Investigating the Impact of Syntax-Enriched Transformers on Quantity Extraction in Scientific Texts

Necva Bölücü, Maciej Rybinski, Stephen Wan
CSIRO Data61
{necva.bolucu;maciek.rybinski;stephen.wan}@csiro.au

Abstract

Measurement extraction is an information extraction subtask focused on extracting quantities and their dependent entities within a given scientific text. Quantity extraction is the first and most important step in measurement extraction. Most existing approaches model the problem as a sequence-labeling task using pre-trained language models (PLMs). However, none of the existing systems have utilised explicit syntactic knowledge to extend the PLM-based modeling. We propose a syntax-enriched extension by integrating dependency tree representations as syntactic knowledge into transformer-based language models to address the task of quantity extraction. We apply our approach to a range of established transformer-based models to evaluate our approach and analyze its impact in experiments on scientific literature datasets. Our experimental results and in-depth analysis show that our approach, syntax-enriched RoBERTa, outperforms the other models, even in situations with scarce training data in the scientific domain. The results demonstrate the adaptability of the proposed model to the tasks, especially useful in low-resource scenarios.[1]

1 Introduction

Current growth rates in scientific publishing increase the interest in extracting information from scientific documents to provide scientists with improved methods for organising, indexing, and querying the vast existing literature (Nasar et al., 2018; Weston et al., 2019; Hong et al., 2021). *Information extraction* (IE) is a task enabling extracting and organising information from large amounts of data from unstructured sources. IE includes several subtasks, such as *named entity recognition* (NER), *relation extraction* (RE), and *relation classification* (RC). Properties specific to scientific documents result in IE subtasks tailored for IE in the scientific literature and applied in various domains, e.g., biomedical (Lewis et al., 2020; Zhang, 2021; Gérardin et al., 2023) or chemistry (Rocktäschel et al., 2012; Luo et al., 2018; He et al., 2020).

One such example is the subtask of extracting measurements and their contexts, as scientific research often relies on precise measurements for the reproducibility of experimental methods. The reproducibility supports extending and building on top of others' work, thus promoting scientific progress. The automatic detection of the measurements and their contexts in scientific texts is a key enabling factor for producing high-quality quantity-centric search systems for scientific literature (Liu et al., 2017; Kang et al., 2017; Kononova et al., 2019).

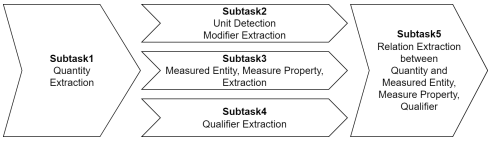


Figure 1: Subtasks of MeasEval shared task (Harper et al., 2021).

Measurement extraction (ME) is a type of IE subtask for scientific documents focused on the identification of quantities and related information and classification of relations between identified quantities and related entities (Göpfert et al., 2022). A large body of research in ME is centered around MeasEval (Harper et al., 2021), a shared task that also introduced a new annotated ME dataset consisting of scientific articles from different scholarly domains. MeasEval decomposes the ME into five finer subtasks, presented in Figure 1.

- *Subtask 1: Quantity Extraction* is the task of identifying quantities—numeric values with corresponding (optional) units of measurement and modifiers[2]. For example, in an expression 'over 5 tonnes', 5 is the numeric

[1]The code is publicly available at https://github.com/adalin16/syntax_NER.

[2]Modifiers are tokens in the quantity span that modify the

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 1–13

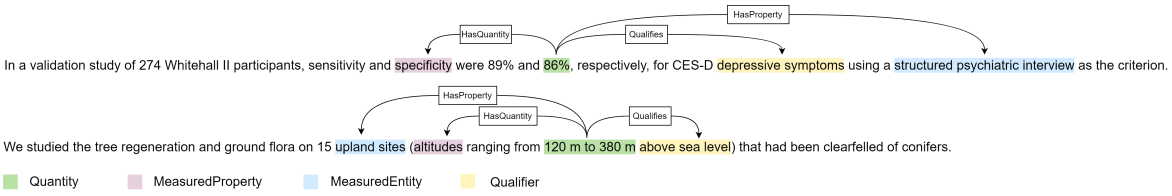


Figure 2: Sample sentences with annotation of quantity and dependent entities.

value, 'tonnes' is the unit of measurement, and 'over' is the modifier.

- *Subtask 2: Unit Detection & Modifier Extraction* has two sub-problems. Unit detection is the task of extraction of units from extracted quantities and Modifier Extraction is the task of classifying quantities into different modifiers (e.g., 'count', 'range', 'mean', etc.).

- *Subtask 3: Measured Entity (ME) & Measured Property (MP) Extraction* is the task of extracting dependent entities that elaborate the extracted quantity (e.g., ME: 'GHQ symptom caseness', 'response categories', etc., MP: 'sensitive', 'scores', 'transit depths', etc.).

- *Subtask 4: Qualifier (QUAL) Extraction* is the task of extracting dependent entities which qualify the extracted quantity (e.g., 'after 13 passages', 'orbits the planet' etc.).

- *Subtask 5: Relation Extraction* is the task of extracting relations ('has quantity', 'has property', 'qualifies') between extracted quantities and dependent entities ('measured properties', 'measured entities', 'qualifiers') and their relations to the extracted quantities.

Here, we focus on the first subtask—quantity extraction—which is required for the other subtasks: its results are directly used for subtasks 2, 3, and 4. Finally, the results of subtask 1 and 4 are used for subtask 5. This highlights the importance of quantity extraction to the overall success of the ME models, as errors incurred at this stage are propagated downstream (Göpfert et al., 2022). Sample sentences for quantities and dependent entities are given in Figure 2.

Existing methods for quantity extraction model the problem as a sequence labeling task and usually fine-tune pre-trained language models (PLMs) (Davletov et al., 2021a; Gangwar et al., 2021b). However, such models do not capture some of the syntactic relations and long-range word dependencies, which have been proven to have a positive impact on natural language understanding (Du et al., 2021). So far, the integration of linguistic knowledge and graph structures into transformer-based PLMs has been proposed for various natural language processing (NLP) problems (e.g., *Machine Translation* (MT) (Bugliarello and Okazaki, 2019; Akoury et al., 2019), *Semantic Textual Similarity* (STS) (Peng et al., 2021)), but not for quantity extraction.

Here, we propose to improve the self-attention mechanism of PLMs to incorporate syntactic information for quantity extraction – Syntax-Enriched Quantity Extraction (SEQE) (§3.2)[3]. Similar to previous studies that used dependency tree representation as syntactic information (Bugliarello and Okazaki, 2019; Guo et al., 2021), we use the dependency tree representation of the input sentence to generate syntax-enriched local attention of the PLM encoder, which provides structural information representing human understanding of the text. Since there are numerous PLMs pre-trained on different NLP data and the size of these models varies in terms of the number of parameters, we test our proposed model SEQE with different PLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and LUKE (Yamada et al., 2020) (see §4). Our method is simple yet effective, improves the task of quantity extraction, and achieves performance gains over baseline PLMs.

Overall, we provide a detailed analysis with prediction interpretation and error analysis pointing to future research directions in measurement extraction (see §5).

2 Related Work

Quantity Extraction In the literature, quantity extraction is often solved as a sequence label-

meaning of the quantity, for example, 'greater than', 'over', 'fewer than'.

[3]"Syntax-enriched" and "syntax-aware" are used interchangeably in the literature implying integration of syntactic information into the systems.

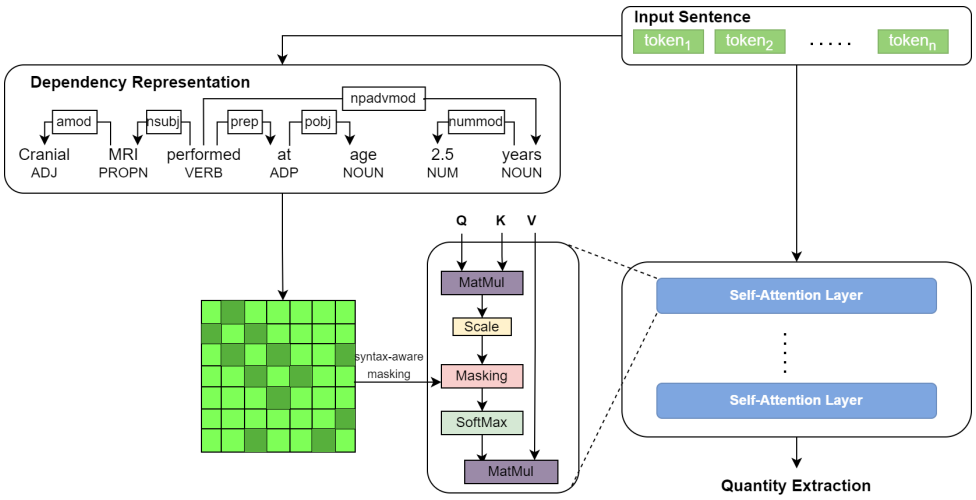


Figure 3: The overall Architecture of SEQE. Note that the syntax mask is generated from the dependency tree representation of the input, where m=1 is used for the sample sentence and the dark green color in the mask represents the value '1' and the light-green color represents the value '0'.)

ing problem using several methods, such as Conditional Random Field (CRF) (Foppiano et al., 2019), Bidirectional Long Short-Term Memory (BiLSTM) (Huang et al., 2015), transformer-based pre-trained language models (PLMs) with fine-tuning (Davletov et al., 2021b; Cao et al., 2021). Most of the systems submitted to the MeasEval shared task use PLMs for the problem. Davletov et al. (2021b) fine-tune LUKE NER model (Yamada et al., 2020) for quantity extraction as sequence labeling problem. Cao et al. (2021) apply a cascaded approach, extracting quantities via RoBERTa (Liu et al., 2019) encoder with an ensembling of PointerNet (Vinyals et al., 2015) and a CRF layers on top of the encoder. Gangwar et al. (2021a) extract quantities using SciBERT with a CRF layer for the sequence labeling problem (SciBERT (Beltagy et al., 2019) is another BERT variant pre-trained on papers from the scientific corpus (semanticscholar.org)). Karia et al. (2021) use a similar approach with BioBERT (Lee et al., 2020)—a BERT variant pre-trained on a biomedical corpus from a BERT checkpoint.

Syntax-Enriched Models Recently, models that integrate syntactic information—so-called syntax-enriched models—have been applied to various NLP problems, such as machine translation (Bastings et al., 2017; Nguyen et al., 2020), semantic role labeling (Wang et al., 2019; Marcheggiani and Titov, 2019), and question answering (Schlichtkrull et al., 2020). These models have gained attention due to their enhanced ability to capture information over long distances, especially between discontinuous constituents (Wang and Li, 2022). In contrast to these models, we incorporate the syntactic information using a distance-based masking approach and use it to alter the activation propagation in the attention heads of PLMs to improve the quantity extraction task. There are also studies that integrate syntactic information into the attention mechanisms of transformer-based models such as LISA (Linguistically-Informed Self-Attention) (Strubell et al., 2018) and Syntax-BERT (Bai et al., 2021). These models inject syntactic information by using only syntactic parents of tokens as masks to the one attention head (Strubell et al., 2018), or by generating 3 masks (parent, child, and sibling masks) from the syntax tree and injecting them into the attention mechanism of PLMS by utilising topical attention layer to aggregate task-oriented representations. Both of these approaches are different from the method proposed in this paper.

Although there is no attempt in the literature to extract quantities using syntactic information, there are studies that show promise in using syntactic information for RE (Tian et al., 2021, 2022; Sun and Grishman, 2022) and NER (Aguilar and Solorio, 2019; Nie et al., 2020; Xiong et al., 2022). However, these approaches do not integrate syntactic information in the attention-level of transformer-based PLMs.

3 Method

In this section, we present the proposed model that exploits syntactic information for quantity extraction. We base our model on the architecture of Transformer (Vaswani et al., 2017) and integrate syntactic information into the encoder with a syntax-enriched local attention mechanism for quantity extraction task. This method allows to incorporate syntactical constraints and long-range syntactic word dependencies into the sentence with syntactic representation without external information for the problem.

First, we describe the self-attention mechanism in Section 3.1. Then, we introduce the syntax-enriched quantity extraction model (SEQE) in Section 3.2.

3.1 Preliminaries

Self-Attention Transformer architecture introduced by Vaswani et al. (2017), has become ubiquitous in modern NLP, as it offers significant effectiveness improvements on many problems. The transformer consists of encoder-decoder blocks and uses stacked self-attention to encode contextual information for input tokens in which three components of queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are learned during training.

Attention is described as a mapping between $\mathbf{Q}$, and ($\mathbf{K}$, $\mathbf{V}$) pairs to obtain an output vector. We describe the simplest form, single-head attention $\mathbf{A}$ which is computed using the *scalar-dot product* between a query and the keys, followed by its softmax to obtain the weights of values:

$$\mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \quad (1)$$

where d is the dimension of keys which is used as a scaling factor in the equation. We note that, in practice, the attention matrix is a series of such attention heads, called multi-head attention, given by $MultiHead(Q, K, V) = concat(head_1, \cdots, head_h)W^O$.

3.2 Syntax-Enriched Quantity Extraction

As mentioned earlier, one limitation of PLMs is that they take a sequence of tokens as input without explicitly incorporating structural information. Some previous works have tried to induce syntactic structure into the self-attention layer (Strubell et al., 2018; Bai et al., 2021). Syntax-Enriched Quantity Extraction (SEQE) is designed to incorporate syntactic information in the self-attention layer of transformer-based PLM for quantity extraction task. The overall architecture of the proposed model is illustrated in Figure 3. As shown in the figure, we generate a syntax mask for the input sentence in a preprocessing step: (1) the dependency tree representation of the input sentence is generated by an external parser, (2) the dependency matrix is extracted from the dependency tree representation given as a graph $G = (V, E, X)$, where V is the set of nodes (skipping ROOT node), E is the set of labeled edges representing dependency relations (without labels), and X is the set of tokens of the sentence. Each token x_i is mapped to a node v_i and the distance, from node v_i to v_j is denoted by $dis(v_i, v_j)$) and $D(i, j) = min\ dis(v_k, v_j),\ k \in [i-1, i+2]$. (3) syntax mask is generated using a dependency matrix as follows:

$$\mathbf{M}_{ij} = \begin{cases} 0 & D(i,j) \leq m \\ -\infty & otherwise, \end{cases} \quad (2)$$

where m is a distance threshold hyperparameter for syntax mask that needs to be fine-tuned.

Next, the sentence is embedded similarly to a standard PLM and given as input to the self-attention layer with a syntax-enriched local attention mechanism. Syntax-enriched local attention, where tokens can attend to other tokens if they are close in the dependency tree representation (m), is computed as follows for a given query $\mathbf{Q}$ and key $\mathbf{K}$:

$$softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M})\mathbf{V} \quad (3)$$

4 Experiments

4.1 Task

Quantity Extraction task is based on the extraction of quantities $q_1, \cdots, q_m$ from a given sentence $s = w_1, \cdots, w_n$ where a quantity q_i is a sequence of words. The problem can be formulated as a token-level classification task in which the model takes a set of input-output pairs $Z = \{(w_1, y_1), \cdots (w_n, y_n)\}$ and try to classify using a function $f : X \rightarrow R$ that maps given words into a set of labels $y \in Y$ (B-Quantity, I-Quantity, O), BIO tags for NER problem.

4.2 Datasets

We use two English datasets for the quantity extraction task:

- **MeasEval**[4] (Harper et al., 2021) dataset contains 110 articles from 10 different subject areas.
- **Grobid (GeneRation Of BIbliographic Data)**[5] (Foppiano et al., 2019) dataset is composed of 32 scientific publications and 3 patents, a total 35 documents, collected across different domains and annotated for quantity and unit extraction.

Table 1 reports the statistics of the datasets.

Dataset	Train	Valid	Test	Avg l
MeasEval	1,284	427	755	8.37
Grobid	5,669	-	1,285	8.68

Table 1: Number of sentences in each dataset with avg l which denotes the average length of quantities

4.3 Evaluations

Our method, SEQE, is an extension of PLMs. For this reason, we use base versions of the PLMs and LISA (Strubell et al., 2018)[6] as the baselines to compare our model against. Conceptually, LISA (also an 'add-on' to other PLMs) is the closest method to SEQE. In LISA syntactic information is injected into only a single attention head, where each token is attending only to its syntactic parent. We run experiments on both datasets, with all models fine-tuned on training subsets. We use the following PLMs in the experiments: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and LUKE (Yamada et al., 2020). We use two variants of each PLM, where the '-base' variant consists of 12 layers, 12 attention heads, and 768 hidden dimensions, while the '-large' variant has 24 layers, 16 attention heads, and 1024 hidden dimensions.

For both experiments, in addition to the baseline PLMs (baseline models), we also compare our results with state-of-the-art models: LIORI (Davletov et al., 2021a) and Grobid (Foppiano et al., 2019).

Evaluation Metric As an evaluation metric, in addition to the token-level macro F_1 score, we also used the macro F_1 score from Seqeval (Nakayama, 2018), span-level evaluation metric, since we try to solve quantity extraction problem as a sequence labeling problem and the important label is only quantity.

[4] https://github.com/harperco/MeasEval

[5] https://github.com/kermitt2/grobid-quantities

[6] https://github.com/strubell/LISA

4.4 Experimental setup

We utilise Hugging Face[7] library for the baseline experiments which are fine-tuning PLMs. We fine-tune the baseline BERT model using Optuna (Akiba et al., 2019), a hyperparameter optimization framework, and apply the same hyperparameters for other PLMs (batch size of 32, max length of 128, the learning rate of 1e-5 and 10 epoch of training). For the proposed model experiments, we extract dependency tree representations from the texts utilising an external deep biaffine dependency parser (Dozat and Manning, 2016)[8] integrated into the SpaCy library[9] (Honnibal and Montani, 2017). We use the English model `en_core_web_sm` of SpaCy in the experiments. Since the nodes in the dependency tree representation are words, in the attention mechanism of SEQE we apply the same masking value (that would have corresponded to the full word) to the sub-word tokens produced by specific tokenisers (WordPiece, byte-level BPE). We finetune the syntax-enriched BERT model using Optuna and apply the same hyperparameters for other syntax-enriched PLMs (distance threshold of 3, batch size of 8, learning rate of 5e-5 and 5 epoch of training). We train all experiments on a single NVIDIA Quadro RTX 5000 GPU.

We train each model five times with different random seeds and report the mean and standard deviation of the results to account for the training variance of the models.

Statistical significance The statistical significance of the differences in macro F_1 score is evaluated with an approximate randomization test (Chinchor, 1992) with $99,999$ iterations and significance level $\alpha = 0.05$ for each baseline PLM and its syntax-enriched version (e.g., BERT $\rightarrow$ Syntax-enriched BERT). For significance testing, we used outputs yielding the 3^{rd}-best results for each of the models (so, a median from the 5 runs reported to account for variance).

5 Results and Discussion

5.1 Main Results

Experimental results are shown in Table 2 and 3 for the base and large models, respectively. We report the results on the test sets of MeasEval and

[7] https://huggingface.co/

[8] The parser achieves 95.7% UAS and 94.1% LAS on the most popular English PTB dataset (Marcus et al., 1993).

[9] https://spacy.io/

		MeasEval		Grobid	
Models	**Params**	**Macro F_1**	**Seq F_1**	**Macro F_1**	**Seq F_1**
Base Models					
BERT (Devlin et al., 2019)	110M	87.26±1.66	57.15±7.24	89.45±1.42	72.28± 6.45
+ LISA		89.45±1.15	68.41± 5.89	89.51±1.31	73.47± 6.18
+ SEQE (Ours)	+ 0.01M	92.38$^{\uparrow}$±1.42	74.17$^{\uparrow}$±6.45	93.45$^{\uparrow}$±1.54	78.36$^{\uparrow}$±6.58
SciBERT (Beltagy et al., 2019)	110M	88.78±1.43	60.41±4.86	90.32±1.25	74.57±5.14
+ LISA		90.18±1.52	67.11±3.52	89.47±1.51	76.25±4.18
+ SEQE (Ours)	+ 0.01M	92.32$^{\uparrow}$±1.30	73.98$^{\uparrow}$±2.36	83.38±1.26	79.22$^{\uparrow}$±3.14
RoBERTa (Liu et al., 2019)	125M	89.63±1.33	65.62±5.54	91.24±1.32	75.42±6.21
+ LISA		90.17±1.25	66.54±5.10	90.89±1.42	75.10±5.89
+ SEQE (Ours)	+ 0.01M	90.58$^{\uparrow}$±1.42	69.05$^{\uparrow}$±4.41	91.25±1.48	75.61±4.48
LUKE (Yamada et al., 2020)	253M	91.22±0.79	72.66±5.06	92.22±0.88	77.68± 4.02
+ LISA		90.23±1.11	73.56±4.99	91.17±1.05	77.15±4.45
+ SEQE (Ours)	+ 0.01M	90.89±1.02	74.57±5.03	91.77±1.11	79.55±5.18

Table 2: Base PLM results on quantity extraction datasets. ↑ means statistically significant improvement over the corresponding baseline PLM. Reported results are averaged over 5 runs.

		MeasEval		Grobid	
Models	**Params**	**Macro F_1**	**Seq F_1**	**Macro F_1**	**Seq F_1**
State-of-the-art Models					
LIORI (Davletov et al., 2021b)	-	90.85	75.13	92.46	76.19
Grobid (Foppiano et al., 2019)	-	86.13	65.16	80.14	54.92
Large Models					
BERT (Devlin et al., 2019)	340M	87.07±1.68	57.75±4.78	88.95±1.54	72.36±5.04
+ LISA		90.45±1.51	68.48±4.15	90.36±1.51	74.47±4.25
+ SEQE (Ours)	+ 0.02M	91.88$^{\uparrow}$±1.42	72.762$^{\uparrow}$±3.78	92.982$^{\uparrow}$±1.50	76.95$^{\uparrow}$±4.11
RoBERTa (Liu et al., 2019)	355M	91.74±0.39	77.01±3.33	93.57±1.32	78.63±4.15
+ LISA		91.18±0.56	76.43±3.14	94.01±1.17	78.44±4.16
+ SEQE (Ours)	+ 0.02M	**92.49**$^{\uparrow}$**±0.78**	77.75$^{\uparrow}$±2.85	**94.28**$^{\uparrow}$**±0.82**	78.52±3.03
LUKE (Yamada et al., 2020)	483M	91.16±0.40	76.22±0.71	93.55±0.52	77.87±1.18
+ LISA		90.89±0.51	76.48±0.69	93.10±0.71	78.15±1.11
+ SEQE (Ours)	+ 0.02M	91.14±0.67	**77.89**±0.73	93.48±0.72	**79.83**$^{\uparrow}$**±1.21**

Table 3: Large PLM results on quantity extraction datasets. ↑ means statistically significant improvement over the corresponding baseline PLM. Reported results are averaged over 5 runs.

Grobid datasets. The results show that the proposed SEQE method achieves consistent gains over the baseline PLMs and LISA for the quantity extraction task, especially for BERT. Even though the baseline RoBERTa performs best among all the baseline models, it shows that the quantity extraction task benefits from injecting syntactic information into the PLMs. The proposed approach outperforms LISA and among the experiments of syntax-enriched PLMs, syntax-enriched RoBERTa achieves the highest score and outperforms baseline RoBERTa with an increase of 0.75 and 0.71 in the Macro F_1 score for the MeasEval and Grobid datasets, respectively. Syntactic information does not result in a notable improvement for LUKE, which is a word- and entity-level model (pre-trained with a large amount of entity-annotated corpus) using entity-aware attention mechanism. SEQE decreased the Macro F_1 score for LUKE-large. However, we obtain the highest span-level Macro F_1 with the syntax-enriched LUKE-large, which mainly shows the weakness of word-level models for this evaluation metric. Importantly, syntax-enriched PLMs with fewer parameters (BERT, SciBERT) outperform their large baseline counterpart PLMs (Wang and Wang, 2020; Yang et al., 2020), showing the importance of syntactic information to the small models.

		Target	
		MeasEval	**Grobid**
Source	**MeasEval**	92.49±0.78	90.45±1.45
	Grobid	90.17±0.95	94.28±0.82

Table 4: Token-level Macro F_1 scores of RoBERTa (large) + SEQE for cross-domain experiments.

5.2 Cross-Domain Results

Cross-domain NER focuses on transferring from a source domain to a target domain. We run

		Predicted		
		B-Q	**I-Q**	**O**
True	**B-Q**	476	31	59
	I-Q	43	889	78
	O	24	42	8403

(a) MeasEval dataset

		Predicted		
		B-Q	**I-Q**	**O**
True	**B-Q**	390	53	99
	I-Q	12	912	49
	O	40	123	7340

(b) Grobid dataset

Table 5: Confusion matrix for the syntax-enriched RoBERTa (large) for quantity extraction task. (B-Quantity (B-Q), I-Quantity (I-Q))

cross-domain experiments with syntax-enriched RoBERTa (large) yielding the best token-level Macro F_1 scores on MeasEval and Grobid datasets. Cross-domain experimental results are shown in Table 4.

When we compare the within-domain and cross-domain results, we observe a slight decrease for both datasets. The macro F_1 scores for the within-domain experiments for MeasEval and Grobid are 92.49% (MeasEval $\rightarrow$ MeasEval)and 94.28% (Grobid $\rightarrow$ Grobid), respectively, while for the cross-domain experiments they are 90.17% (Grobid $\rightarrow$ MeasEval) and 90.45% (MeasEval $\rightarrow$ Grobid). Despite the effectiveness decrease, the results are still comparable to those of the baseline models.

5.3 Error Analysis

In Table 5, we show the confusion matrices for the predictions of the model with the best results (syntax-enriched RoBERTa) for the MeasEval (Table 5a) and Grobid (Table 5b) datasets. Typically, the model does not confuse the Quantity tags (B-Quantity, I-Quantity), but instead makes errors in deciding whether a token is a quantity or not. This makes sense, since the number of O tags is higher than the number of Quantity tags. We perform a comprehensive analysis of the errors made by the models to understand which quantity formats the model performs well on, and which it performs poorly on.

We categorise the prediction errors made by the model by exploring the properties of individual tokens for which the model made incorrect predictions for each of the datasets. For MeasEval, of the 24 tags for which the model confused an O tag for a B-Quantity, 7 are punctuation characters and 7 are numbers written as numeric or alphabetic, and the others are modifiers for quantities that occur frequently in the datasets, such as *approximately, low*, etc. Out of the 42 tags where the model confused an O tag for an I-Quantity, 10 are units, and 6 are numbers written as numeric or alphabetic.

For the Grobid dataset, of the 40 tags where the model confused a O tag for a B-Quantity, 7 are numbers written as numeric or alphabetic, and 10 are punctuations. Interestingly, 10 of the mislabeled tokens are units, such as *m, %*. Out of the 123 tags where the model confused a O tag for an I-Quantity, 16 are units, 24 are numbers written as numeric or alphabetic and 33 are punctuations.

After analyzing all the errors made by the models, we found that the syntax-enriched model tends to find longer quantity spans compared to the baseline PLMs. The common errors made by both models can be divided into 3 categories: (1) labeling modifier words as O (e.g., range, between), (2) labeling numbers written as numeric or alphabetic as B-Quantity, (3) labeling stop words in quantities as O (e.g., a, the).

5.4 Discussion

Based on the results, we analyse the impact of the syntax-enriched attention mechanism on the problem by visualising the model's decision. For this purpose, we used the transformers-interpret[10], a post-hoc explanation tool compatible with models from the transformers package designed for the sequence labeling problem. Tokens are assigned an importance score indicating how their presence contributes to the prediction of a particular positive token (Attribution Label) with the cumulative importance scores (Attribution score) for that token. Tokens highlighted in green have a positive contribution to the model's decision, while tokens highlighted in red have a negative contribution.

We randomly select a few sentences from the test set and analyze the predictions of the best-performing model (syntax-enriched RoBERTa-large) and its baseline version (RoBERTa-large)[11]. Figure 4 shows the visualisation of the models for the sentence *"In addition, the number of emerged Striga plants for each plot was recorded at 67, 101 and 121 das."* with the quantity *"67, 101 and 121 das"*. While the baseline model correctly predicts

[10] https://github.com/cdpierse/transformers-interpret

[11] Dependency tree representations of the sentences are given in Figure 6.

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-Quantity	B-Quantity (0.99)	67	2.92	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.99)	,	1.53	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (1.00)	101	2.34	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.99)	and	2.13	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.99)	121	1.45	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.99)	d	1.64	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.98)	as	1.26	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s

(a) RoBERTa-large

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-Quantity	B-Quantity (0.94)	67	1.97	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.77)	,	1.42	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	B-Quantity (0.79)	101	0.82	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.75)	and	1.77	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	B-Quantity (0.69)	121	1.06	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.87)	d	2.13	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s
I-Quantity	I-Quantity (0.56)	as	0.84	#s In addition , the number of emerged Str iga plants for each plot was recorded at 67 , 101 and 121 d as . #/s

(b) Syntax-enriched RoBERTa-large

Figure 4: Visualisation of the sentence *"In addition, the number of emerged Striga plants for each plot was recorded at 67, 101 and 121 das."*

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-Quantity	B-Quantity (0.99)	>	1.63	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s
I-Quantity	B-Quantity (0.96)	10	3.24	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s
I-Quantity	I-Quantity (0.48)	%	1.80	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s

(a) RoBERTa-large

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
B-Quantity	B-Quantity (0.92)	>	2.49	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s
I-Quantity	B-Quantity (0.87)	10	1.41	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s
I-Quantity	I-Quantity (0.75)	%	2.40	#s The colored letters indicate a comparatively high expression level of the M HC - I allele , comprising > 10 % of c DNA sequence reads . #/s

(b) Syntax-enriched RoBERTa-large

Figure 5: Visualisation of the sentence *"The colored letters indicate a comparatively high expression level of the MHC-I allele, comprising >10% of cDNA sequence reads."*

the quantity, lots of tokens have positive and negative effects on the prediction of token labels, especially some distant tokens (e.g., the word 'addition'). In the syntax-enriched model, on the other hand, the contributing tokens are closer together, due to dependency relations extracted from the sentence's dependency tree and incorporated in the attention mechanism. In particular, the syntax-enriched model appears to base its decision on the positive contribution of a predicate syntactically close the quantity span (here, 'recorded').

We observe similar results in Figure 5 for the sentence *"The colored letters indicate a comparatively high expression level of the MHC-I allele, comprising >10% of cDNA sequence reads."* with the quantity *">10%"*. Since numbers written as numeric or alphabetic are usually placed at the beginning of quantities, both models tend to label 10 as B-Quantity. Apart from this result, we see that close tokens have a positive effect in predicting token labels for the syntax-enriched model.

6 Conclusion

We introduce the SEQE model that integrates syntactic information into the Transformer attention mechanism to provide a complementary structure for the quantity extraction modeled as a sequence labeling problem. We demonstrate the effectiveness of the proposed SEQE model, which uses syntactic information, by comparing it to baseline

PLMs on the quantity extraction task. We find that the proposed method outperforms the baseline PLMs and SOTA models and the syntax-enriched RoBERTa achieves the best effectiveness among all evaluated methods. We also find that syntactic information added at the attention-level of the PLMs contributes to more accurate entity span extraction, which is also very important for other (downstream) subtasks of ME, as these other subtasks depend directly on the quality of quantity extraction. Finally, the SEQE model is versatile in a sense that it can be easily integrated into all tasks that use pre-trained transformer models.

In future work, we will explore the performance of the transformer models extended using semantic representations such as AMR (Banarescu et al., 2013), UMR (Van Gysel et al., 2021), and UCCA (Abend and Rappoport, 2013).

Our work aims to explicitly extract quantity extraction using linguistic knowledge as syntactic information integrated into the attention mechanism of the PLMs encoder. We focus on autoencoding models (BERT, RoBERTa, LUKE) that rely on the encoder part of the original transformer. However, autoregressive models (e.g., GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019)) and seq2seq models (e.g., BART (Lewis et al., 2019), T5 (Raffel et al., 2020)) are widely used in the literature for the token classification problem. In addition, non-autoregressive models (Gu et al., 2017) have become popular due to their fast inference speed, as they omit the sequential dependencies in inference. We hope to extend our study on syntax-enriched masking for quantity extraction to these models.

Finally, we will investigate the impact of our approach on downstream subtasks of ME defined in the MeasEval shared task (Harper et al., 2021).

Limitations

Even though our proposed model outperforms the baselines, there are still limitations, mainly based on the syntax-enriched mask integrated into PLMs. We utilised dependency tree representations in the syntax-enriched attention mechanism. Although the labels of the dependency arcs give the syntax type of the relation between the connected words, we ignore the arc labels to keep the masking simple. In addition, our model depends on the effectiveness of the dependency parser model used 'off-the-self' in our method.

Ethical Statement

The datasets used in our experiments are publicly available. Both these datasets are focused on processing (publicly available) scientific literature, thus constituting a low-risk setting.

References

Omri Abend and Ari Rappoport. 2013. UCCA: A Semantics-based Grammatical Annotation Scheme. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 1–12.

Gustavo Aguilar and Thamar Solorio. 2019. Dependency-aware named entity recognition with relative and global attentions. *arXiv preprint arXiv:1909.05166*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. 2019. Syntactically supervised transformers for faster neural machine translation. *arXiv preprint arXiv:1906.02780*.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. *CoRR*, abs/1704.04675.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Emanuele Bugliarello and Naoaki Okazaki. 2019. Enhancing machine translation with dependency-aware self-attention. *arXiv preprint arXiv:1909.03149*.

Jiarun Cao, Yuejia Xiang, Yunyan Zhang, Zhiyuan Qi, Xi Chen, and Yefeng Zheng. 2021. CONNER: a cascade count and measurement extraction tool for scientific discourse. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1239–1244.

Nancy Chinchor. 1992. The statistical significance of the muc-4 results. In *Proceedings of the 4th Conference on Message Understanding, MUC 1992*, pages 30–50.

Adis Davletov, Nikolay Arefyev, Denis Gordeev, and Alexey Rey. 2021a. LIORI at SemEval-2021 task 2: Span prediction and binary classification approaches to word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 780–786, Online. Association for Computational Linguistics.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021b. LIORI at SemEval-2021 task 8: Ask transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Chunning Du, Jingyu Wang, Haifeng Sun, Qi Qi, and Jianxin Liao. 2021. Syntax-type-aware graph convolutional networks for natural language understanding. *Applied Soft Computing*, 102:107080.

Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019. Automatic identification and normalisation of physical measurements in scientific literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021a. Counts@ iitk at semeval-2021 task 8: Scibert based entity and semantic relation extraction for scientific data. *arXiv preprint arXiv:2104.01364*.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021b. Counts@IITK at SemEval-2021 task 8: SciBERT based entity and semantic relation extraction for scientific data. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1232–1238, Online. Association for Computational Linguistics.

Christel Gérardin, Yuhan Xiong, Perceval Wajsbürt, Fabrice Carrat, and Xavier Tannier. 2023. Impact of translation on biomedical information extraction from real-life clinical notes.

Jan Göpfert, Patrick Kuckertz, Jann Weinand, Leander Kotzur, and Detlef Stolten. 2022. Measurement Extraction with Natural Language Processing: A Review. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2191–2215.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Qipeng Guo, Xipeng Qiu, Xiangyang Xue, and Zheng Zhang. 2021. Syntax-guided text generation via graph neural network. *Science China Information Sciences*, 64(5):1–10.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr, and Paul Groth. 2021. Semeval-2021 task 8: Measeval–extracting counts and measurements and their related contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316.

Jiayuan He, Dat Quoc Nguyen, Saber A Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, et al. 2020. Overview of ChEMU 2020: named entity recognition and event extraction of chemical reactions from patents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 237–254. Springer.

Zhi Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. 2021. Challenges and advances in information extraction from scientific literature: a review. *JOM*, 73(11):3383–3400.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071.

Neel Karia, Ayush Kaushal, and Faraaz Mallick. 2021. KGP at SemEval-2021 Task 8: Leveraging Multi-Staged Language Models for Extracting Measurements, their Attributes and Relations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 387–396.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):203.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Sijia Liu, Liwei Wang, Donna Ihrke, Vipin Chaudhary, Cui Tao, Chunhua Weng, and Hongfang Liu. 2017. Correlating lab test results in clinical notes with structured lab data: a case study in HbA1c and glucose. *AMIA Summits on Translational Science Proceedings*, 2017:221.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Diego Marcheggiani and Ivan Titov. 2019. Graph Convolutions over Constituent Trees for Syntax-Aware Semantic Role Labeling. *CoRR*, abs/1909.09814.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117:1931–1990.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2020. Differentiable Window for Dynamic Local Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6589–6599.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving named entity recognition with attentive ensemble of syntactic information. *arXiv preprint arXiv:2010.15466*.

Qiwei Peng, David Weir, and Julie Weeds. 2021. Structure-aware sentence encoder in bert-based Siamese network. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 57–63.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. 2020. Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking. *CoRR*, abs/2010.00577.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*.

Huiyu Sun and Ralph Grishman. 2022. Employing Lexicalized Dependency Paths for Active Learning of Relation Extraction. *Intelligent Automation & Soft Computing*, 34(3).

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. Dependency-driven relation extraction with attentive graph convolutional networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4458–4471.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Improving Relation Extraction through Syntax-induced Pre-training with Dependency Masking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1875–1886.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3-4):343–360.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Chang Wang and Bang Wang. 2020. Encoding sentences with a syntax-aware self-attention neural network for emotion distribution prediction. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 256–266. Springer.

Haitao Wang and Fangbing Li. 2022. A text classification method based on LSTM and graph attention network. *Connection Science*, 34(1):2466–2480.

Yufei Wang, Mark Johnson, Stephen Wan, Yifang Sun, and Wei Wang. 2019. How to best use syntax in semantic role labelling. *arXiv preprint arXiv:1906.00266*.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.

Ying Xiong, Hao Peng, Yang Xiang, Ka-Chun Wong, Qingcai Chen, Jun Yan, and Buzhou Tang. 2022. Leveraging Multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network. *Journal of Biomedical Informatics*, 128:104035.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Baosong Yang, Derek F Wong, Lidia S Chao, and Min Zhang. 2020. Improving tree-based neural machine translation with dynamic lexicalized dependency encoding. *Knowledge-Based Systems*, 188:105042.

Linlin Zhang. 2021. ZJU's IWSLT 2021 speech translation system. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 144–148, Bangkok, Thailand (online). Association for Computational Linguistics.

A Multilingual PLMs

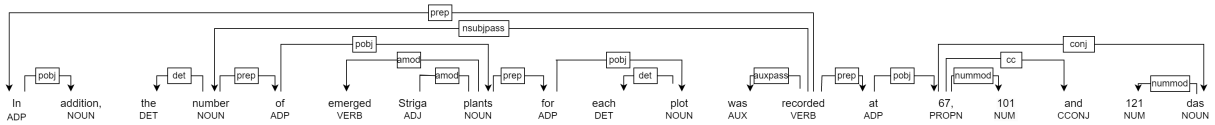
We primarily use monolingual PLMs for our experiments. However, syntax-enriched multilingual PLMs are applied to various tasks. Therefore, we perform experiments with multilingual PLMs: XLM (Cao et al., 2021) and the multilingual version of LUKE (Ri et al., 2022). The results are given in Table 6. We observe that the improvements of quantity extraction with multilingual PLMs are relatively smaller than with monolingual PLMs.

B Syntactic Representation

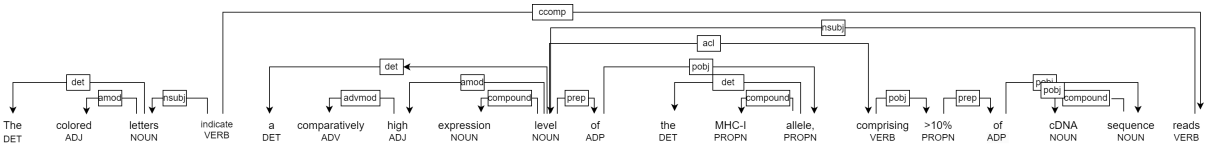
Figure 6 shows the dependency tree visualisation of sentences given in Section 5.4.

		MeasEval		Grobid	
Base-size Baseline Models					
XLM	125M	88.58±0.82	61.34±5.18	89.57±0.97	73.18±5.89
mLUKE	585M	88.98±0.72	62.59±3.21	88.75±0.75	73.61±3.18
Large-size Baseline Models					
XLM	355M	89.37±0.79	67.22±2.96	90.20±0.75	75.69± 3.47
mLUKE	868M	88.83±0.66	63.81±3.68	87.94±0.44	74.15±3.08
Syntax-Enriched Base-size Models					
XLM	125M + 0.01M	89.45±1.15	68.32±4.25	90.03±0.98	74.66±4.67
MLUKE	585M + 0.01M	87.55±0.82	62.05±2.36	87.16±0.78	73.18 ±1.94
Syntax-Enriched Large-size Models					
XLM	355M + 0.02M	90.22±0.56	76.21±0.92	91.36±0.61	78.35±1.45
MLUKE	868M + 0.02M	88.62±0.65	64.56±2.45	88.03±0.66	74.31±2.51

Table 6: Multilingual PLM results on quantity extraction datasets. Reported results are averaged over 5 runs.



(a) In addition, the number of emerged Striga plants for each plot was recorded at 67, 101 and 121 das.



(b) The colored letters indicate a comparatively high expression level of the MHC-I allele, comprising >10% of cDNA sequence reads.

Figure 6: Dependency tree visualisation of sentences given in Section 5.4