information

Article

Leveraging Semantic Text Analysis to Improve the Performance of Transformer-Based Relation Extraction

Marie-Therese Charlotte Evans [1], Majid Latifi [2], Mominul Ahsan [2,*] and Julfikar Haider [3]

[1] Solution Consultant, IDHL Group, Central House, Otley Road, Harrogate HG3 1UF, UK; m.t.c.evans@outlook.com
[2] Department of Computer Science, University of York, Deramore Lane, York YO10 5GH, UK; majid.latifi@york.ac.uk
[3] Department of Engineering, Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester M1 5GD, UK; j.haider@mmu.ac.uk
* Correspondence: mominul.ahsan2@gmail.com

Abstract: Keyword extraction from Knowledge Bases underpins the definition of relevancy in Digital Library search systems. However, it is the pertinent task of Joint Relation Extraction, which populates the Knowledge Bases from which results are retrieved. Recent work focuses on fine-tuned, Pre-trained Transformers. Yet, F1 scores for scientific literature achieve just 53.2, versus 69 in the general domain. The research demonstrates the failure of existing work to evidence the rationale for optimisations to finetuned classifiers. In contrast, emerging research subjectively adopts the common belief that Natural Language Processing techniques fail to derive context and shared knowledge. In fact, global context and shared knowledge account for just 10.4% and 11.2% of total relation misclassifications, respectively. In this work, the novel employment of semantic text analysis presents objective challenges for the Transformer-based classification of Joint Relation Extraction. This is the first known work to quantify that pipelined error propagation accounts for 45.3% of total relation misclassifications, the most poignant challenge in this domain. More specifically, Part-of-Speech tagging highlights the misclassification of complex noun phrases, accounting for 25.47% of relation misclassifications. Furthermore, this study identifies two limitations in the purported bidirectionality of the Bidirectional Encoder Representations from Transformers (BERT) Pre-trained Language Model. Firstly, there is a notable imbalance in the misclassification of right-to-left relations, which occurs at a rate double that of left-to-right relations. Additionally, a failure to recognise local context through determiners and prepositions contributes to 16.04% of misclassifications. Furthermore, it is highlighted that the annotation scheme of the singular dataset utilised in existing research, Scientific Entities, Relations and Coreferences (SciERC), is marred by ambiguity. Notably, two asymmetric relations within this dataset achieve recall rates of only 10% and 29%.

Keywords: Joint Relation Extraction (JRE); digital libraries; Named Entity Recognition (NER); Relation Extraction (RE); Pre-trained Language Model; transformer; SCIBERT; Scientific Entity Relation and Coreferences (SciERC); PL-Marker; semantic text analysis; global context

Citation: Evans, M.-T.C.; Latifi, M.; Ahsan, M.; Haider, J. Leveraging Semantic Text Analysis to Improve the Performance of Transformer-Based Relation Extraction. *Information* **2024**, *15*, 91. https://doi.org/10.3390/info15020091

Academic Editor: Gennady Agre

Received: 12 January 2024
Revised: 2 February 2024
Accepted: 2 February 2024
Published: 6 February 2024

1. Introduction

Search result relevancy for Digital Libraries is a pertinent challenge due to the growing number of scholarly publications. Additionally, there exists a distinct disinvestment compared to the commercial semantic search sector. Software-as-a-service companies and technology giants are financially incentivised to research the field due to the potential for advertising gains and contractual agreements. Conversely, the success of Open Access has led to archaic methods employed for Digital Libraries. Notably, such disparity results in technological research advancement ironically limited by the algorithmic performance [1].

To improve relevancy in the Digital Library domain, much research into recommender systems has been undertaken [2]. However, architecturally, semantic search result relevancy

is highly dependent on sequential information extraction tasks, Named Entity Recognition (NER)-deriving nouns, and Relation Extraction (RE)-obtaining relationships between nouns. These tasks populate Knowledge Bases from which search results are ultimately retrieved. As expected, approaches to such tasks have evolved in line with artificial intelligence advancements, from rudimentary, rule-based methods to supervised learning, with notable improvements in F1 scores. However, the widespread challenge for large, annotated input datasets for supervised approaches limits both performance and efficiency.

Commercial investment from Google and OpenAI has driven emerging research into performant Transformer-based Pre-trained Language Models, with encoder–decoder question-answering interfaces gaining great publicity in recent times. Consequently, this unsupervised pre-training reduces the input dataset requirement to a smaller, task-specific dataset requirement for fine-tuning.

Furthermore, NER and RE tasks have been combined into one task, formulating Joint Relation Extraction (JRE), in a bid to improve F1 scores and efficiency. JRE methods undertake either a joint or sequential learning approach. However, JRE F1 scores for literature-based datasets underperform versus the general domain, particularly for scientific literature. Furthermore, the lack of investment in research on Pre-trained Language Models for Digital Libraries is evidenced by the small number of fine-tuning datasets available.

Established linguistic theory underpins the well-known ambiguity for semantic and discourse-level classification tasks, such as JRE. It is therefore logical to assume a difficulty in the derivation of pragmatics. Conversely, lexical and syntactic Part-of-Speech (POS) tagging [3] and dependency parsing [3,4] evidence near-human accuracy when employing Transformer-based approaches. Yet, the reliance of fine-tuned classifiers for scientific JRE in current research on these linguistic theoretical challenges of language ambiguity [5] and context derivation [6–8] presents an undeniable limitation to existing research.

Subsequently, this paper presents a novel methodology to derive evidence-based challenges limiting JRE F1 scores in the scientific Digital Library domain. Current research fails to evidence the rationale for model optimisations, and, subsequently, improvements in F1 scores are negligible. Indeed, the novelty of this work extends beyond the task and domain under study. In fact, due to a widespread lack of scientific evidence underpinning Transformer fine-tuning, this work presents a semantic text analysis framework to evidence task-specific fine-tuned model optimisations. Employing the state-of-the-art fine-tuned scientific JRE classifier, PL-Marker [6], this study uses confusion matrices to analyse relation types with low recall. Semantic text analysis is subsequently undertaken on a relation sample to draw a correlation between semantic challenges derived from existing work and model improvements derived objectively from POS tagging due to the high task accuracy. To conclude, the resulting contributions are as follows:

- Firstly, the proposed framework evidences the subjective limitations of the model optimisations proposed in current research. In fact, global context and shared knowledge account for just 10.4% and 11.2% of the relation sample misclassifications, respectively.
- Secondly, this work is novel in quantifying pipelined error propagation within state-of-the-art JRE approaches for scientific literature as the largest performance limitation, accounting for 45.3% of misclassifications.
- In addition, the singular task-specific dataset available is proven to limit F1 score due to annotation scheme ambiguities, with [Feature-Of] and [Part-Of] achieving just 10% and 29% recall, respectively. There also exists a failure to appropriately represent complex noun phrases, accounting for 25.47% of total misclassifications, as the dataset contains solely paper abstracts. As such, a new test and train dataset is recommended.
- Finally, a weakness in BERT architecture for accurate modelling of localised language representations is highlighted in 16.04% of misclassifications. Future research into BERT bidirectional encodings is proposed.

The structure of this paper is outlined as follows. Section 2 presents a critical analysis of relevant work, highlighting the gap in existing research: the evaluation of current approaches to Transformer Pre-trained Language Models, recent supervised fine-tuning

approaches to JRE, and their methods for global context representation. Section 3 poses a novel methodology for semantic text analysis of model predictions, subsequently identifying the prevailing semantic challenges and model improvements, supported by POS tagging. Section 4 presents the results and analysis, and Section 5 critically analyses the findings with regards to the limitations of current research. Section 6 concludes the study findings, proposing areas for future research.

2. Recent Advancements

2.1. Transformer Pre-Training

BERT Pre-trained Language Model advanced language encoding improving F1 score by +7pp through bidirectional modelling of both right-to-left and left-to-right representations [9]. Conversely, previous state-of-the-art ELMo concatenates unidirectional representations [9,10] and highly publicised OpenAI GPT [11–14] fail to encode right-to-left context at all. Such comparative architectures inappropriately represent the semantics of natural language. Furthermore, domain-specific re-pretraining increased target task Relation Extraction (RE) F1 scores. BIOBERT, for the biomedical domain, improved F1 by +12.36pp [15], and SCIBERT achieved +2.51pp uplift for scientific literature, with a vocabulary overlap of just 42% with the general domain [16]. Naturally, semantic ambiguity is a core challenge for language encoding tasks, such as JRE. For example, 'aim', 'estimate', 'object', and 'use' could be nouns or verbs. In fact, domain-specific re-pre-training aims to resolve even more complex scientific ambiguities, such as complex noun phrases, containing prepositions, determiners, adjectives, and nouns, frequently labelled as entities in their entirety [17], such as SciERC terms 'space of candidate regions' and 'regular expressions' [18]. Such noun phrases are frequently present in scientific language [17], with the potential for entities nested inside these noun phrases [10]. This highlights a potential challenge in entity boundary definition for scientific Named Entity Recognition.

Target task RE is further abstracted syntactically, as relational verbs may present ambiguity. For example, 'is related to' pertains to several possible definitions, with the directionality also impacting the meaning. Indeed, relations may span multiple sentences [6,7] or only be determined through inference [6]. In such cases, syntactic ambiguity is prevalent, leading to a requirement to derive situated meaning outside of sentence structure or localised grammar. As a result, core challenges of context and shared knowledge have been assumed in recent work [6–8,10].

2.2. Fine-Tuning Approaches

To resolve boundary definition challenges, all recent SCIBERT fine-tuning approaches redefine entities as spans [1,6–8,10], building on BERT's BILOU tag representations [9,10]. Such improvements have resulted in structural extensions to the BERT encoder, enabling a token to pertain to 0 to many entities [6], as opposed to just one. In the example 'space of candidate regions' [18], span-based approaches derive the entire term, as well as the nested term 'candidate regions', as entities, whilst BILOU tags derive only a single entity.

A further fine-tuning approach is the evolution of multitask frameworks. DyGIE++ [8], SpERT [10], and SpERT.PL [1] jointly learn NER and RE simultaneously, as opposed to the traditional pipelining of NER classifications into the RE task [6,7]. Surprisingly, pipelining achieves state of the art [6]. However, lightly researched solutions for error propagation from NER to RE have been unfruitful, including training with predicted entities and increasing the relation sample [7].

2.3. Global Context

Despite such well-documented scientific language complexities, recent work assumes BERT's limited localised language representations and subsequent failure to encode global context are key areas of focus. Local context precedes or follows the entity boundaries, denoting the relation type, and it is frequently defined syntactically through prepositions and determiners. However, global context is relevant text denoting the relation, which is

situated further away from the target entity, perhaps in a different sentence entirely. Whilst multitask models, SpERT [10] and SpERT.PL [1], deem global context irrelevant, state-of-the-art classifier PL-Marker [6], amongst other top-performing approaches, including DyGIE++ [8] and PURE [7], incorporates methods for its encoding, such as a three-sentence context window (see Table 1). The multitask framework, DyGIE++, additionally learns coreference resolution, advocating a decrease in error propagation and an increase in learning propagation [8]. Yet, the pipelined model, PURE, has more recently demonstrated the negligible value of sharing entity representations globally through coreferences, increasing F1 by just +0.1pp, as local context is unique [7]. Indeed, specifically for scientific literature classification, DyGIE++ acknowledges that this approach introduces just as many errors as were resolved [8]. For example, whilst the entity 'three-dimensional objects' pertains to two relations, as shown in Figure 1, the local context token 'for' is only relevant for the [Used-For] relation. As shown in Table 1, it is notable that over time, fine-tuning optimisations have incorporated an ever-increasing number of global context methods.

Table 1. Comparison of global context methods in fine-tuning.

Algorithm	Year	Approach	Global Context Method					
					Markers		Batch Computation	
			Context Window	Coreference Resolution	Solid	Levitated	Inference	Training
DyGIE++ [8]	2019	Multi-task	x	x				
SpERT [10]	2019							
SpERT.PL [1]	2021							
Pure (Full) [7]	2020	Pipelined	x		x			
Pure (Approx.) [7]	2020		x	x	x	x	x	
PL-Marker [6]	2021		x		x	x	x	x

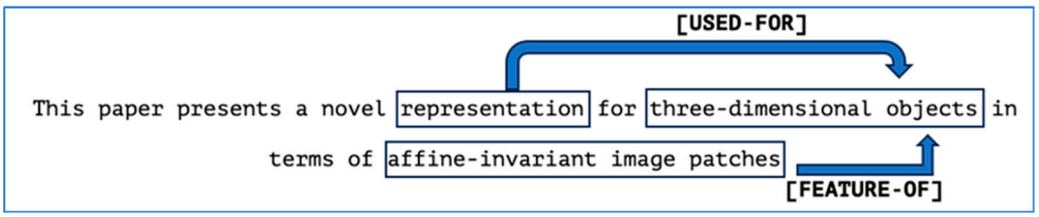

Figure 1. Example of unique local entity context for different relation types containing the same entities using SciERC.

PURE obtains global context by modelling interrelated objects of a subject, using solid markers for labelling [7] (see Table 1). Most recently, state-of-the-art PL-Marker advances encodings with the packed levitated marker. This approach associates multiple subject and interrelated object entities, employing novel packing strategies for combined logical inference [6]. Both approaches are demonstrated in Figure 2. PURE's solid markers generate three independent span pairs containing the subject 'Copenhagen' and objects 'David', 'workers', and 'teammates'. However, the packed levitated marker packs the objects and subject together to benefit from the semantic interrelation between the entities within the sentence.

Indeed, emerging research has continued to optimise pipelined models for global context. Cascade-SRN investigates the decomposition of the RE task into a subject extraction task and a subject-oriented object extraction task [19]. PREFER further explores the use of attention mechanisms to derive spans and context [20]. Yet neither approach has achieved state of the art.

Sentence: 'David assaulted a pair of restaurant workers during a night out with national squad teammates in Copenhagen'

Solid Markers:

[S]Copenhagen[/S]	[O]David[/O]
[S]Copenhagen[/S]	[O]workers[/O]
[S]Copenhagen[/S]	[O]teammates[/O]

Packed Levitated Markers:
[S]Copenhagen[/S] [O]David[/O] [O]workers[/O] [O]teammates[/O]

Figure 2. PL-Marker-packed levitated markers compared to PURE solid markers.

2.4. Research Gap

Despite the focus of emerging research on encoding global context [6–8], performance improvements are negligible, with just a +4.8pp increase in F1 score for target task RE since 2019, as shown in Figure 3. State-of-the-art PL-Marker achieves an F1 score of 53.2. However, in comparison, F1 scores for the general domain news and digital forum dataset, ACE-05, achieve up to 69 [6].

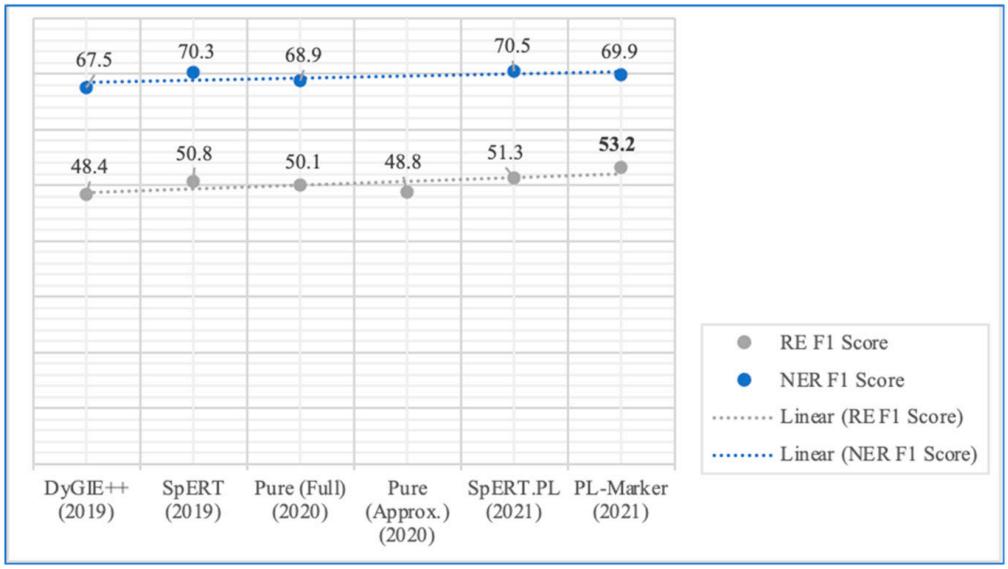

Figure 3. JRE F1 Scores for state-of-the-art, fine-tuned classifiers using SCIBERT Pre-trained Language Model and SciERC dataset.

Current research fails to evaluate potential performance constraints related to the singular scientific dataset employed in all research to date, Scientific Entities, Relations and Coreferences (SciERC). SemEval 2017 and 2018 [17,18] are noted as two alternatives. However, the former is proven to have a bias towards NER, and the latter only annotates intra-sentence relations [18], which is unrepresentative of natural language.

Compared to the general domain, F1 scores indicate prevailing challenges for semantic encoding. Yet, existing research fails to evidence the rationale for optimisations to resolve purported challenges in language ambiguity [5], logical inference [6], and global context [6–8]. Furthermore, whilst solutions to the pipelined error propagation issue have been lightly investigated, no known work has quantified the impact of this challenge on the model F1 score.

Subsequently, this work presents a semantic text analysis framework to scientifically evidence the prevailing challenges and subsequent semantic model optimisations required to improve F1 scores for scientific JRE.

3. Semantic Text Analysis Framework

The framework is outlined in Figure 4. The state-of-the-art, fine-tuned pipelined classifiers, PL-Marker [6], take as input the test set for the NER task and the test set and entity predictions for the RE task. Relation types with low recall, which drive low model F1 score, are identified through confusion matrices, forming the relation sample. POS tag pre-processing facilitates the objective annotation of syntactic identifiers during text analysis. Predictions and POS tags form the input for semantic text analysis, from which a deductive hierarchy of theme annotations is made to ultimately draw a correlation between prevailing semantic challenges and syntactic improvement proposals for future model optimisation. Additional analysis of the SciERC dataset distribution supports the findings on dataset limitation conclusions.

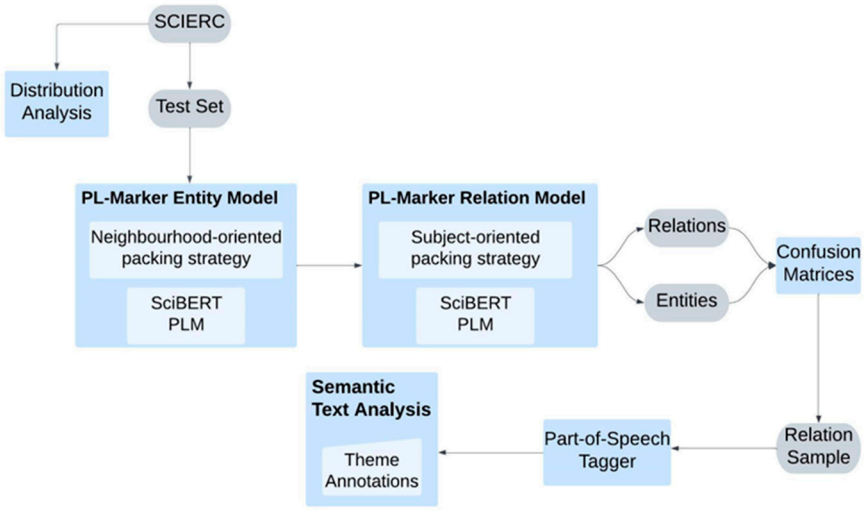

Figure 4. Semantic text analysis framework.

3.1. Dataset

SciERC contains 500 abstracts from 12 NLP conferences, and the source data have been retrieved from Semantic Scholar. The dataset includes annotations for named entities, relations, and coreferences [18]. In line with PL-Marker, the original SciERC test set is employed to evaluate the pre-trained model, and coreference annotations are excluded. SciERC Kappa statistics state a named entity annotations score of 76.9%, and relation annotations score 67.8%. Examples of the seven target relation types are displayed in Figure 5, highlighting the directionality of each relation type. For example, [Compare] and [Conjunction] are symmetric relation types denoted by the syntax 'and', whilst 'sufficient computational resources' are a [Feature-Of] 'devices', with an asymmetric right-to-left relation type denoted in this example.

3.2. Task Problem Definitions

For the NER task, an entity is correctly predicted when the start and end tokens and entity type meet the gold label provided. The NER problem definition is detailed in Equation (1) below.

$$E = \{t_m, t_n, entity_{type}\} \tag{1}$$

For target task RE, boundaries evaluation is employed, defining a relation as correctly predicted when the relation type and the subject and object entity bounds match the gold label provided. The RE task definition is detailed in Equation (2) below.

$$R = \{E_1 t_m, E_1 t_n, E_2 t_m, E_2 t_n, relation_{type}\} \tag{2}$$

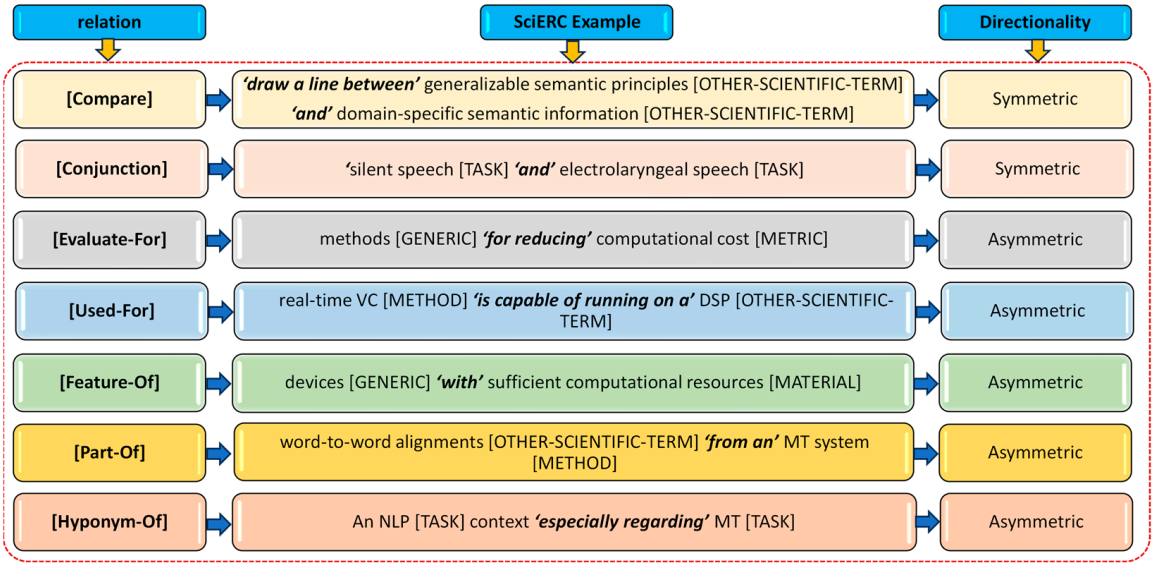

Figure 5. SciERC relation type examples, including relation directionality.

Equation (2) demonstrates the importance of the token positionings for asymmetric relation types, where the directionality of the relation is denoted by the order in which the entities in the formula are defined by the gold label. For instance, in the SCIERC example, 'A domain independent model is proposed for the automated interpretation of nominal compounds in English entities 'domain independent model' and 'automated interpretation of nominal compounds' are defined'. Equation (3) demonstrates the gold label relation.

$$R = \{1, 3, 8, 12, [\text{Used} - \text{For}]\} \tag{3}$$

In this case, [Used-For] is an asymmetric left-to-right relation. Whilst it may be argued that a relation exists between these two entities from right to left, the relation type would not be [Used-For], as changing the positioning of the entities within the gold label equation would not have the same semantics as the above sentence.

3.3. Distribution Analysis

SciERC dataset distribution is assessed by comparing entity and relation labels across the dev, test, and train sets. This analysis quantifies potential dataset performance limitations when analysed in conjunction with the confusion matrices defined in Section 3.8, such as potential under- or over-fitting. However, generalisations of entity and relation representation to the scientific field cannot be made due to the lack of comparative datasets available. Such distribution analysis for JRE datasets is novel, not only in the scientific but also the general domain. Subsequently, the results may be used in future research as supplementary insight for model performance to identifying possible dataset bias, which negatively impacts F1 scores.

3.4. PL-Marker Model Architecture

3.4.1. SCIBERT Encoders

The NER and RE models each contain a SCIBERT encoder within their architecture, with the predicted entity output from the NER model fed as input to the RE model. SCIBERT is pre-trained unsupervised on 1.14 million Semantic Scholar journals, aided by scientific vocabulary, SCIVOCAB [16]. Structural extensions to encoders are commonplace in fine-tuned model architectures [7,8]. PL-Marker employs extensions to SCIBERT to facilitate span and span pair representations, a three-sentence context window, and the model-specific packing strategies defined below. Initially, PL-Marker pre-training parameters are used to initialise each encoder. The SciERC dev set is then employed for parameter fine-

tuning [16,18]. Finally, the SciERC training set is fed through the encoder, creating localised language representations, which incorporate bidirectional language understanding [9]. The outputs are then classified during fine-tuning. In this work, the pre-trained PL-Marker model is employed using the evaluation bash scripts provided and the SciERC test set.

3.4.2. Entity Model Packing Strategy

The NER model structural extension categorises spanss into groups using a neighbourhood-oriented packing strategy; those that share the same start or end token are defined as nested entities [6]. For example, the entity 'Bank of China' is packed together with nested token 'China' [6]. Secondly, levitated markers are applied to the span, which then forms input to the SCIBERT NER encoder. Finally, span filtering classifies relevant spans as a specific entity type.

3.4.3. Relation Model Packing Strategy

The relation model structural extension employs a subject-oriented packing strategy to combine entity objects associated with a subject into a single instance for batch training and inference. During this process, solid and levitated markers are assigned, respectively. Figure 6 demonstrates the subject 'direction-giving task', which is packed with objects 'eye gaze', 'head nods', and 'attentional focus' to derive common relations [18]. If a [Conjunction] relation is derived between objects 'eye gaze', 'head nods', and 'attentional focus' and one of the three [Part-Of] relations to the subject is derived, then the other object-to-subject relations can be inferred.

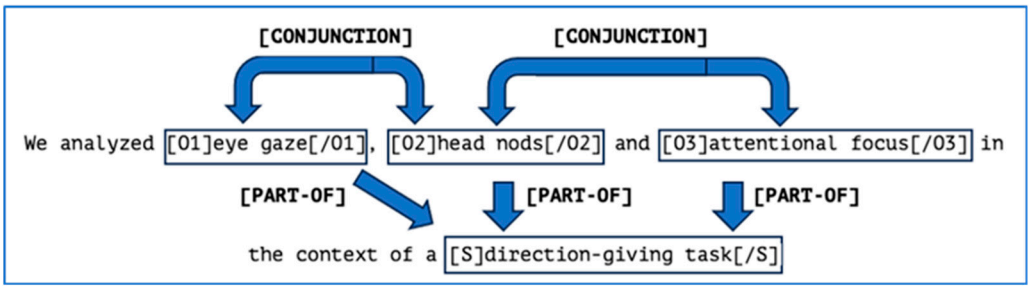

Figure 6. Example of PL-Marker subject-oriented packing strategy using SciERC.

The contextualised span pair representation outputs are then classified as relation types. Model F1 scores are evaluated against the standard deviations reported in the original PL-Marker experiment to ensure validity in conclusions.

3.5. Part-Of-Speech Tagging

NLTK's default Perceptron tagger is employed due to its high accuracy compared to TNT and CRF taggers [21]. State-of-the-art POS tagging with the Penn Treebank dataset achieves an F1 Score of 98.3 [3]. Such pre-processing provides insight into syntactic–semantic relationships. For example, a [Part-Of] relation is indicated by the preposition 'from' and the determiner 'an', which are directed from the entity 'word-to-word alignments' to the entity 'MT system' (Figure 7). Furthermore, the semantic text analysis framework example in Figure 8 shows that POS tagging highlights the entity 'iterative deformation of a 3-D surface mesh' is a complex noun phrase, inclusive of an adjective, determiner, and preposition [18].

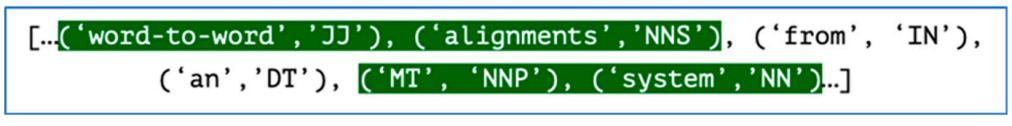

Figure 7. Example of POS tagging using Natural Language Toolkit (NLTK) and SciERC.

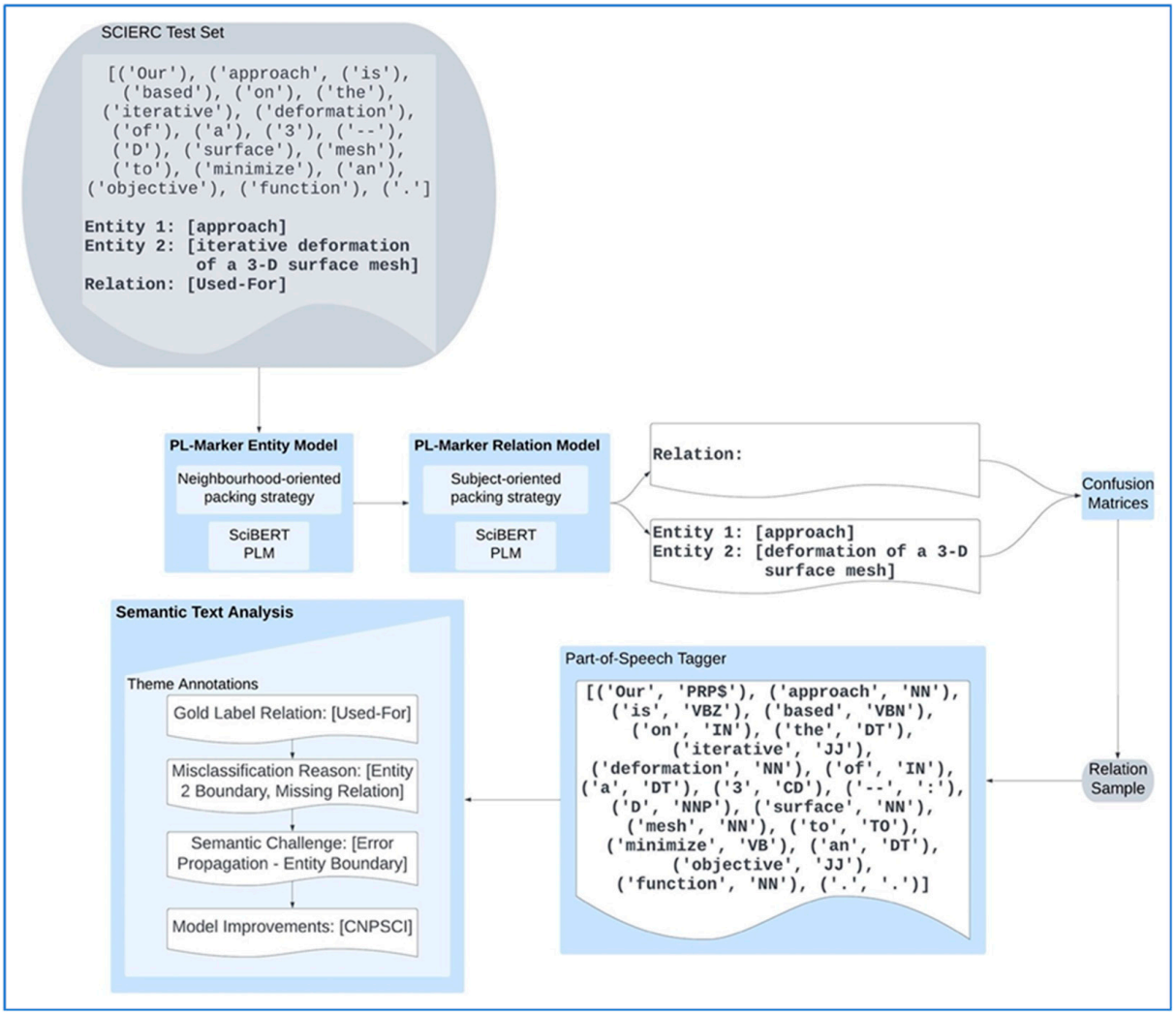

Figure 8. Example of the semantic text analysis process across the hierarchy of all 4 themes using SciERC.

3.6. Semantic Text Analysis

The semantic text analysis process applied to the relation sample is shown in the lower left of Figure 8. The method follows a deductive hierarchical approach, where each theme is informed by the previous coded theme. Two evidence-based themes are initially coded. Firstly, the SciERC gold relation labels are coded, such as [Feature-Of] and [Part-Of]. Secondly, misclassification reason codes—such as missing entity or misclassified relation type—are derived using the task problem definitions listed in Section 3.2 and the analysis steps listed in Section 3.8. The purpose of the misclassification reason theme is to define objective reasons as to why the prediction does not match the gold label provided. Two content codes are subsequently derived hierarchically: semantic challenges, followed by model improvements. Semantic challenges, such as error propagation, global context, or local context, outlined in Section 2.1, are annotated. Finally, the improvement theme identifies syntactic improvements, aided by POS tags. For example, these include the positioning of syntax indicating relations within local context, demonstrated in Figure 7. The codebooks are presented in the following Appendices: Appendices A–D Table A4. Combinations of semantic challenges and improvements are assessed for statistical significance using theme co-occurrence and cluster analysis. Both normalised percentages and raw frequencies are analysed to determine pertinent themes. In addition, relations are annotated twice due to the small sample, resulting in Kappa scores for reproducibility.

3.7. Implementation Details

The framework is built in Python using Google Colaboratory as an IDE, and it is executed on a T4 GPU. HuggingFace Transformers library facilitates the import of Pre-trained Language Model scibert-scivocab-uncased, whilst customised Transformer scripts within the PL-Marker models facilitate the structural extensions detailed in Section 3.4. The framework employed is PyTorch, in line with PL-Marker. Pandas library enables efficient data manipulation, whilst scikit-learn and Matplotlib libraries generate and visualise confusion matrices, respectively. Natural Language Toolkit (NLTK) facilitates POS tagging. For semantic text analysis, qualitative data analysis software, NVivo, is used due to its functionality for matrix coding for theme co-occurrence. In addition, NVivo allows for the generation of Kappa statistics and hierarchical cluster analysis for correlation coefficients, defined in Section 3.8.

The code and analysis for reproducing the results of this framework are publicly available on GitHub: https://github.com/mtclevans/semantictextanalysis (accessed on 5 February 2024).

3.8. Evaluation Metrics

Confusion matrices are used to measure model predictions against SciERC gold labels. Section 3.2 details the industry task formulations, which are broken down further to identify classification issues. As such, Figure 9 details the specific NER analysis steps used. For example, pipelined error propagation is quantified by negating the requirement for either the correct start or end token. Furthermore, the analysis of nested entities evaluates whether the PL-Marker neighbourhood packing strategy is performant. Figure 10 details the analysis steps for the RE task, which allow for the identification of error propagation due to missing entities through the negation of either start or end entity requirements.

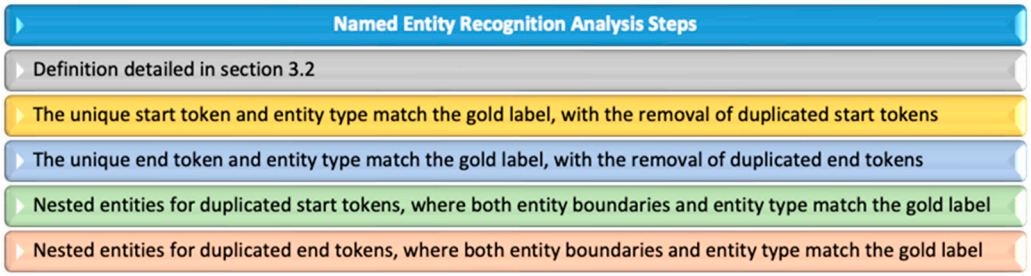

Figure 9. Named Entity Recognition analysis steps.

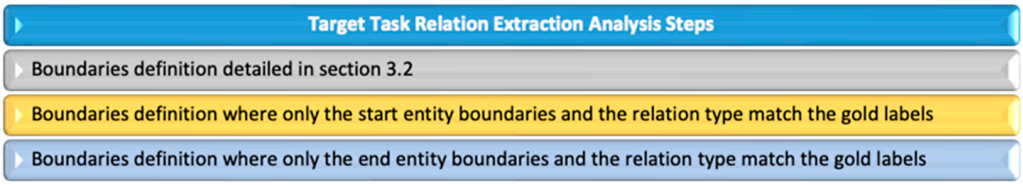

Figure 10. Relation Extraction analysis steps.

A relation sample for semantic text analysis is defined as those relation types with the lowest recall to contain the scope of the research. In recent work, and in the general domain, overall model classification accuracy is measured using F1 score. However, in this work, recall highlights the proportion of gold labels that were classified correctly through confusion matrices. This results in the ability to evaluate the labels that have the lowest proportion of true positives, where true positives pertain to gold labels. Subsequently, labels for improvements to overall model performance are identified. *Recall* is defined through Equation (4) below.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Theme co-occurrence matrices of semantic annotations are assessed using normalised percentages to allow for relative comparison. Pearson's Correlation Coefficient, defined through Equation (5) below, enables comparison of each combination of challenge and improvement relative to other combined pairs.

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \Sigma(Y_i - \overline{Y})^2}} \quad (5)$$

4. Results and Analysis

4.1. PL-Marker Performance

Challenges in identifying entity boundaries are prevalent with almost all labels, as demonstrated by the improvement upon removal of either the start or end token requirement (Figure 11). Those with the lowest standard recall, [Metric] and [Other-Scientific-Term], see the greatest improvements upon removal of the requirement for a correct start token. In fact, most labels see a greater increase in recall in this scenario. [Metric] accounts for just 4% of the train and test sets, alongside obtaining the highest proportion of standard NER Not Predicted values. Indeed, such entity error propagation is notable, as certain named entities often pertain to certain relation types, such as [Metric] within [Evaluate-For]. Slight underfitting is noted from train to test for [Other-Scientific-Term], which accounts for 31% of the test set. In the absence of alternative datasets, it is infeasible to determine whether this high proportion is representative of the scientific domain. Further analysis demonstrates [Other-Scientific-Term] entities are frequently common and compound nouns, as well as complex noun phrases with leading adjectives. This suggests entity boundary error propagation challenges for scientific terminology, which is investigated further in Section 4.2.1.

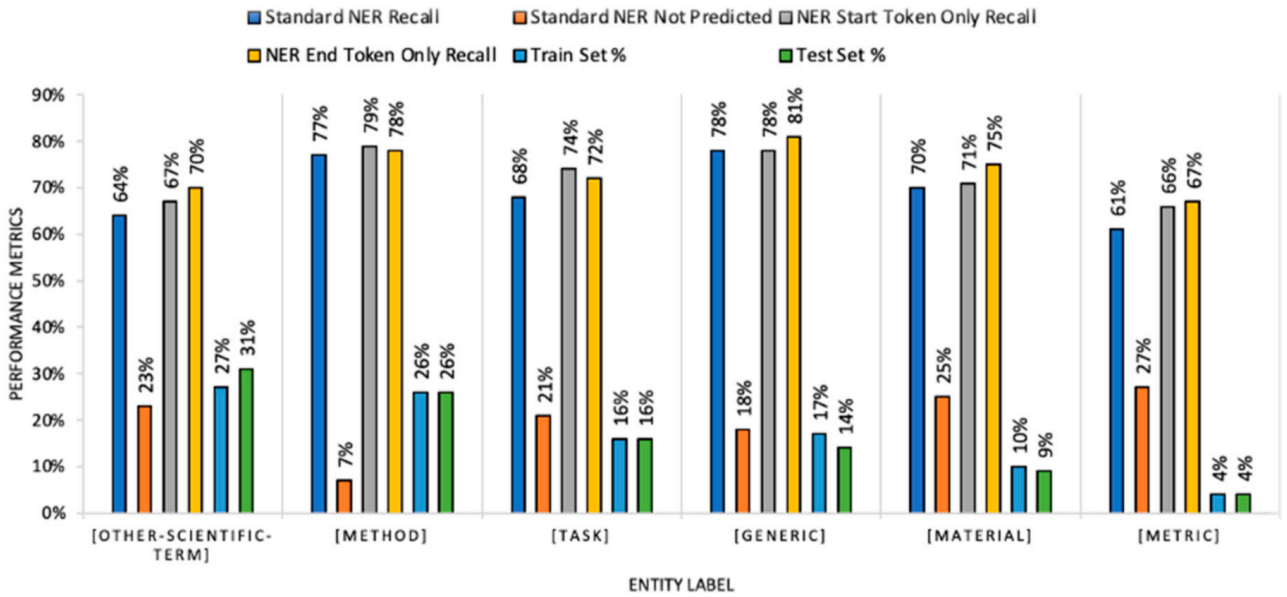

Figure 11. Comparison of NER Standard Recall, Start and End Token Recall, and Standard NER Not Predicted values, alongside dataset distribution statistics.

Nested entity samples are small and invalid for analysis, yet PL-Marker fails to predict the majority. Indeed, SciERC is argued to be more representative than SemEval 2017 and 2018 [18], yet it fails to evaluate nested entity distribution. This low representation of nested entities results in the inability to quantify the performance of the PL-Marker neighbourhood-oriented packing strategy, as well as the evaluation of span-based approaches in general.

Moving to relation analysis, [Compare] accounts for just 4% of the test set (Figure 12), whilst [Conjunction] recall overperforms at 68%. This highlights the ease of symmetric clas-

sification. Indeed, the annotation scheme guidelines define simple syntactic conjunctions 'and' and 'or' as denoting the [Conjunction] relation type. Furthermore, it is notable that symmetric and asymmetric relations are only misclassified as their respective types or not at all. This suggests BERT architecture distinguishes directionality but classifies symmetric relations more accurately.

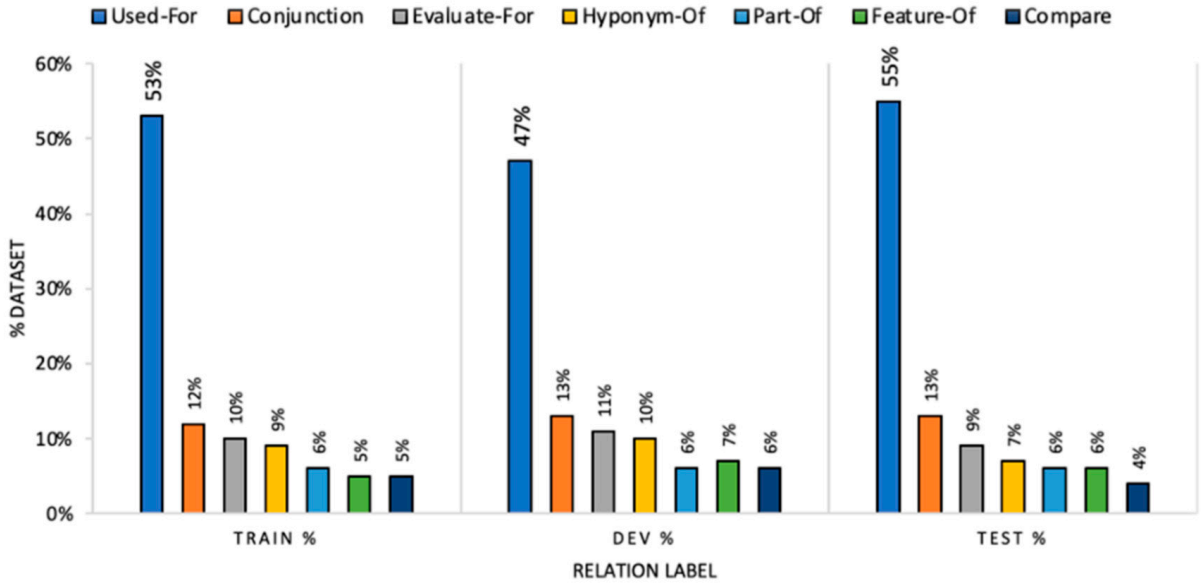

Figure 12. SciERC dataset statistics in Python—Relations.

The lowest recall scores are obtained by asymmetric relations, [Feature-Of] and [Part-Of], which also have the highest proportion of Not Predicted values (Table 2). Both labels account for only 6% of the test set, respectively (Figure 12). Misclassifications for [Part-Of] are often predicted as [Feature-Of] and [Hyponym-Of]. This is notable, as these relations share semantic similarity and may present ambiguity, even for a human annotator. [Feature-Of] and [Hyponym-Of] may be defined as component relations to [Part-Of], which suggests the model may lean towards a more specific, rather than generalised, classification. These two relation types form the semantic text analysis sample.

Table 2. RE confusion matrix recall—standard task formulation. Key statistics referenced are highlighted in bold.

True Label / Predicted Label	Compare	Conjunction	Evaluate-For	Feature-Of	Hyponym-Of	Part-Of	Used-For	Not Predicted
Compare	**53%**	**11%**	0%	0%	0%	0%	0%	**36%**
Conjunction	**1.6%**	**68%**	0%	0%	0%	0%	0%	**30%**
Evaluate-For	1.1%	0%	46%	1.1%	0%	0%	6.6%	45%
Feature-Of	0%	0%	3.4%	**10%**	0%	1.7%	6.8%	**78%**
Hyponym-Of	0%	0%	0%	0%	54%	3%	3%	40%
Part-Of	0%	0%	0%	**4.8%**	**6.3%**	**29%**	6.3%	**54%**
Used-For	0%	0.19%	0.38%	0.19%	0%	0.38%	57%	42%

4.2. Semantic Text Analysis

The average weighted Kappa score for semantic text analysis is 88%, as detailed in Table 3. The gold relation theme is derived from ground truth annotations, and thus a score of 100% is expected. Misclassification reasons are derived objectively from, for example, incorrect entity boundaries, missing relations, and incorrect directionality. Thus, it is expected that the Kappa statistics for each sequential theme decline in line with the inductive approach.

Table 3. Semantic text analysis Kappa statistics.

Theme	Theme Annotation Hierarchy	Kappa Score
Gold Relation	1	100%
Misclassification Reason	2	87%
Semantic Challenge	3	84%
Improvement	4	82%
Average Weighted Kappa Score	-	88%

4.2.1. Error Propagation

NER error propagation is quantified as the largest challenge, accounting for 45.3% of total relation misclassifications. This results in missing entities and entity boundary misclassifications. The key improvement themes are displayed in Table 4. It is important to note that not all themes have been listed within the result tables to ensure that a succinct summary of the key themes is presented. Whilst it is evident that the model frequently fails to predict common nouns (CN), such as 'task', 46.5% of total entity error propagation pertains to misclassifications of complex noun phrases (Table 4). In a third of cases, missing entities are due to the failure to predict such a term (CNP (GEN), (CNP (SCI)). The complex syntactic composition of such missing entities as complex noun phrases is demonstrated by the adjectival noun 'productive affixations of derivational and inflectional suffixes' in Figure 13a. In 66.7% of entity boundary cases, complex noun phrases are predicted where they do not exist, incorporating adjectives, prepositions, and determiners into the entity start and end token boundaries (INCORRECT CNP). This further substantiates the misclassification of leading adjectives in complex noun phrases, initially identified in Section 4.1. Figure 13b demonstrates misclassified entity boundary cases, where both entities 'weighted sum' and 'precision', the conjoining relational local preposition 'of', and the determiner 'in' are predicted as one term.

Table 4. NER error propagation semantic challenges and improvement themes. Notable statistics are highlighted in bold.

NER Error Propagation / Code Label	Missing Entities			Entity Boundaries			Total	
	Raw	Theme Co-Occurrence	Correlation Coefficient	Raw	Theme Co-Occurrence	Correlation Coefficient	Raw	Theme Co-Occurrence
CNP (GEN)	5	12.5%	0.82	-	-	-	5	**8.6%**
CNP (SCI)	8	20.0%	0.90	2	11.1%	0.81	10	**17.2%**
CN	15	**37.5%**	**0.94**	-	-	-	15	25.9%
CPN (SCI)	4	10.0%	0.72	-	-	-	4	6.9%
CPN	3	7.5%	0.72	-	-	-	3	5.2%
INCORRECT CNP	-	-	-	12	**66.7%**	0.97	12	**20.7%**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total Theme Annotations	**40**	-	-	**18**	-	-	**58**	-

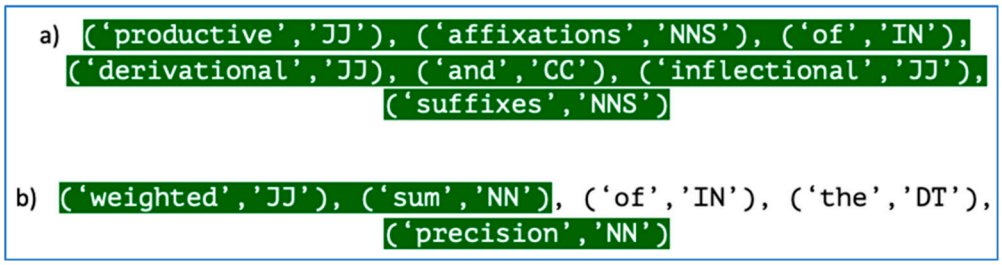

Figure 13. Entity error propagation (**a**) missing complex noun phrase (**b**) boundary misclassification predicting two entities and the relation as one complex noun phrase.

4.2.2. Relational Ambiguity

Disambiguation of shared knowledge accounts for a large proportion of relational ambiguity cases (Table 5). However, pertaining to only 11.2% total misclassifications, this work demonstrates that its significance within current research is unwarranted. In fact, relational ambiguity is heavily distorted by the misclassification of [Part-Of] as [Feature-Of] and [Hyponym-Of], previously identified in Section 4.1. In Figure 14, 'layers' is classified as a [Feature-Of] 'CNN models', and 'Kalman filter' as a [Hyponym-Of] 'prediction techniques', whilst both gold relations are defined as [Part-Of]. However, both examples demonstrate that [Hyponym-Of] and [Feature-Of] are component relations to [Part-Of]. This demonstrates clear SciERC annotation scheme ambiguities in differentiating between these relation types. It is unsurprising, therefore, that the SciERC relation Kappa score drops to 67.8% from the NER score of 76.9% [18].

Table 5. Relational ambiguity semantic challenge and improvement themes.

Relational Ambiguity / Improvement Theme	Raw	Theme Co-Occurrence	Correlation Coefficient
SHARED KNOWLEDGE INFER	11	42.3%	0.87
PART-OF as FEATURE-OF	3	11.5%	0.75
PART-OF as HYPONYM-OF	4	15.4%	0.72
FEATURE-OF as PART-OF	1	3.8%	0.51
⋮	⋮	⋮	⋮
Total	**26**		

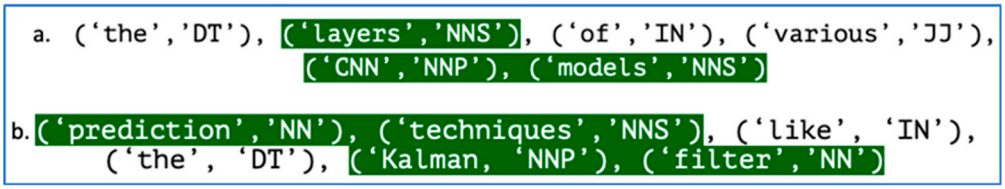

Figure 14. Examples of [Part-Of] predicted as (**a**) [Feature-Of] and (**b**) [Hyponym-Of].

4.2.3. Context Derivation

Local context cases are defined by correct entity classification, with a failure to derive local context denoting a relation. As such, 40.9% of misclassifications contain prepositions and an optional determiner, denoting a [Feature-Of] or [Part-Of] relation (In(IN) + a(DT) POST, (IN) BETWEEN ENTITIES) (Table 6). Indeed, a further 27.3% of cases pertain to a misclassified relation directionality (In(IN) DIRECTIONAL). Figure 15a,b demonstrates the relation [Part-Of] misclassified as [Used-For], regardless of local prepositions 'in' and 'from', whilst in Figure 15c, 'in' denotes the directionality of 'VLSI' as a [Feature-Of] 'dynamics'. In addition, further analysis of cases where no relation is classified, or the wrong directionality is assigned, demonstrates asymmetric left-to-right relations are half as likely to be misclassified versus right-to-left. Whilst Section 4.1 details BERT's success in distinguishing between symmetric and asymmetric relations, this more detailed insight challenges BERT's bidirectional architectural approach.

In a small number of overall context cases, 8.11%, the local preposition is insufficient to derive the relation alone, and context within the wider sentence or paragraph is required (CONTEXTPRE + (IN)). However, such a low proportion indicates the PL-Marker cross-sentence mechanism may be performant.

Global context accounts for just 10.4% of total misclassifications, with almost 50% due to the failure of the subject-oriented packing extension within PL-Marker (PACKINGSUB-REL). Nevertheless, as multiple relations pertain to one packed instance, the sample size is too small to derive conclusions.

Table 6. Local and global context semantic challenges and improvement themes.

Context / Improvement Theme	Local Context			Global Context			Total	
	Raw	Theme Co-Occurrence	Correlation Coefficient	Raw	Theme Co-Occurrence	Correlation Coefficient	Raw	Theme Co-Occurrence
In(IN) DIRECTIONAL	6	27.3%	0.85	-	-	-	6	16.2%
In(IN) + a(DT) POST	5	22.7%	0.83	-	-	-	5	13.5%
(IN) BETWEEN ENTITIES	4	18.2%	0.82	-	-	-	4	10.8%
CONTEXTPRE + (IN)	1	4.5%	0.64	2	13.3%	0.69	3	8.1%
CONJ(CC) + PRONOUN PRP($)	1	4.5%	0.61	-	-	-	1	2.7%
PACKINGSUBREL	-	-	-	7	46.7%	0.93	7	18.9%
SHARED KNOWLEDGE INFER	-	-	-	1	6.7%	0.66	1	2.7%
⋮	-	-	-	⋮	⋮	⋮	⋮	⋮
Total	**22**	**-**	**-**	**15**	**-**	**-**	**37**	**-**

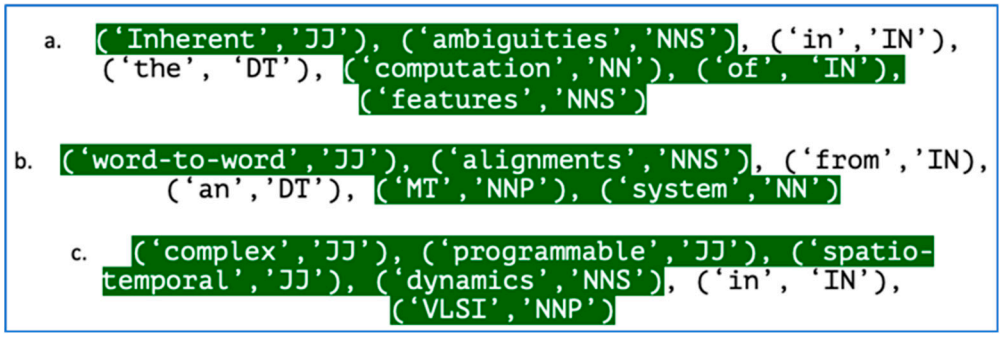

Figure 15. Instances of PL-Marker misclassifications, despite the presence of local prepositions, such as (**a**) 'in', (**b**) 'from', and (**c**) 'in'.

5. Critical Analysis and Discussion

5.1. Strengths

Contrary to the research of state-of-the-art JRE classifiers [6–8], this framework evidences global context and shared knowledge as impertinent challenges. In fact, SciERC is proven to limit performance, with annotation scheme ambiguities for [Part-Of], [Feature-Of], and [Hyponym-Of] relations. Furthermore, the small, nested entity sample renders the performance of the PL-Marker neighbourhood-oriented packing strategy, as well as overarching span-based approaches over BILOU tags, inconclusive for all emerging classifiers [1,6–8,10].

Most importantly, this research highlights PL-Marker's failure to represent, and subsequently predict, the presence of complex noun phrases, or the lack thereof, resulting in pipelined error propagation. Syntax within complex noun phrases is often restructured, formulating new terms. As such, a larger test and train set of more abstracts will fail to resolve this issue. However, the repetition of such terms across an entire paper results in a recommendation for a new dataset encompassing full papers. This is evidenced by the frequency of the term 'complex noun phrase' in this entire paper compared to the frequency in the abstract. This work is the first known work to scientifically quantify the pipelined error propagation issue.

The prevalence of NER common noun misclassifications suggests that the DyGIE++ multitask framework should perform better due to the incorporation of coreference resolution into the fine-tuning approach. Whilst the authors of PURE claim that sharing localised entity representations for coreferences incurs further error propagation [7], this research demonstrates that these common noun entities were not classified at all. Subsequently, they were not provided as input to the relation model. Further evaluation of annotations of common nouns as entities in SciERC may unearth further weaknesses in the dataset.

For misclassifications where the NER model is performant, local prepositional context denoting the relation type and directionality is largely present. SCIBERT should encode

such context within its representations. However, the improved performance of symmetric over asymmetric relations, and the likelihood of misclassified right-to-left relations, suggest a weakness in BERT's purported bidirectional encodings [9]. Further research into the performance of BERT's bidirectionality is required, alongside comparisons with other domain datasets for similar asymmetric relation types.

5.2. Limitations

Acknowledged limitations of this work are the employment of the sole scientific dataset, and the subsequent absence of conclusions drawn against the wider domain. Furthermore, as the study scope defines two relation types for semantic text analysis, this work provides no generalisation to the entire SciERC relation sample, and no qualitative comparison of symmetric relation types has been undertaken. Finally, this study is limited to conclusions drawn against misclassifications, and, as such, it provides no comparison of true positives in its semantic challenge conclusions.

6. Conclusions and Future Work

This study advances previous work by establishing a framework for semantic text analysis of JRE classification. It has been demonstrated that current research optimisations of fine-tuned Transformers are based on subjective assumptions that global context derivation is the key issue. However, this semantic challenge only accounts for 10.4% of misclassifications. Equally, just 11.2% of misclassifications relate to a lack of shared knowledge. The key performance limitation for scientific JRE is evidenced as pipelined error propagation, accounting for 45.3% of misclassifications. Such difficulties result from complexities in the identification of complex noun phrases, which alone account for 25.47% of relational misclassifications. The sole dataset employed in research to date is proven to present annotation scheme ambiguities, which result in difficulty in distinguishing between [Feature-Of], [Part-Of], and [Hyponym-Of] relation types. This is evidenced by the low recall of the former two labels, at just 10% and 29%, respectively, and high false positives for [Part-Of] as [Feature-Of] and [Hyponym-Of]. Furthermore, key encoding issues with BERT Pre-trained Language Model are identified, with the misclassification of right-to-left relations versus left-to-right relations increasing twofold.

Subsequently, future work may create a new representative test, and train scientific journal dataset, and further investigate the ability of BERT-based encoders to appropriately represent localised entity context and model semantics of right-to-left relation types. Such improvements in Pre-trained Language Model approaches to JRE for scientific literature will have profound effects on the research domain. As the core information retrieval task within a Digital Library architecture, advancements in JRE F1 score will lead to more efficient and relevant referencing, upon which scientific researchers will base their own work. Subsequently, algorithmic improvements to such semantic search tools have the potential to transform the rate of scientific technological advancement.

Author Contributions: All authors made an equal contribution in preparing and finalising the manuscript. Conceptualisation, M.-T.C.E. and M.L.; methodology, M.-T.C.E., M.L. and M.A.; validation, M.-T.C.E., M.L., M.A. and J.H.; formal analysis, M.-T.C.E., M.L., M.A. and J.H.; investigation, M.-T.C.E., M.L., M.A. and J.H.; data curation, M.-T.C.E.; writing—original draft preparation, M.-T.C.E. and M.L.; writing—review and editing, M.-T.C.E., M.L., M.A. and J.H.; supervision, M.L., M.A. and J.H. 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: Publicly available datasets were analysed in this study. This data can be found here: https://nlp.cs.washington.edu/sciIE/ (accessed on 1 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Theme level 1 codebook—gold label relations.

Code Label	Definition	Example SCIERC
FEATURE-OF	The gold relation label defined in the SCIERC test set is of the type FEATURE-OF, which denotes an entity that describes a feature another entity.	devices [GENERIC] 'with' sufficient computational resources [MATERIAL]
PART-OF	The gold relation label defined in the SCIERC test set is of the type PART-OF, which denotes a part–whole relation, where one entity forms a part of another whole entity.	a priori geometric constraints [OTHER-SCIENTIFIC-TERM] 'in a' 3D stereo reconstruction scheme [METHOD]

Appendix B

Table A2. Theme level 2 codebook—misclassification reasons.

Code Label	Name	Definition	Example SCIERC
ENT1	Missing entity 1	The first entity in the relation is not predicted.	Gold: ['English', 'nominal compounds', 'FEATURE-OF'] Prediction: [N/A, 'nominal compounds', N/A]
ENT2	Missing entity 2	The second entity in the relation is not predicted.	Gold: ['ambiguity', 'determiners' 'FEATURE-OF'] Prediction: ['ambiguity', N/A, N/A]
ENT1BOUND	Entity 1 boundaries	The first entity is predicted. However, the boundaries are not correctly predicted, resulting in the misclassified start AND/OR end token.	Gold: ['high-density, low-power analog array', 'externally digital architecture', 'PART-OF'] Predictions: ['low-power analog array', 'externally digital architecture', 'PART-OF']
ENT2BOUND	Entity 2 boundaries	The second entity is predicted. However, the boundaries are not correctly predicted, resulting in the misclassified start AND/OR end token.	Gold: ['generalization ability', 'learned metric', 'FEATURE-OF'] Prediction: ['generalization ability', 'metric', 'FEATURE-OF']
REL	Missing relation	The relation is not predicted.	Gold: ['outlier removal', 'stereo vision', 'PART-OF'] Prediction: []
RELDIRECT	Incorrect relation direction	The relation is predicted. However, the direction of the relation is misclassified, resulting in the second entity listed first and the first entity listed second.	Gold: ['indirect lighting', 'robustness', 'FEATURE-OF'] Prediction: ['robustness', 'indirect lighting', 'FEATURE-OF']
RELTYPE	Incorrect relation type	The relation is predicted. However, the type of relation is misclassified.	Gold: ['layers', 'CNN models', 'PART-OF'] Prediction: ['layers', 'CNN models', 'FEATURE-OF']

Appendix C

Table A3. Theme level 3 codebook—semantic challenges.

Code Label	Name	Definition	Example SCIERC	Exceptions
CONT	Global Context	The relation is semantically understood from the global context and situated outside of the relevant text segment, where the text segment spans from the start entity to the end entity.	(‘Recognition’, ‘NN’), (‘of’, ‘IN’), (‘proper’, ‘JJ’), (‘nouns’, ‘NNS’), (‘in’, ‘IN’), (‘Japanese’, ‘JJ’), (‘text’, ‘NN’), (‘has’, ‘VBZ’), (‘been’, ‘VBN’), (‘studied’, ‘VBN’), (‘as’, ‘IN’), (‘a’, ‘DT’), (‘part’, ‘NN’), (‘of’, ‘IN’), (‘the’, ‘DT’), (‘more’, ‘RBR’), (‘general’, ‘JJ’), (‘problem’, ‘NN’), (‘of’, ‘IN’), (‘morphological’, ‘JJ’), (‘analysis’, ‘NN’)	Does not contain local context.
COREF	Coreference	One or both entities within the relation are coreferences.	(‘The’, ‘DT’), (‘objects’, ‘NNS’), (‘can’, ‘MD’), (‘be’, ‘VB’), (‘complex’, ‘JJ’), (‘in’, ‘IN’), (‘that’, ‘DT’), (‘they’, ‘PRP’), (‘may’, ‘MD’), (‘be’, ‘VB’), (‘composed’, ‘VBN’), (‘of’, ‘IN’), (‘multiple’, ‘JJ’), (‘layers’, ‘NNS’)	N/A
EPIBOUND	Error Propagation Issue—misclassified Boundaries	An error in entity boundary prediction for NER, for one or both entities in the relation, resulting in error propagation to RE. Entity boundaries defined as the start and end tokens of the entity.	See ENT1BOUND or ENT2BOUND misclassification reason codebook.	N/A
EPIMISSING	Error Propagation Missing Entity	A missing entity in the entity model, for one or both entities in the relation, resulting in error propagation to RE.	See ENT1 or ENT2 misclassification reason codebook.	Missing nested entity.
LOCAL	Local Context	There is a failure to derive the semantics of the local context.	(‘a’, ‘DT’), (‘feature’, ‘NN’), (‘in’, ‘IN’), (‘a’, ‘DT’), (‘higher’, ‘JJR’), (‘dimensional’, ‘JJ’), (‘space’, ‘NN’)	N/A
EPINESTED	Error Propagation Missing Nested Entity	One or both entities in the entity pair is a nested entity OR contains a nested entity.	(‘automated’, ‘JJ’), (‘interpretation’, ‘NN’), (‘of’, ‘IN’), (‘nominal’, ‘JJ’), (‘compounds’, ‘NNS’), (‘in’, ‘IN’), (‘English’, ‘NNP’)	N/A
RELAMBIG	Semantically Ambiguous Relation	Semantics denoting the relation are ambiguous and could pertain to more than one relation type.	(‘the’, ‘DT’), (‘predicative’, ‘JJ’), (‘information’, ‘NN’), (‘associated’, ‘VBN’), (‘with’, ‘IN’), (‘nominals’, ‘NNS’)	N/A
OTHER	Other	The relation does not fit any other code and is therefore deemed out of the scope of this study.	(‘Branch’, ‘NN’), (‘and’, ‘CC’), (‘bound’, ‘NN’), (‘strategies’, ‘NNS’), (‘have’, ‘VBP’), (‘previously’, ‘RB’), (‘attempted’, ‘VBN’), (‘to’, ‘TO’), (‘curb’, ‘VB’), (‘this’, ‘DT’), (‘complexity’, ‘NN’), (‘whilst’, ‘VBZ’), (‘maintaining’, ‘VBG’), (‘global’, ‘JJ’), (‘optimality’, ‘NN’)	N/A

Appendix D

Table A4. Theme level 4 codebook—improvements.

Code Label	Definition	Example SCIERC	Exceptions
CONJ(CC) + PRONOUN (PRP($))	A conjunction (CC) and a possessive pronoun (PRP$) are present as local context between entities, indicating a 'belonging-to' relation type [PART-OF or FEATURE-OF]. The syntax should be the sole indicator of the relation and it should not rely on further context.	Example syntax 'and their' Example SCIERC ('affine-invariant', 'JJ'), ('image', 'NN'), ('patches', 'NNS'), ('and', 'CC'), ('their', 'PRP$'), ('spatial', 'JJ'), ('relationships', 'NNS')	Personal pronouns I, you, he, she, it, we, they, me, him, her, us
((IN) + (DT)) POST	A preposition (IN) and determiner (DT) are present as local context between entities, indicating a 'belonging-to' relation type [PART-OF or FEATURE-OF]. The present syntax should be the sole indicator of the relation and it should not rely on further context.	Example syntax 'like a/the' 'from a/the' 'in a/the' 'of a/the' 'from a/an' 'on a/the' Example SCIERC ('a', 'DT'), ('feature', 'NN'), ('in', 'IN'), ('a', 'DT'), ('higher', 'JJR'), ('dimensional', 'JJ'), ('space', 'NN')	A misclassified relation direction. (IN) = 'and', 'with', 'on', 'like' PART-OF or FEATURE-OF predicted as each other, or HYPONYM-OF
(IN) BETWEEN ENTITIES	A preposition (IN) is present as local context between entities, indicating a 'belonging-to' relation type [PART-OF or FEATURE-OF]. The present syntax should be the sole indicator of the relation and it should not rely on further context.	Example syntax 'in' 'as' 'from' Example SCIERC ('syntactic', 'JJ'), ('structure', 'NN'), ('from', 'IN'), ('parse-trees', 'NNS')	A misclassified relation direction. (IN) = 'and', 'with', 'on', 'like' PART-OF or FEATURE-OF predicted as each other, or HYPONYM-OF
(IN) DIRECTIONAL	A preposition (IN) and determiner (DT) are present as local context between entities, indicating a 'belonging-to' relation type [PART-OF or FEATURE-OF], AND the relation direction is misclassified. The present syntax should be the sole indicator of the relation and it should not rely on further context.	Example syntax Of indicates PART OF Example SCIERC ('images', 'NNS'), ('extracted', 'VBN'), ('from', 'IN'), ('modern', 'JJ'), ('computer', 'NN'), ('games', 'NNS'), ('.', '.')	(IN) = 'and', 'with', 'on', 'like'
CONTEXTPRE + (IN)	A preposition (IN) and optional determiner (DT) are present as local context between entities, indicating a 'belonging-to' relation type [PART-OF or FEATURE-OF], AND context prior to the first entity is required to understand the relation.	Example syntax With Example SCIERC ('is', 'VBZ'), ('an', 'DT'), ('agglutinative', 'JJ'), ('language', 'NN'), ('with', 'IN'), ('word', 'NN'), ('structures', 'NNS')	A misclassified relation direction. (IN) = 'and' PART-OF or FEATURE-OF predicted as each other, or HYPONYM-OF
COLON	The relation spans across a colon, where one entity is on the left-hand side of the colon, and the other entity is on the right-hand side of the colon.	Example SCIERC ('Amorph', 'NNP'), ('recognizes', 'VBZ'), ('NE', 'NNP'), ('items', 'NNS'), ('in', 'IN'), ('two', 'CD'), ('stages', 'NNS'), (':', ':'), ('dictionary', 'JJ'), ('lookup', 'NN'), ('and', 'CC'), ('rule', 'NN'), ('application', 'NN'), ('.', '.')	N/A

Table A4. *Cont.*

Code Label	Definition	Example SCIERC	Exceptions
PARENTHESES	One or both entities contained in the relation are between parentheses.	Example SCIERC ('-LRB-', 'VBP'), ('grammars', 'NNS'), ('with', 'IN'), ('regular', 'JJ'), ('expressions', 'NNS'), ('at', 'IN'), ('the', 'DT'), ('right', 'JJ'), ('hand', 'NN'), ('side', 'NN'), ('-RRB-', 'NN')	N/A
COREF	One or both entities within the relation are coreferences, OR the relation is not identified due to being later referred to as a coreference	Example SCIERC ('they', 'PRP'), ('may', 'MD'), ('be', 'VB'), ('composed', 'VBN'), ('of', 'IN'), ('multiple', 'JJ'), ('layers', 'NNS')	N/A
CPN	One or both entities within the relation are compound nouns, which are not predicted at all or not predicted in full, where the definition of a compound noun contains one or more nouns (NN).	Example SCIERC ('ranking', 'VBG'), ('blog', 'NN'), ('posts', 'NNS'), ('with', 'IN'), ('respect', 'NN'), ('to', 'TO'), ('their', 'PRP$'), ('relevance', 'NN'),	CPNSCI CNPSCI CNPGEN
CPNSCI	One or both entities within the relation are scientific compound nouns, which are not predicted at all or not predicted in full, where the definition of a compound noun contains one or more nouns (NN).	Example SCIERC ('decompose', 'VB'), ('the', 'DT'), ('data', 'NNS'), ('matrix', 'NN'), ('to', 'TO'), ('a', 'DT'), ('low', 'JJ'), ('rank', 'NN'), ('part', 'NN')	CPN CNPGEN CNPGEN
CNPSCI	One or both entities within the relation are scientific complex noun phrases, which are not predicted at all or not predicted in full, where the definition of a complex noun phrase contains one or more nouns (NN) AND any combination of Adjectives (JJ), prepositions and conjunctions (IN), conjunctions (CC), and verbs (VB).	Example SCIERC ('At', 'IN'), ('the', 'DT'), ('core', 'NN'), ('of', 'IN'), ('the', 'DT'), ('externally', 'JJ'), ('digital', 'JJ'), ('architecture', 'NN'), ('is', 'VBZ'), ('a', 'DT'), ('high-density', 'NN'), (',', ','), ('low-power', 'JJR'), ('analog', 'NN'), ('array', 'IN')	CNPGEN
CNPGEN	One or both entities within the relation are scientific complex noun phrases, which are not predicted at all or not predicted in full, where the definition of a complex noun phrase is as per CNPSCI.	Example SCIERC ('they', 'PRP'), ('may', 'MD'), ('be', 'VB'), ('composed', 'VBN'), ('of', 'IN'), ('multiple', 'JJ'), ('layers', 'NNS')	CNPSCI
INCORRECTCNP	One or both entities within the relation are misclassified as scientific OR general complex noun phrases, where none were present and where the definition of a complex noun phrase is as per CNPSCI.	Example SCIERC ('NIST', 'NNP'), ('sentence', 'NN'), ('boundary', 'JJ'), ('detection', 'NN'), ('task', 'NN'), ('in', 'IN'), ('speech', 'NN')	N/A
CN	One or both entities within the relation are common nouns, which are not predicted at all or not predicted in full, where the definition of a common noun is a general term for classes of things, rather than a specific term, and it can be modified by determiners or adjectives, OR a general term, which is not always classed as a named entity.	Example SCIERC ('the', 'DT'), ('data', 'NN'), ('has', 'VBZ'), ('large', 'JJ'), ('intra-class', 'JJ'), ('variations', 'NNS')	CPN CPNSCI CNPSCI CNPGEN

Table A4. *Cont.*

Code Label	Definition	Example SCIERC	Exceptions
PACKINGSUBJREL	The sentence contains multiple interrelated objects of a subject and one or more of the entities are not predicted, resulting in a failure of the subject-oriented packing strategy for the entity model, OR all entities are correctly predicted but the subject-oriented packing strategy fails to associate the interrelated objects.	Example SCIERC ('outlier', 'JJR'), ('removal', 'NN'), ('and', 'CC'), ('quality', 'NN'), ('improvement', 'NN'), ('in', 'IN'), ('stereo', 'JJ'), ('vision', 'NN')	N/A
PACKINGNEIGHENT	One or both entities within the relation are a nested entity parent or child, which are not predicted at all or not predicted in full, where the definition of a nested entity is an entity that shares the same start or end token with the parent entity.	Example SCIERC ('automated', 'JJ'), ('interpretation', 'NN'), ('of', 'IN'), ('nominal', 'JJ'), ('compounds', 'NNS'), ('in', 'IN'), ('English', 'NNP'), ('.', '.')	N/A
FEATURE OF AS PART OF	The gold relation is FEATURE OF, but the predicted relation is PART OF.	Example SCIERC ('syntactic', 'JJ'), ('structure', 'NN'), ('features', 'NNS'), ('embedded', 'VBN'), ('in', 'IN'), ('a', 'DT'), ('parse', 'NN'), ('tree', 'NN')	N/A
PART OF AS FEATURE OF	The gold relation is PART OF, but the predicted relation is FEATURE OF.	Example SCIERC ('the', 'DT'), ('layers', 'NNS'), ('of', 'IN'), ('various', 'JJ'), ('CNN', 'NNP'), ('models', 'NNS')	N/A
PART OF AS HYPONYM OF	The gold relation is PART OF, but the predicted relation is HYPONYM OF.	Example SCIERC ('prediction', 'NN'), ('techniques', 'NNS'), ('like', 'IN'), ('the', 'DT'), ('Kalman', 'NNP'), ('filter', 'NN')	N/A
SHARED KNOWLEDGE	The relation can only be derived from shared knowledge outside of any context within the sentence or the wider abstract.	Example SCIERC ('three-dimensional', 'JJ'), ('objects', 'NNS'), ('in', 'IN'), ('terms', 'NNS'), ('of', 'IN'), ('affine-invariant', 'JJ'), ('image', 'NN')	N/A
OTHER	The relation does not fit any other code and is therefore deemed out of the scope of this study.	See OTHER Semantic challenge codebook.	N/A

References

1. Santosh, T.Y.S.S.; Chakraborty, P.; Dutta, S.; Sanyal, D.K.; Das, P.P. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. In Proceedings of the EEKE@JCDL, 21—Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents, Virtual Event, 13 September 2021.
2. Yadav, P.; Pervin, N. Towards efficient navigation in digital libraries: Leveraging popularity, semantics and communities to recommend scholarly articles. *J. Inf.* **2022**, *16*, 101336. [CrossRef]
3. Jung, J.; Jung, S.; Seo, H.; Namgung, H.; Kim, S. Sequence Alignment Ensemble with a Single Neural Network for Sequence Labeling. *IEEE Access* **2022**, *10*, 73562–73570. [CrossRef]
4. Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K. Automated concatenation of embeddings for structured prediction. *arXiv* **2020**, arXiv:2010.05006.
5. Lu, X. Machine Learning for Text, by Charu, C. Aggarwal, New York, Springer, 2018. ISBN 9783319735306. XXIII+ 493 pages. *Nat. Lang. Eng.* **2022**, *28*, 541–543. [CrossRef]
6. Ye, D.; Lin, Y.; Li, P.; Sun, M. Packed levitated marker for entity and relation extraction. *arXiv* **2021**, arXiv:2109.06067.
7. Zhong, Z.; Chen, D. A frustratingly easy approach for entity and relation extraction. *arXiv* **2020**, arXiv:2010.12812.
8. Wadden, D.; Wennberg, U.; Luan, Y.; Hajishirzi, H. Entity, relation, and event extraction with contextualized span representations. *arXiv* **2019**, arXiv:1909.03546.

9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
10. Eberts, M.; Ulges, A. Span-based joint entity and relation extraction with transformer pre-training. *arXiv* **2019**, arXiv:1909.07755.
11. Radford, K.N.A.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *OpenAI*, 2018, *early access*.
12. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
13. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
14. OpenAI, R. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
15. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
16. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv* **2019**, arXiv:1903.10676.
17. Jehangir, B.; Radhakrishnan, S.; Agarwal, R. A survey on Named Entity Recognition—datasets, tools, and methodologies. *Nat. Lang. Process. J.* **2023**, *3*, 100017. [CrossRef]
18. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv* **2018**, arXiv:1808.09602.
19. Wang, A.; Liu, A.; Le, H.H.; Yokota, H. Towards effective multi-task interaction for entity-relation extraction: A unified framework with selection recurrent network. *arXiv* **2022**. [CrossRef]
20. Liu, Z.; Li, H.; Wang, H.; Liao, Y.; Liu, X.; Wu, G. A novel pipelined end-to-end relation extraction framework with entity mentions and contextual semantic representation. *Expert Syst. Appl.* **2023**, *228*, 120435. [CrossRef]
21. Goh, T.T.; Jamaludin NA, A.; Mohamed, H.; Ismail, M.N.; Chua, H.S. A Comparative Study on Part-of-Speech Taggers' Performance on Examination Questions Classification According to Bloom's Taxonomy. *J. Physics. Conf. Ser.* **2022**, *2224*, 012001. [CrossRef]