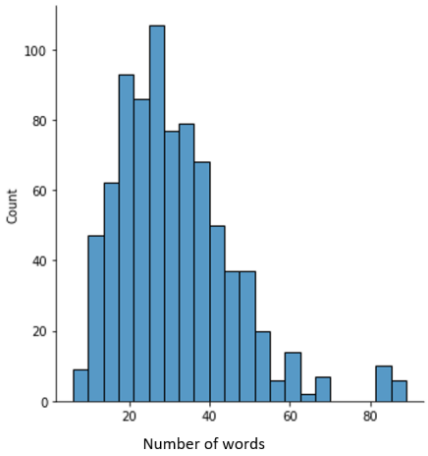

Figure 2. Distribution of number of words in the sentences in the LLL dataset.

The limitations of relation classification highlighted above, caused by the structural complexity of biological text, can be alleviated by simplifying complex sentences, ensuring that no information content is lost. Early attempts to simplify sentences used syntax and part-of-speech tags such as nouns and verbs [19,20]. Bach et al. treated sentence simplification for relation extraction as a statistical problem [21]. These authors divided a full sentence into a set of all possible simpler sentences and used a probability distribution score to select the set that preserved the most information from the original sentence. Hakenberg et al. proposed paraphrasing sentences by eliminating unwanted filler words and simplifying the syntax of the sentence [22]. Syntactic simplification can be relatively ineffective in the biological domain, due to complicated nature of sentence formation, named entities, and relation descriptions. Moreover, such sentence reduction techniques can even alter the complete meaning of a sentence. Miao et al. proposed a methodology that relied on a prior knowledge base to extract relevant noun and verb entities to generate a simpler sentence formation [23]. This method relies heavily on the availability of an up-to-date knowledge repository, and is challenging due to a significant increase in newly discovered entities and their interactions. Recent attempts at medical sentence simplification have used trained model-based systems [24]. These efforts also focus on abstract-level paraphrasing. Devraj et al. retrieved review articles from online databases to produce a simplified summary of technical abstracts and used the extracted data to train an encoder-decoder model for text summarisation [25]. Wang et al. attempted to train a recurrent neural network encoder-decoder model to perform neural machine translation [26]. Although both the statistical and neural network models possess considerable potential for biological text simplification and summarisation, the lack of availability of training corpora continues to pose a challenge for these approaches. Related to biological text simplification, studies have been conducted in relation to the usefulness of dependency parsing [27–29]. Junagadh et al. proposed bioSimplify, a method that uses dependency-relation classification among nodes to identify noun phrases in a sentence and normalise named entities appearing in a text [20]. After named entity normalisation, bioSimplify attempts to further reduce the structural complexity by splitting sentences by commas to identify independent clauses. The method can lose important biological entity relational context by consuming interaction definitions embedded in noun phrases. For example, the noun phrase "Spo0A-dependent spoIIG operon promoter" when replaced with the entity tag GENE1 loses the information about the causal relation between entities Spo0A and spollG. Percha et al. attempted a biological entity relation extraction using a Stanford dependency parser for capturing nodes appearing in a dependency path between two entities in a text [30]. Such simplification, by ignoring relevant nodes that do not appear in the shortest dependency path between the entities, may alter the meaning of the sentence due to the loss of relevant information. Also, solely relying on a parser dependency node selection process for accurate relation extraction is insufficient because it lacks the semantic and contextual understanding needed to handle

ambiguities and complex relations in text. The use of robust large language models (LLMs) combined with dependency parsing for sentence simplification has the potential to improve the accuracy of relation extraction, particularly in complex domains like biology.

In this work, we deploy transformer-based parser dependency mapping to produce biological sentence simplification for improved relation extraction and minimum loss of information. We propose a novel Knowledge-based Intelligent Text Simplification (KITS) method that performs informed detection of segments within a sentence relevant to the relation under examination and evaluates the simplified sentence to ensure the preservation of its original meaning. KITS distinguishes itself not merely by utilising parser dependency graphs but by innovatively employing them for sentence simplification. In contrast to conventional methods where text simplification occurs as a pre-processing step before named entity recognition, our approach harnesses named entity recognition prior to the simplification process. Additionally, for dependency parsing, we leverage spaCy's advanced transformer-based parser, introduced in 2020, capitalizing on the benefits offered by transformers. Unlike most approaches that rely solely on the shortest dependency path between entities for text simplification, our method takes a different approach by considering both the parent and child nodes of the entity. Visually, this means not only accounting for the shortest dependency path between entities, composed of their respective parent nodes, but also including the neighbouring dependent nodes. Depending solely on the shortest dependency path proves inadequate in addressing the intricacies of text simplification, particularly in a biological context. By incorporating both parent and child nodes of entities, our method enhances the ability to precisely capture the relational context of entity interactions. Additionally, certain methods have employed the shortest dependency paths for relation extraction [30]. However, dependency parsing, while offering insights into word relationships, may fall short in relation extraction due to its inherent lack of semantic understanding. Our proposed KITS implementation improves relation prediction by integrating sophisticated relation classifier models. To our knowledge, the majority of biological text simplification methods have not incorporated an assessment of potential information loss resulting from the removal of significant portions of content. These approaches have typically been applied uniformly to all sentences for simplification, rather than employing a selective approach. KITS conducts a selective text simplification using a controlling function using the positional distance between each entity's simplified sequential set of words to identify the reliability of simplified text.

The proposed method was used in conjunction with a Decision Tree Sequence (DTC) classification model and with a BioBERT model for experimentation. Experiments with the DTC produced a 14% improvement in precision, with only 42% of sentences simplified, compared to relation classification using full sentences. With BioBERT, our proposed approach reported F-scores of 87.67% and 88.67% using the benchmark gene/protein interaction corpora LLL [16] and HPRD50 [31]. This approach consistently demonstrates improvements in the accuracy of relation extraction models across various datasets and models. Our proposed method can be extended to other domains with similar text characteristics, such as the presence of complex terminologies, structural intricacies, lengthy sentences, and ambiguous relational contexts.

The remainder of the paper is organised as follows: Section 2 gives a brief description of dependency parsing and relation extraction. Section 3 details our proposed contributions. Section 4 discusses the experiments and results. Finally, Section 5 presents the conclusion and future directions.

2. Preliminaries

2.1. Relation Extraction

In general, a relation between two entities can be lexical, negation, coreference, or semantic [32]. Biological relation extraction algorithms attempt to accurately identify semantic relationships among various entities [17]. The simplest form of relation extraction (RE) is defined as $R{:=}\, r(e_1, e_2, e_{3\ldots}, e_n)$ where r is the relation classification among n entities

(e_i). The nested form of RE, a more complicated form, is defined as $R := r(s_1, s_2, s_3 \ldots, s_n)$, where r is the relation type and s can be either a simple form or a nested form of RE. While a binary relation between two biological entities deploys a binary classification to predict their semantic relationship, the multiclass relation estimates a semantic interaction and classifies its type [33]. For example, an interaction between two gene/protein entities can be either inhibitory or activatory. As shown in Figure 3 below, the simplest biological RE is a three-step process including (i) pre-processing; (ii) entity tagging; and (iii) relation extraction [34–36]. In the initial stage, pre-processing techniques are employed to refine the extensive biological corpus. This involves selecting and formatting the literature pieces relevant to the domain of study. Next, a Named Entity Recognition (NER) model scans the pre-processed set of biological sentences, diligently identifying named entities of interest. For example, in the sentence "AccB functions to negatively regulate transcription of the accBC operon", the NER model recognises AccB and accBC as gene/protein entities. These identified entities are replaced and labelled with placeholders like "GENE1" and "GENE2" for the purpose of anonymisation or to generalise the entities during analysis. These replacement labels ensure the focus is on the relationship between entities. After labelling the identified entities, the next step in the relation extraction process is relation classification. Relation classification involves determining the specific type or category of relationship that exists between the labelled entities. This step typically employs machine-learning models or rule-based systems to assign a relationship label to the pair of entities, indicating the nature of their interaction or connection within the context of the text. the relation classification step would determine whether they are related through a specific biological relationship, such as regulates, interacts with, inhibits, or encodes.

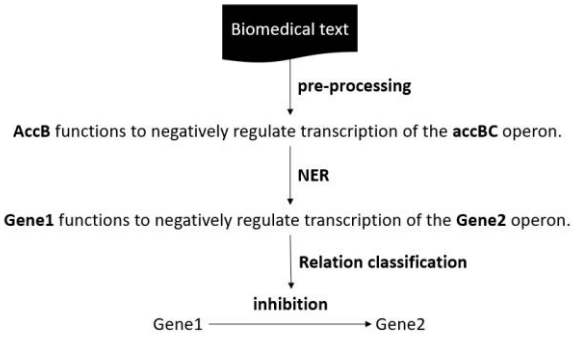

Figure 3. An example of biological relation extraction.

For this work, RE is accomplished using two well-known NLP relation classification models: the Decision Tree Classifier (DTC) and Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT). The Decision Tree Classifier, a supervised machine-learning technique, is well studied in the field of biological relation extraction. DTCs are straightforward, efficient, and easy to implement [37]. They exhibit robust performance in scenarios with sparse data and demonstrate resilience against overfitting on small datasets. However, the current limitation of DTCs lies in handling the complexity of relations, particularly in biomedical terms. BioBERT, built upon BERT's transformer architecture with a focus on biomedical and clinical text processing, has gained widespread popularity in the field of biological relation extraction [38]. However, BioBERT faces challenges not only in dealing with the intricate structural nature of biomedical relations but also the associated computational burden. For NER, we used a pre-trained transformer model, Advanced Biological Entity Recognition and Normalisation (BERN2) [39]. BERN2 supports NER and Named Entity Normalisation (NEN) by deploying a single multi-task pre-trained model to identify nine biological entities including species, genes and proteins, DNA, diseases, and chemicals. BERN2 reported an F-score of 83.7 when identifying gene/protein entities from the BC2GM dataset [38].

2.2. Dependency Parsing

A parsing dependency tree (z) represents the syntactic dependencies, such as subjects, objects, roots and modifiers, of words in an input sequence sentence x of length n, i.e., $x = \{x_1, x_2, x_3, \ldots\ldots, i_n\}$ where x_i is the i^{th} word in the sequence [40]. A dependency tree comprises directed edges and nodes (Figure 4). The nodes represent words and the edges represent grammatical dependencies among the words. An edge is directed from a parent node (also known as a head or controlling word) to a child (or dependent) node. Extensive research has been conducted into parsing techniques and their usefulness [41]. Recent advancements in data-driven machine- and deep-learning methods have increased interest in research into grammatical dependency.

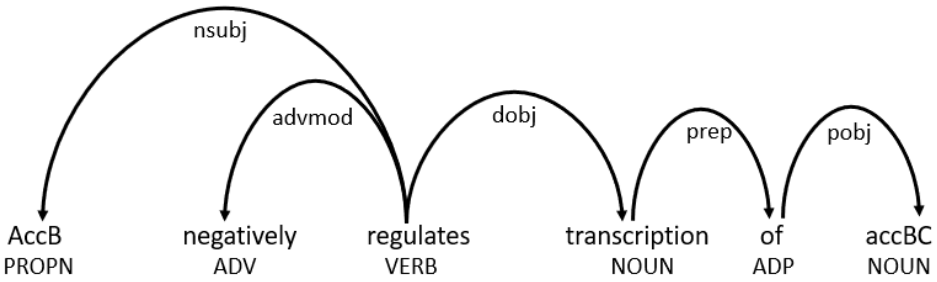

Figure 4. Parser dependency tree representation of a sentence using spaCy's pretrained linguistic model: en_core_web_trf. nsubj: nominal subject, asvmod: adverbial modifier, dobj: direct object, prep: prepositional modifier; pobj: object of preposition; PROPN: proper noun, ADV: adverb, ADP: adposition.

Several libraries and tools are available to facilitate parsing dependencies, including spaCy, Natural Language Toolkit (NLTK) [42] and Stanford CoreNLP [43]. Stanford CoreNLP offers a comprehensive and robust combination of rule-based and statistical NLP modelling pipelines used for a range of biological text dependency parsing tasks [44]. spaCy offers an efficient, easy to use, pre-trained, and customizable NLP pipeline, and is widely used for dependency parsing [45].

spaCy is an NLP library in Python that provides efficient tools for various tasks, including dependency parsing. Among the array of pre-trained models in spaCy dedicated to dependency parsing, en_core_web_trf stands out. This particular model adopts a transformer-based approach, leveraging the BERT architecture to enhance its understanding of English. The introduction of the transformer-based pipeline in spaCy v3.0 marked a significant advancement, resulting in a 3% increase in overall accuracy compared to its non-transformer counterpart, en_core_web_lg [46]. The parser pipeline achieved an accuracy of 95.1% on the OntoNotes 5.0 corpus. The en_core_web_trf dependency parsing pipeline has found application in various research studies for the initial reduction in irrelevant information [47–49]. This utilisation helped to reduce the model's processing time when dealing with extensive datasets and is concurrently leading to improvements in overall accuracies. Hence, harnessing the capability to handle complex linguistic structures afforded by a transformer serving as the foundation of the pipeline, and considering its successful applications for tasks in other domains, we utilised spaCy's en_core_web_trf parser dependency pipeline in this study.

3. Materials and Methods

In this section, we describe the processes carried out for the proposed biological relation extraction using an informed sentence simplification technique and its inherent self-evaluation in detail. Unlike many existing information extraction algorithms that simplify sentences as part of a pre-processing step (Figure 3), KITS integrates sentence simplification directly with NER outputs. This integration allows for a more targeted and contextually relevant simplification, as it considers the specific entities and relationships being analysed. One of the key distinctions of KITS in the realm of relation extraction and

text simplification is its incorporation of self-evaluation. Unlike many traditional methods that focus solely on simplifying sentences or extracting relations, KITS goes a step further by systematically assessing the quality of the simplified sentences, improving the reliability and accuracy of biological relation extraction.

As depicted by Figure 5, KITS includes three processes: (i) NER, (ii) sentence simplification, and the (iii) controlling function-based evaluation of simplified sentences. KITS uses labelled entities to select dependency nodes, constructing the simplified sentence to precisely represent the contextual relationship between these labelled entities within the sentence. Therefore, each simplified sentence varies depending on the specific labelled entity. The controlling evaluation of simplified sentences represents the potential deviation in meaning from the original sentence, with a higher value suggesting a greater likelihood of meaning change. Simplified sentences with evaluated values exceeding a predefined threshold are rejected, and the original sentence is retained for relation extraction instead. The different processes involved in KITS are discussed in following section.

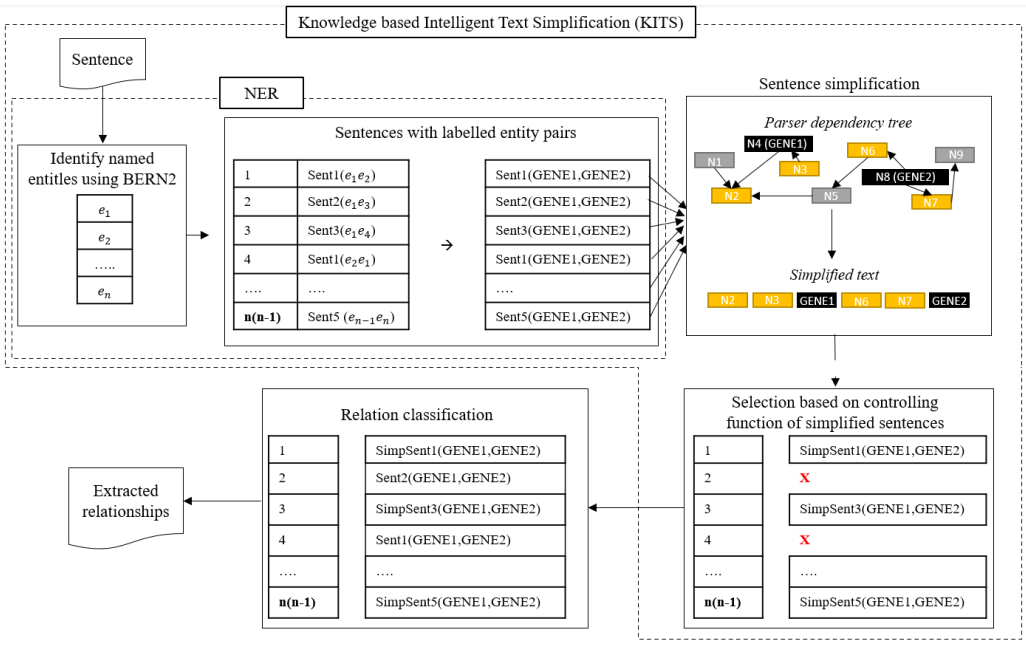

Figure 5. Framework of biological relation extraction using Knowledge-based Intelligent Text Simplification (KITS). KITS comprises three steps: named entity recognition (NER), sentence simplification, and selection using controlling function evaluation. The red 'x' represents the rejected simplified sentence.

3.1. Named Entity Recognition (NER)

The proposed framework starts with tagging gene/protein pairs in sentences. As we are using BERN2, its NEN and NER capability identifies not only a single-word entity but also a group of words that are directly related to the named entity. For instance, consider the following sentence:

> "A protein initially called Hst23 was identified as a product of the yvyD gene of *Bacillus subtilis*"

Here, instead of identifying *yvyD* as the gene/protein entity, BERN2 will suggest the phrase "yvyD gene of *Bacillus subtilis*" as the named entity. Applying normalisation ability of BERN2, we are able to reduce structural complexity without compromising the lexical integrity of the sentence. To represent the agent and target units, we replace the NER-identified entities (either single words or phrases) with replacement labels GENE1 and GENE2. Once the tagging of the set of sentences with all entities is completed, the independent clauses present in a sentence are identified. The sentences are split by a semicolon (;) which connects the related but independent clauses without a conjugational relation in a sentence. In such sentences, entities involved in a causal relationship are

unlikely to appear in separate independent clauses. It may be noted that prior to sentence simplification, the set of entity-tagged sentences are pre-processed to identify sentences containing independent clauses. We consider clauses containing both tagged entities as the simplified version of the sentence. If the entities appear in separate independent clauses in a sentence, we assume that these are unlikely to dictate a functional link, and thus they are eliminated from further processing. As the pre-processing removes noisy data, it will thereby contribute to improving model accuracy. The NER procedure is presented in Algorithm A1 in Appendix A.

3.2. Text Simplification

To simplify the text without information loss, we leverage the spaCy's pre-trained English pipeline en_core_web_trf parser to exploit the importance of directly related parent and child nodes present in the dependency parser tree. Parent nodes directly related to child nodes are important for understanding the lexical significance of a child word in a sentence. To extract all controlling and dependent words relevant to tagged entities, and thus to accurately capture their causal interactions, we identify the parent nodes that contain a tagged entity as their dependent child. After isolating the nodes that are directly related to tagged entity words, the identified parent and child nodes are arranged as per their position in the full sentence sequence. This selected and sequentially arranged set of words is now the simplified version of the original complex sentence.

An example of our node selection process controlling entities in a sentence is given in Figure 6. The method simplified the sentence "Transcription of GENE2 was dependent on GENE1, and the mRNA was detectable from 2 h after the cessation of logarithmic growth (T2 of sporulation)" to "Transcription of GENE2 was dependent on GENE1", accurately capturing the functional link between the entities without losing relevant information.

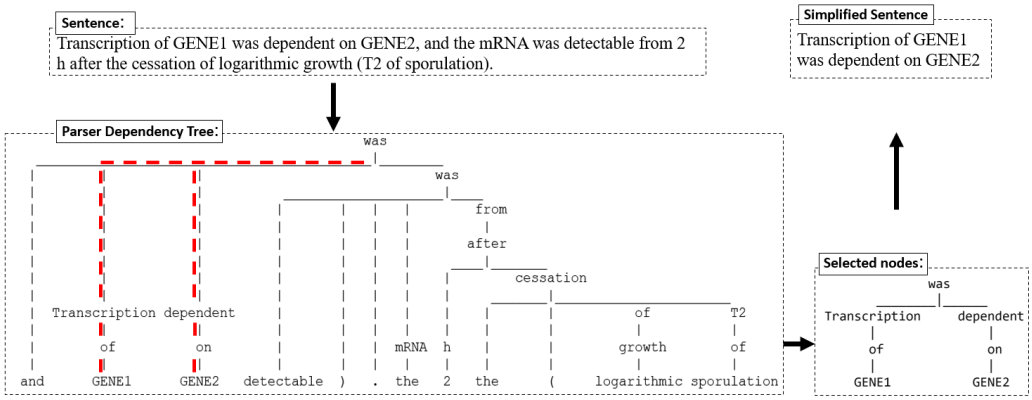

Figure 6. Text simplification process using a dependency parser tree illustrating the selection of nodes based on the labelled entities. The selection path for proposed text simplification is highlighted in red.

Another prominent issue to be considered is that the essential and non-essential elements of a sentence depend on the entities and relationships we wish to extract. A long sentence involving multiple entities may contain multiple interactions. The proposed technique is dependent on the positioning of the tagged entity. This awareness of positioning allows text simplification to capture the interaction context precisely. An example of text simplification of the same sentence with multiple entity pairs is given in Figure 7. The sentence "Both SigK and GerE were essential for ykvP expression, and this gene was transcribed from T5 of sporulation" has three gene entities: GerE, ykvP, and SigK. Of the three variations of the sentence with labelled entities given in Figure 7, labelled entities in sentence (i) and (ii) indicate a casual relation between GENE1 and GENE2, whereas sentence (iii) does not indicate a regulatory relation.

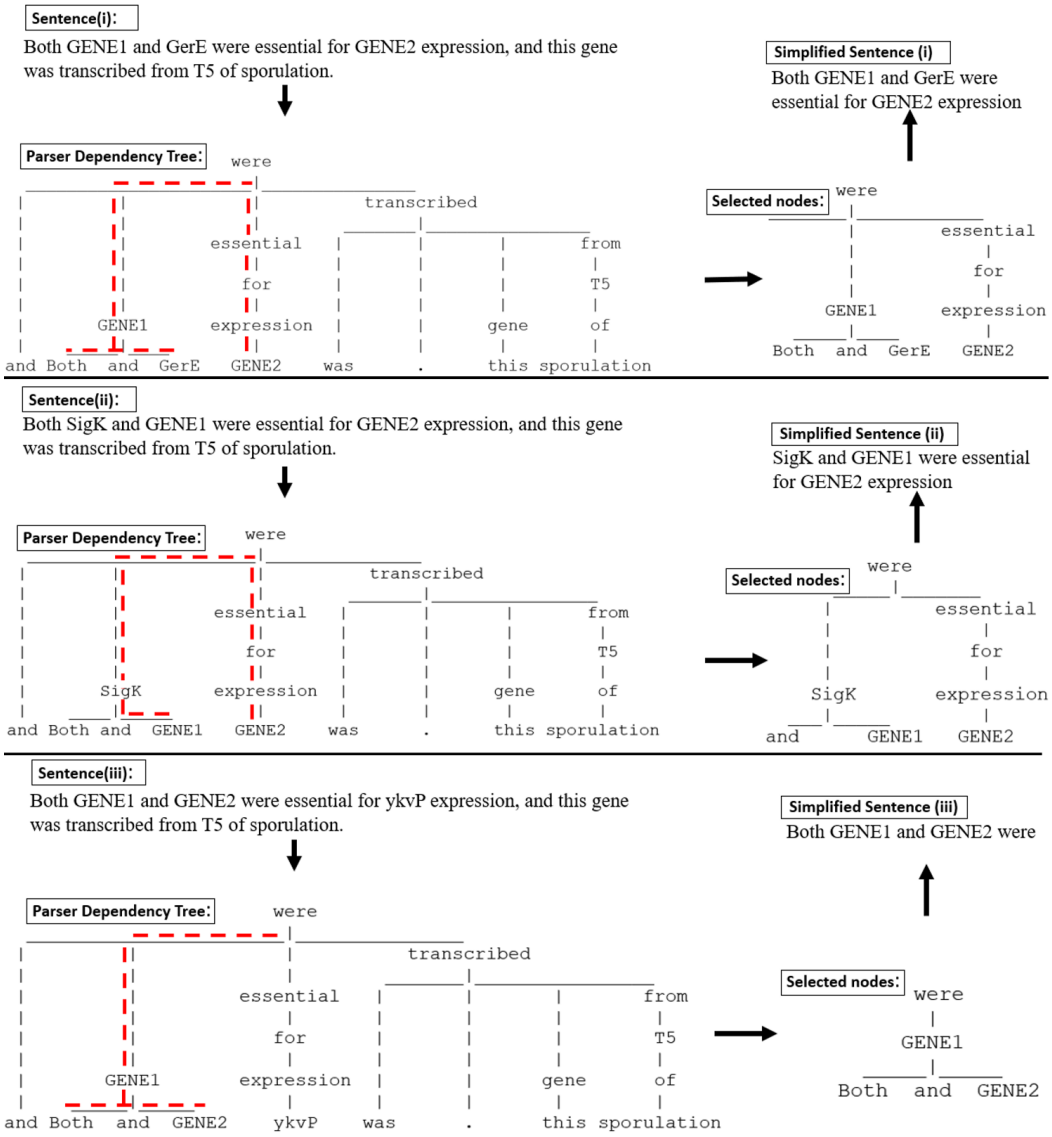

Figure 7. Text simplification of same sentence with three labelled entity relations, illustrating different simplified text generated based on the relation between labelled entities. The selection path for proposed text simplification is highlighted in red.

The parser dependency parent and dependent node selection method to simplify biological sentences was able to capture the functional interaction between GENE1 and GENE2 in sentences (i) and (ii). It was also able to capture the conjunction/compound relation between GENE 1 and GENE2 in sentence (iii). The text simplification procedure is presented in Algorithm A2 in Appendix A.

3.3. The Controlling Function for Simplified Sentences

As depicted in Figure 7, the simplified text more accurately represents meaningful relationships between entities when such relationships exist in the original sentence (Sentences (i) and (ii)). However, in the absence of any relationship between entities, the simplified sentence is not meaningful (Sentence (iii)). This is mainly attributed to the emphasis on capturing dependent and causal relations, which are non-existent in cases where there is no such relationship. This phenomenon is effective for the relationship classification model, facilitating the identification of patterns in non-meaningful sentences and the absence of relationships. However, when dealing with longer sentences in which entities are widely separated without a direct relationship, the simplification of such complex sentences can

inadvertently alter the meaning. This alteration occurs because the technique may simplify the sentence into a phrase that conveys a relationship due to the omission of relevant nodes. To mitigate this effect, we introduce a controlling function that considers the potential omission of vital information from the sentence, which could lead to modifications. The controlling function (f) between entities GENE1 and GENE2 (represented by e_1 and e_2), is calculated as follows:

$$f_{e_1e_2} = \left| i_{l_{e_1}} - i_{f_{e_2}} \right| \quad (1)$$

where i is the index position of the word in the original sentence, l_{e_1} is the last word of selected nodes that are related to e_1 (GENE1), and f_{e_2} is the first word of selected nodes that are related to e_2 (GENE2) and $f_{e_1e_2}$ represents the distance between the two independent phrases comprising the selected nodes for each entity. As illustrated in Figure 8, the simplification of the sentence "A low concentration of GerE activated cotB transcription by final GENE1 RNA polymerase, whereas a higher concentration was needed to activate transcription of cotX or GENE2" results in the simplified form "activated transcription by GENE1 polymerase needed activate transcription of cotX GENE2". This simplification erroneously suggests a functional or dependent relationship between the activated transcription action of GENE1 and the active transcription of GENE2. In reality, the original sentence indicates no direct relationship between GENE1 and GENE2. The simplification technique, in cases where there is no direct relationship, may miss important nodes and generate sentences that imply a relationship that does not exist.

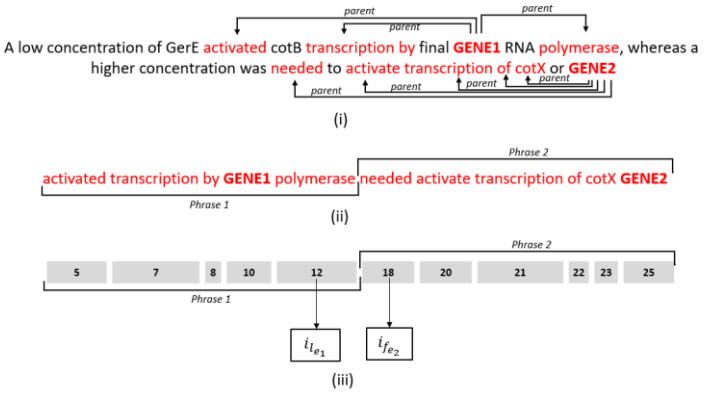

Figure 8. (**i**) Illustrates the chosen parents/dependent nodes for each entity (GENE1 and GENE2) for text simplification using the proposed method in red (there were no entity-dependent nodes for the sentence). (**ii**) Provides the simplified version of the sentence, emphasising the two distinct phrases created from the selected nodes related to GENE1 and GENE2, represented as Phrase1 and Phrase2; and (**iii**) indicates the index position of each selected node in each phrase as per the original sentence, including the indexes representing $i_{l_{e_1}}$ and $i_{f_{e_2}}$ in Equation (1).

Simplified text with an f value above a threshold ($t_\varnothing$) is more likely to alter the context of the interaction. Thus, we reject the simplified sentences with $f > t_\varnothing$. The controlling function evaluation procedure for simplified sentences is presented in Algorithm A3 in Appendix A.

KITS allows the parser dependency entity-controlling nodes to simplify complex biological sentences. The post-NER simplification technique attempts to accurately capture the precise regulatory interactions between tagged entities for sentences with multiple causal links. The simplified sentences can improve the prediction accuracy of the relation classification model. The experiments and results of relation extraction from established PPI corpora using the proposed method are discussed in the following section.

4. Experiments and Results

4.1. Datasets

For the experimentation, we used three well-known gene/protein interaction benchmark corpora: LLL [16], HPRD50 [31] and BioInfer [50], with their special features given in Table 1. The dataset LLL contained *Bacillus subtilis* gene interactions made publicly available during the Learning Language in Logic 2005 challenge. The dataset HPRD50 was created from 50 Human Protein Reference Database (HPRD)-referenced abstracts. The BioInfer corpus, which is the largest of the four, was created from abstracts referenced by the Database of Interacting Proteins as containing at least one interacting protein pair. As BioInfer is highly imbalanced, we selected 1000 positive and 1500 negative sentences for training and the rest for validation.

Table 1. Characteristics of the PPI datasets.

Dataset	Positive	Negative	Unique Sentences
BioInfer	2534	7132	1100
HPRD50	163	270	145
LLL	164	166	77

4.2. Experimental Setup

Three well known performance metrics were used for model comparison: recall (R) and specificity (S_p) measure the rate of true positives and true negatives, respectively, whereas precision (P) defines the accuracy of the model for the prediction of true positives. The F-score (F) is commonly used to measure the overall model accuracy, balancing uneven distributions.

The sentences unique to each of the three corpora provided in Table 1 were subjected to named entity recognition (NER) using BERN2, utilising the RESTful API web interface accessible at http://bern2.korea.ac.kr/ (accessed on 2 October 2023). Dependency parsing was performed using the transformer-based parser en_core_web_trf, which is part of spaCy's pipelines. The input data for both dependency parsing and the relation classification model consists of sequences (sentences) in their original and simplified forms, respectively. These sequences contain generic labelled entities, emphasising the entity pair under consideration.

To assess the impact of the proposed text simplification technique on enhancing relation extraction accuracy, we conducted experiments using two distinct models: (i) a straightforward, computationally lighter, and easily understandable Decision Tree Classifier (DTC), and (ii) a complex, sophisticated, pretrained, and computationally expensive model, BioBERT. For the experiment with DTC, tokenisation and classification were performed using the Scikit-learn's countVectorizer and DecisionTreeClassifier packages in Python. For the experiment with BioBERT, we deployed the relation classification model in Google Notebook with a Pytorch framework. The pretrained BioBERT model *biobert_v1.1_pubmed* with 768 dimensions of embedding vectors had the maximum token length to 128. For optimisation, we used the BertAdam optimiser with settings of 2×10^{-5} for learning rate, 0.1 as the warmup rate and learning rate decay (weight decay rate) set to 0.01. The model ran for 10 epochs with a batch size of 8. To evaluate the influence of the proposed KITS on the relation extraction methods, we selected classic baseline models incorporating parser dependency in their techniques for comparison using 10-fold validation with the *KFold* library from Scikit-learn, a widely used technique in machine learning for evaluating a model's performance and generalisation ability. The chosen machine-learning methods for assessing KITS performance with DTC included the following.

ASM (Approximate Subgraph Matching) and APG (All Path Graph kernel) as studied by Panyam et al. [51], explore various graph kernels in conjunction with parser dependency for enhanced biomedical relation extraction.

PIPE, introduced by Zhang et al. [52], is a module that extracts protein–protein interaction passages using an Interaction Pattern Generation structure to capture comprehensive information, involving pruning sentences through middle clause removal with parser dependency paths.

The selected neural network methods for assessing KITS performance with BioBERT included the following.

DNN [53], proposed by Zhao et al., is a protein–protein interaction extraction method utilising a deep neural network with greedy layer-wise unsupervised learning for parameter initialisation, enabling the model to learn intricate features from unlabelled data, thereby improving performance and considering the shortest path between entities during feature extraction.

Zhang et al. presented **RNN + CNN** [54], a hybrid model for biomedical relation extraction, combining Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to learn features from sentence and dependency sequences, and utilising the shortest dependency path (SDP) to detect and extract biomedical relations.

Ahmed et al.'s "**iLSTM + Attn**" model [55] is an approach for identifying protein–protein interactions that employs Structured Attention and LSTM to independently generate dependency structure information, learning the structure through the model without direct access to the actual dependency tree.

PIPE has been included in the results' comparison with the KITS+BioBERT evaluation due to its high recall score for the HPRD50 dataset.

4.3. The Experiment to Determine Threshold Value ($t_{\varnothing}$)

To obtain most suitable value for the controlling evaluations threshold $t_{\varnothing}$, we conducted several experiments with different values of $t_{\varnothing}$. Table 2 presents experimental results with the LLL, HPRD50, and BioInfer datasets, exploring the impact of the threshold value ($t_{\varnothing}$) on the performance of the relation extraction model using 10-fold cross-validation.

Table 2. Tenfold cross-validation results using a Decision Tree classifier (DTC) (in %) on 3 PPI corpora. P: Precision, F: F-score and KITS: Knowledge-based Intelligent Text Simplification. Best results are highlighted in bold.

Dataset	**LLL**				**HPRD50**				**BioInfer**			
Model	DTC		BioBERT		DTC		BioBERT		DTC		BioBERT	
Threshold	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*
$t_{\varnothing} = 2$	68.23	61.35	79.34	78.88	-	-	-	-	-	-	-	-
$t_{\varnothing} = 3$	75.57	76.23	81.23	82.43	-	-	-	-	-	-	-	-
$t_{\varnothing} = 4$	79.89	72.45	84.56	83.24	57.27	54.51	72.41	72.35	74.23	61.32	**78.76**	65.78
$t_{\varnothing} = 5$	**82.96**	**79.87**	**86.37**	**87.67**	**64.76**	**68.11**	**86.28**	**88.02**	76.10	**65.98**	77.65	**73.16**
$t_{\varnothing} = 6$	81.21	76.45	85.67	86.21	64.76	68.11	84.54	87.21	**78.23**	63.04	71.54	72.42
$t_{\varnothing} = 7$	72.34	68.11	79.96	81.31	-	-	-	-	-	-	-	-
$t_{\varnothing} = 8$	67.45	64.98	75.43	74.56	-	-	-	-	-	-	-	-

LLL, the smallest dataset with only 77 sentences, was chosen for a thorough investigation of threshold values. The optimal threshold value identified in LLL was then validated in the larger datasets, HPRD50 and BioInfer. The model achieved its maximum f-score at $t_{\varnothing}$= 5 for all datasets, where the performance of threshold value 5 exhibited greater consistency than other values. The setting $t_{\varnothing}$ = 5 was chosen as it produced the best results.

4.4. Results

Among the 77 sentences in the LLL dataset, 3 sentences contained phrases separated by semicolons. The first sentence had 21 relationships between two phrases, the second sentence had 36 relationships among five phrases, and the third sentence had 1 relationship with two phrases, a total of 58 relationships across the three sentences. We split these sen-

tences by semicolons and treated them as individual sentences for relation extraction. This splitting reduced the relationship count to 16, eliminating 42 relations in which the entities were not present in the same phrase. This step not only reduced the number of sentences for relation extraction but also eliminated false positives. All 42 eliminated relations were true negatives. In the HPRD50 dataset, there was only one sentence with two phrases separated by a semicolon and one relationship, which was present in one phrase. The phrase not containing either of the two entities was eliminated. BioInfer contained 30 sentences with phrases separated by semicolons, containing 222 interactions. The largest sentences included seven phrases and 13 relationships. Upon splitting, the relationship count was reduced to 88, eliminating 134 relationships as entities were placed in separate phrases. Of these 134 eliminated relationships, 8 were false negatives. Our method, designed to identify relationships where entities are situated in separate phrases separated by semicolons, demonstrated an accuracy of 95.48% in identifying true negatives. The remaining relationships and their corresponding sentences underwent further processing for text simplification and relation extraction.

Table 3 shows the overall performance of our model on the three PPI corpora compared to various baseline RE methods that employ statistical and machine-learning techniques, with and without sentence simplification, in terms of recall, precision, and F-score using a Decision Tree Classifier. While our decision classifier model on its own could not show major improvement compared to the previous statistical and machine-learning approaches, KITS, when combined with the Decision Tree Classifier, exhibited a significantly improved prediction accuracy. The use of the Decision Tree with our text simplification method (S_{sim}) produced a 20.95%, 13.47%, and 9.9% increase in precision for LLL, HPRD50 and BioInfer, respectively.

Table 3. Tenfold cross-validation results using Decision Tree Classifier (DTC) (in %) on 3 PPI corpora. P: Precision, R: Recall, F: F-score and KITS: Knowledge-based Intelligent Text Simplification. Best results are highlighted in bold.

Dataset	**BioInfer**			**HPRD50**			**LLL**		
Method	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
ASM	67.20	22.60	33.80	**66.00**	58.30	61.90	79.3	28.00	41.4
APG	68.60	28.60	40.40	62.30	69.90	65.90	**84.70**	57.30	68.30
PIPE	57.60	59.90	58.70	62.50	**83.30**	**71.40**	73.20	**89.60**	**80.60**
DTC (w/o KITS)	66.20	62.01	64.25	50.77	64.24	59.35	62.01	66.76	64.30
DTC (w KITS)	**76.10**	**64.96**	**65.98**	64.76	78.13	68.11	82.96	76.74	79.87

This method surpassed other machine-learning approaches on BioInfer, the largest dataset among the three, showcasing the model's capacity to effectively manage larger and more diverse datasets. This underscores the model's robustness and generalisability, affirming its suitability for tasks involving extensive and varied data. While PIPE exhibited higher F-score and recall, it fell short in precision compared to our model. The combination of a high F-score and recall with a low precision is often linked to low specificity. In biological relation extraction, where non-relations constitute the dominant class, accurately identifying the absence of interactions is crucial. Our model demonstrates this ability, as evidenced by its improved performance on the larger and imbalanced BioInfer dataset. These experiments show that the proposed KITS, used with a Decision Tree model, has a performance comparable to the advanced and more complex existing state-of-the-art RE techniques.

The performance of the method using a fine-tuned BioBERT sequence classifier in contrast to several baseline RE methods that leverage deep-learning techniques, for both with and without the proposed sentence simplification, is given in Table 4. The results from our simplified set of sentences produced a substantial improvement in classification performance compared to using the original full sentences. We note that the BioBERT's prior

training within the biological context makes it a more accurate model for biological relation extraction compared to a Decision Tree Classifier. However, there may be a possibility that sentences with complex terminology and multiple entities can shift the focus from the tagged entities, affecting BioBERT's ability to correctly capture the interactional context solely related to the entities in question. With only 25% of full sentences simplified for LLL, BioBERT achieved a 15.57% increase in precision and a 6.93% increase in recall. We compared our results with previous deep-machine-learning methods. The proposed simplification method with BioBERT outperformed existing state-of-the-art approaches in RE precision and F-score. The interactive pattern generation module PIPE achieved a high recall score with HPRD50. However, PIPE's low precision score indicates a high false positive rate, thereby impairing its overall prediction ability. Our method has a higher precision and F-score, indicating our model's ability to capture most of the actual positive instances, minimizing false negatives. For BioInfer, we recorded a recall score of 85.66, a 13% improvement over DNN. On all three datasets, the difference in precision and recall of BioInfer is apparent. This difference can be attributed to almost three-times the number of negative classifications as compared to the number of positive sentences. BioInfer was the largest and most unbalanced dataset of the three PPI datasets under consideration, and therefore most methods struggled to achieve high accuracies with this dataset.

Table 4. Tenfold cross-validation results using BioBERT (in %) on 3 PPI corpora. P: Precision, R: Recall, F: F-score and KITS: Knowledge-based Intelligent Text Simplification. Best results are highlighted in bold.

Dataset	**BioInfer**			**HPRD50**			**LLL**		
Method	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
PIPE	57.60	59.90	58.70	62.50	**83.30**	71.40	73.20	89.60	80.60
DNN	53.90	72.90	61.60	58.70	92.40	71.30	76.00	91.00	81.40
RNN + CNN	56.70	67.30	61.30	69.60	82.70	75.10	72.50	87.20	76.50
iLSTM + tAttn	61.80	54.20	57.60	78.60	78.70	78.50	84.80	84.30	84.20
BioBERT (w/o KITS)	70.14	79.25	74.31	76.45	80.36	75.24	70.80	84.23	83.81
BioBERT (w KITS)	**73.16**	**85.66**	**77.65**	**86.28**	81.43	**88.02**	**86.37**	**91.16**	**87.67**

The number of sentences successfully simplified for each data set is given in Table 5. Although our sentence simplification technique resulted in the improved performance of statistical and deep-learning models, the number of successfully simplified sentences was limited. Our strict threshold ($t_{\varnothing} = 5$) allowed 25%, 42%, and 37% of sentence simplifications to replace original full sentences for the relation extraction for LLL, HPRD50, and BioInfer. Certain sentences within the datasets were already in their optimal simplified form. In these instances, we observed that the text simplification process preserved all words from the original sentences, resulting in the simplified sentence being identical to the original. This phenomenon likely occurs when all words in these sentences are relevant to the context of the relationship. LLL, HPRD50, and BioInfer contained 12, 6, and 95 such sentences, respectively.

Table 5. Number of sentences simplified with proposed method for each dataset.

Dataset	**Number of Sentences Successfully Simplified**
LLL	84
HPRD50	185
BioInfer	3566

Table 6 presents the performance results for both DTC and BioBERT for the successfully simplified set of sentences among the three PPI corpora, with and without the proposed KITS. There was a consistent and substantial improvement in performance across

all datasets and models. This finding emphasises the importance of simplifying the text to retain pertinent information without altering its meaning, leading to a significant enhancement in the predictive capacity of the relation classification model. Specificity (true negative rate) exhibited the most substantial improvement of all measures, indicating the effectiveness of KITS in assisting models to accurately identify non-relations. In the context of biological relation extraction, this is a crucial aspect as non-interactions are the dominant classification.

Table 6. Tenfold cross-validation results using DTC and BioBERT (in %) for the successfully simplified sentences among three PPI corpora. P: Precision, R: Recall, Sf: Specificity, F: F-score and KITS: Knowledge-based Intelligent Text Simplification. Best results are highlighted in bold.

Dataset	**BioInfer**				**HPRD50**				**LLL**			
Method	*P*	*R*	S_f	*F*	*P*	*R*	S_f	*F*	*P*	*R*	S_f	*F*
DTC (w/o KITS)	53.37	46.01	68.25	50.74	53.75	66.71	72.82	50.47	72.02	71.76	47.23	70.23
DTC (w KITS)	**75.30**	**65.38**	**82.47**	**68.49**	**69.49**	**68.71**	**87.16**	**59.46**	**86.96**	**82.47**	**91.18**	**85.67**
BioBERT (w/o KITS)	73.67	71.76	78.43	77.84	79.24	74.85	83.75	76.78	79.21	83.54	78.91	80.45
BioBERT (w KITS)	**86.31**	**89.67**	**91.57**	**84.76**	**84.32**	**89.81**	**94.69**	**87.34**	**91.92**	**94.69**	**97.23**	**93.45**

4.5. Error Analysis

In this section, we examine our model's error analysis on the BioInfer, HPRD50, and LLL corpora, summarised as follows.

- Indirect relationships and the presence of negative terms such as 'unable' or 'incapable' make it more difficult for the model to accurately identify and extract positive relationships between entities. For instance, in the sentence "However, the mutant was unable to stimulate transcription by final GENE2-RNA polymerase from the GENE1-dependent spoIIG operon promoter", the direct mention of the relationship between GENE1 and GENE2 is absent. Instead, the relationship between GENE1 and GENE2 is mediated through "the mutant" and "spollG". Also, the presence of the negative term 'unable' poses challenges for the model to accurately classify this relationship as true.
- SpaCy's 'en_core_web_trf' model may overlook the identification of all directly dependent nodes in certain cases. For example, the sentence "In this work, we show that GENE1 and GENE2 specifically interact with the Cdk1/CyclinB1 complex, but not with other Cdk/Cyclin complexes, in vitro and in vivo" was simplified to "show GENE1 and GENE2 interact", resulting in the omission of important directly related nodes like "Cdk1/CyclinB1 complex". This oversight could be attributed to entities being placed in a conjunctive form. To address this issue, an additional evaluation of conjunctive entity placement in sentences is necessary.
- While most phrases split by semicolons in the three datasets were independent clauses, the incorrectly rejected eight relations for BioInfer were from sentences in which the purpose of the semicolon was to separate complex items in a list. This issue can be mitigated by verifying the type of the phrase before elimination.
- Some sentences contain incorrect annotations. For instance, in the sentence "Quantitation of the appearance of X22 banding in primary cultures of myotubes indicates that it precedes that of other myofibrillar proteins and that assembly takes place in the following order GENE2 myosin heavy chain GENE1", the annotation depicts a positive relationship between GENE1 and GENE2. However, the sentence conveys a placement order of the entities without implying a causal relationship.

5. Conclusions

In this paper, we propose a novel text simplification method for improved biological relation extraction called Knowledge-based Intelligent Text Simplification (KITS). Unlike most sentence simplification methods, this technique is deployed after NER tagging, enabling the proposed model to retain the relevant labelled entity information needed for relation extraction. We leveraged a dependency parsing method to identify the dependent and controlling nodes of named entities for text simplification. Our method includes a novel controlling function evaluation measure to represent the ability of simplified text to retain the true context. The proposed method was tested on three PPI benchmark datasets and obtained improved performances. The experimental results of the proposed KITS with both BioBERT and DTC demonstrate the method's efficacy in enhancing the accuracy of both a basic and straightforward model as well as a sophisticated Large Language Model, showcasing its versatility and effectiveness across different complexity levels. Future work can focus on extending the proposed method to improve the performance of large-scale text mining frameworks. Further, studies can be performed on the introduction of semantic understanding of causal interactions embedded in a text to improve the text simplification of nested and indirect entity relationships.

Author Contributions: Conceptualisation, J.G., M.C., S.L. and J.H.; methodology, J.G. and M.C.; software, J.G.; validation J.G.; formal analysis, J.G., M.C., S.L. and J.H.; investigation, J.G. and M.C.; resources, J.G., M.C., S.L. and J.H.; data curation, J.G.; writing—original draft preparation J.G.; writing—review and editing, J.G., M.C., S.L. and J.H.; visualisation, J.G., M.C., S.L. and J.H.; supervision, M.C., S.L. and J.H.; project administration, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Github at https://github.com/BNLNLP/PPI-Relation-Extraction/tree/main/datasets/PPI/original (accessed on 6 February 2023) [56].

Acknowledgments: The first author acknowledges the support for the tuition fee waiver scholarship from Federation University and the stipend scholarship from Health Innovative and Transformation Centre (HITC), Federation University Australia.

Conflicts of Interest: Author J.H. is the Director of BioThink Pty Ltd. BioThink Pty Ltd. has no financial involvement in this research. Author J.H. is also an adjunct staff member within the Health Innovation and Transformation Centre at Federation University, Australia. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflict of interest.

Appendix A

In this appendix, we provide the pseudocode of three key algorithms used in the paper. Algorithm A1 relates to the identification of named entities, Algorithm A2 relates to text simplification and Algorithm A3 is for controlling the function evaluation of simplified sentences. Algorithms A1–A3 are used in Sections 3.1–3.3, respectively.

Algorithm A1: NamedEntityRecognition(S)

Input: **S = Set of sentences**
Output: **L = Refined NER Tagged Sentences**

L ← Initialise an empty List variable to contain refined tagged sentences of S_i
For *i* = 1 to S **Do**
    E_i ← Get all recognised gene/protein entity using BERN2
    P_i (A_i, T_i) ← Identify all possible pairs from E_i
    L_i ← Initialise an empty List variable to contain tagged variations of S_i
    For *p* = 1 to P_i **Do**
        S_{ip} ← Replace A_i with GENE1 and B_i with GENE2 in S_i
        If S_{ip} contains a semicolon **Do**
            S_{split} ← Split S_{ip} by delimitator
            S_{temp} = "" ← Initialise an empty String variable
            For ss = 1 to S_{split} **Do**
                If GENE1 and GENE2 in S_{ss} **Do**
                    $S_{temp} = S_{ss}$ ← Replace the tagged sentence with independent clause
                    Break loop
                End If
            End For
            If S_{temp} != "" **Do**
                $S_{ip} = S_{temp}$
            End If
        End If
        Append S_{ip} to L_i ← Add the refined tagged sentence to L_i
    End For
    Append L_i to L ← Add the refined tagged sentence for S_i to L
End For

Algorithm A2: TextSimplification(S)

Input: **S = Sentence**
Output: **S_{simp} = Simplified sentence**
P_{simp} = Index position of words in simplified sentence

nlp ← load *'en_core_web_sm'* from *spaCy*
Doc = nlp(S) ← Tokenise the sentences
S_{simp} ← Initialise an empty list variable to save words of the simplified sentence
P_{simp} ← Initialise an empty list variable to save position of words of the simplified sentence
For token in Doc **Do**
    S_c = Get the dependent nodes of token
    S_h = Get the parent nodes of token
    If "GENE1" in S_c or "GENE2" in S_c **Do**
        Append token to S_{simp}
        Append Position(token) to P_{simp}
    Else If "GENE1" in token or "GENE2" in token **Do**
        Append S_h to S_{simp}
        Append Position(S_h) to P_{simp}
    End If
End For
S_{simp} ← Rearrange S_{simp} as per their token position in S_{simp}

Algorithm A3: ControllingFunctionEvaluation(S_{simp}, P_{simp}, S)

Input: S_{simp} **= Simplified sentence**
P_{simp} **= Index position of words in simplified sentence**
S **= Original sentence**
Output: S_f **= Sentence used for relation classification**

Seq_{gene1} ← Identify phrase containing GENE1
Seq_{gene1} ← Identify phrase containing GENE2
If Seq_{gene1} not equal to Seq_{gene1} **then**
    Seq_{gene1_last} ← Get the position of last word in Seq_{gene1}
    Seq_{gene2_first} ← Get the position of first word in Seq_{gene2}
    If abs(Seq_{gene2_first}-Seq_{gene1_last}) **then**
        $S_f = S$
    End If
Else
    $S_f = S_{simp}$
End If

References

1. Naseem, U.; Khushi, M.; Khan, S.K.; Shaukat, K.; Moni, M.A. A Comparative Analysis of Active Learning for Biomedical Text Mining. *Appl. Syst. Innov.* **2021**, *4*, 23. [CrossRef]
2. Simon, C.; Davidsen, K.; Hansen, C.; Seymour, E.; Barnkob, M.B.; Olsen, L.R. BioReader: A text mining tool for performing classification of biomedical literature. *BMC Bioinform.* **2019**, *19*, 57. [CrossRef] [PubMed]
3. Gamage, H.N.; Chetty, M.; Shatte, A.; Hallinan, J. Ensemble Regression Modelling for Genetic Network Inference. In Proceedings of the 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Ottawa, ON, Canada, 15–17 August 2022.
4. Nair, A.; Chetty, M.; Wangikar, P.P. Improving gene regulatory network inference using network topology information. *Mol. BioSystems* **2015**, *11*, 2449–2463. [CrossRef] [PubMed]
5. Morshed, N.; Chetty, M.; Vinh, N.X. Simultaneous learning of instantaneous and time-delayed genetic interactions using novel information theoretic scoring technique. *BMC Syst. Biol.* **2012**, *6*, 62. [CrossRef]
6. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E.A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24*, 102155. [CrossRef] [PubMed]
7. Corlan, A.D. Medline Trend: Automated Yearly Statistics of PubMed Results for Any Query. Available online: http://dan.corlan.net/medline-trend.html (accessed on 14 February 2023).
8. Mercatellia, D.; Scalambra, L.; Triboli, L.; Ray, F.; Giorgi, F.M. Gene regulatory network inference resources: A practical overview. *Biochim. Et Biophys. Acta (BBA)-Gene Regul. Mech.* **2020**, *1863*, 194430. [CrossRef]
9. Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Sun, Y.; Xu, B.; Zhao, Z. Neural network-based approaches for biomedical relation classification: A review. *J. Biomed. Inform.* **2019**, *99*, 103294. [CrossRef]
10. BioCreative. BioCreative VI Challenge and Workshop. Available online: https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/biocreative-vi-challenge/ (accessed on 12 November 2023).
11. Peng, Y.; Rios, A.; Kavuluru, R.; Lu, Z. Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database J. Biol. Databases Curation* **2018**, *2018*, bay073. [CrossRef]
12. Wang, H.; Qin, K.; Zakari, R.Y.; Lu, G.; Yin, J. Deep neural network-based relation extraction: An overview. *Neural Comput. Appl.* **2022**, *34*, 4781–4801. [CrossRef]
13. Zhao, S.; Lu, C.S.Z.; Wang, F. Recent advances in biomedical literature mining. *Brief. Bioinform.* **2021**, *22*, bbaa057. [CrossRef]
14. Kilicoglu, H. Biomedical text mining for research rigor and integrity: Tasks, challenges, directions. *Brief. Bioinform.* **2018**, *19*, 1400–1414. [CrossRef]
15. Fleuren, W.W.; Alkema, W. Application of text mining in the biomedical domain. *Methods* **2015**, *75*, 97–106. [CrossRef] [PubMed]
16. Nédellec, C. Learning language in logic—Genic interaction extraction challenge. In Proceedings of the Learning Language in Logic Workshop (LLL05), Bonn, Germany, 1 April 2005.
17. Huang, C.-C.; Lu, Z. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Brief. Bioinform.* **2016**, *17*, 132–144. [CrossRef] [PubMed]
18. Singhal, A.; Leaman, R.; Catlett, N.; Lemberger, T.; McEntyre, J.; Polson, S.; Xenarios, I.; Arighi, C.; Lu, Z. Pressing needs of biomedical text mining in biocuration and beyond: Opportunities and challenges. *Database* **2016**, *2016*, baw161. [CrossRef] [PubMed]
19. Peng, Y.; Torii, M.; Wu, C.H.; Vijay-Shanker, K. A generalizable NLP framework for fast development of pattern-based biomedical relation extraction systems. *BMC Bioinform.* **2014**, *15*, 285. [CrossRef]

20. Jonnalagadda, S.; Tari, L.; Hakenberg, J.; Baral, C.; Gonzalez, G. Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. *arXiv* **2010**, arXiv:1001.4277.
21. Bach, N.; Gao, Q.; Vogel, S.; Waibel, A. TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 2 June 2011.
22. Hakenberg, J.; Leaman, R.; Vo, N.H.; Jonnalagadda, S.; Sullivan, R.; Miller, C.; Tari, L.; Baral, C.; Gonzalez, G. Efficient extraction of protein-protein interactions from Full-Text Articles. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2010**, *7*, 481–494. [CrossRef]
23. Miao, Q.; Zhang, S.; Zhang, B.; Yu, H. Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Bali, Indonesia, 7–10 November 2012.
24. Ondov, B.; Attal, K.; Demner-Fushman, D. A survey of automated methods for biomedical text simplification. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 976–1988. [CrossRef]
25. Devaraj, A.; Marshall, I.J.; Wallace, B.C.; Li, J.J. Paragraph-level Simplification of Medical Texts. In Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Online, 6–11 June 2021.
26. Wang, T.; Chen, P.; Rochford, J.; Qiang, J. Text Simplification Using Neural Machine Translation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
27. Siddharthan, A. Text Simplification using Typed Dependencies: A Comparison of the robustness of different generation strategies. In Proceedings of the 13th European Workshop on Natural Language Generation, Nancy, France, 28–31 September 2011.
28. Siddharthan, A. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In Proceedings of the 13th European Workshop on Natural Language Generation, Nancy, France, 28–31 September 2011.
29. Chatterjee, N.; Agarwal, R. DEPSYM: A Lightweight Syntactic Text Simplification Approach using Dependency Trees. In Proceedings of the CTTS@ SEPLN, Málaga, Spain, 21–24 September 2021.
30. Percha, B.; Altman, R.B. A global network of biomedical relationships derived from text. *Bioinformatics* **2018**, *34*, 2614–2624. [CrossRef]
31. Fundel, K.; Küffner, R.; Zimmer, R. RelEx—Relation extraction using dependency parse trees. *Bioinformatics* **2007**, *23*, 365–371. [CrossRef]
32. Zhou, D.; Zhong, D.; He, Y.I.S. Biomedical Relation Extraction: From Binary to Complex. *Comput. Math. Methods Med.* **2014**, *2014*, 298473. [CrossRef] [PubMed]
33. Yang, X.; Yu, Z.; Guo, Y.; Bian, J.; Wu, Y. Clinical Relation Extraction Using Transformer-based Models. *arXiv* **2021**, arXiv:2107.08957.
34. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
35. Goyal, A.; Gupta, V.; Kumar, M. Recent Named Entity Recognition and Classification techniques: A systematic review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]
36. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48. [CrossRef]
37. Raul Garreta, G.M.T.H.G.H. *Scikit-Learn: Machine Learning Simplified: Implement Scikit-Learn into Every Step of the Data Science Pipeline*; Packt Publishing Ltd.: Birmingham, UK, 2017.
38. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J.W.J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
39. Sung, M.; Jeong, M.; Choi, Y.; Kim, D.; Lee, J.; Kang, J. BERN2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **2022**, *38*, 4837–4839. [CrossRef]
40. Vacariu, A.V. A High-Throughput Dependency Parser. 2017. Available online: https://summit.sfu.ca/item/17739 (accessed on 4 September 2023).
41. Siddharthan, A. A survey of research on text simplification. *ITL-Int. J. Appl. Linguist.* **2014**, *165*, 259–298. [CrossRef]
42. Millstein, F. NLTK, Natural Language Processing with Python: Natural Language Processing Using. 2020. Available online: https://scholar.google.com.hk/scholar?hl=zh-TW&as_sdt=0,5&q=NLTK,+Natural+Language+Processing+with+Python:+Natural+Language+Processing+Using&btnG=#d=gs_cit&t=1702266004906&u=/scholar?q=info:Rrd7HVVyN8IJ:scholar.google.com/&output=cite&scirp=0&hl=zh-TW (accessed on 4 September 2023).
43. Nazaruka, E.; Osis, J.; Griberman, V. Using Stanford CoreNLP Capabilities for Semantic Information Extraction from Textual Descriptions. In Proceedings of the International Conference on Evaluation of Novel Approaches to Software Engineering, Heraklion, Greece, 4–5 May 2019.
44. Okhapkin, V.P.; Okhapkina, E.P.; Iskhakova, A.O.; Iskhakov, A.Y. Constructing of Semantically Dependent Patterns Based on SpaCy and StanfordNLP Libraries. In Proceedings of the Futuristic Trends in Network and Communication Technologies: Third International Conference, FTNCT 2020, Taganrog, Russia, 14–16 October 2020.
45. Vasiliev, Y. *Natural Language Processing with Python and spaCy: A Practical Introduction*; No Starch Press: San Francisco, CA, USA, 2020.
46. Honnibal, M.; Montani, I.; Landeghem, S.V.; Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. 2020. Available online: https://github.com/explosion/spaCy (accessed on 4 September 2023).

47. Ramesh, S.; Tiwari, A.; Choubey, P.; Kashyap, S.; Khose, S.; Lakara, K.; Singh, N.; Verma, U. BERT based Transformers lead the way in Extraction of Health Information from Social Media. In Proceedings of the Sixth Social Media Mining for Health Workshop, Mexico City, Mexico, 10 June 2021.
48. Algamdi, S.; Albanyan, A.; Shah, S.K.; Tariq, Z. Twitter Accounts Suggestion: Pipeline Technique SpaCy Entity Recognition. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022.
49. Kandji, A.K.; Ndiaye, S. Design and realization of an NLP application for the massive processing of large volumes of resumes. In Proceedings of the IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD), Bobo-Dioulasso, Burkina Faso, 23–25 February 2023.
50. Pyysalo, S.; Ginter, F.; Heimonen, J.; Björne, J.; Boberg, J.; Järvinen, J.; Salakoski, T. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinform.* **2007**, *8*, 50. [CrossRef]
51. Panyam, K.V.N.C.; Cohn, T.; Ramamohanarao, K. Exploiting graph kernels for high performance biomedical relation extraction. *J. Biomed. Semant.* **2018**, *9*, 7. [CrossRef]
52. Chang, Y.-C.; Chu, C.-H.; Su, Y.-C.; Chen, C.C.; Hsu, W.-L. PIPE: A protein-protein interaction passage extraction module for BioCreative challenge. *Database J. Biol. Databases Curation* **2016**, *2016*, 101. [CrossRef] [PubMed]
53. Zhang, H.; Guan, R.; Zhou, F.; Liang, Y.; Zhan, Z.-H.; Huang, L.; Feng, X. A protein-protein interaction extraction approach based on deep neural network. *IEEE Access* **2019**, *7*, 89354–89365. [CrossRef]
54. Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Zhang, S.; Sun, Y.; Yang, L. A hybrid model based on neural networks for biomedical relation. *J. Biomed. Inform.* **2018**, *81*, 83–92. [CrossRef] [PubMed]
55. Ahmed, M.; Islam, J.; Samee, M.R.; Mercer, R.E. Identifying Protein-Protein Interaction using Tree LSTM and Structured Attention. In Proceedings of the 2019 IEEE 13th international conference on semantic computing (ICSC), Newport Beach, CA, USA, 30 January–1 February 2019.
56. Park, G.; McCorkle, S.; Soto, C.; Blaby, I.; Yoo, S. Extracting Protein-Protein Interactions (PPIs) from Biomedical Literature using Attention-based Relational Context Infor-mation. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 2052–2061. [CrossRef]