Heliyon 9 (2023) e20505

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

# Transformer-based approach to variable typing

Charles Arthel Rey, Jose Lorenzo Danguilan, Karl Patrick Mendoza, Miguel Francisco Remolona *

*Chemical Engineering Intelligence Learning Laboratory, Department of Chemical Engineering, University of the Philippines Diliman, Quezon City, 1101 Philippines*

ARTICLE INFO

*Keywords:*
Natural language processing
Transformers
Entity recognition
Relation extraction
Variable typing
Machine learning
Mathematical knowledge

ABSTRACT

The upsurge of multifarious endeavors across scientific fields propelled Big Data in the scientific domain. Despite the advancements in management systems, researchers find that mathematical knowledge remains one of the most challenging to manage due to the latter's inherent heterogeneity. One novel recourse being explored is variable typing where current works remain preliminary and, thus, provide a wide room for contribution. In this study, a primordial attempt to implement the end-to-end Entity Recognition (ER) and Relation Extraction (RE) approach to variable typing was made using the BERT (Bidirectional Encoder Representations from Transformers) model. A micro-dataset was developed for this process. According to our findings, the ER model and RE model, respectively, have Precision of 0.8142 and 0.4919, Recall of 0.7816 and 0.6030, and F1-Scores of 0.7975 and 0.5418. Despite the limited dataset, the models performed at par with values in the literature. This work also discusses the factors affecting this BERT-based approach, giving rise to suggestions for future implementations.

## 1. Introduction

The emergence of Big Data in the scientific domain is inevitable and unprecedented [1–3]. In 2016, Elsevier received approximately 1.3 million articles from about 1.8 million unique authors around the world [4] – these numbers form part of the 2.5 million articles published annually across approximately 30, 000 active scholarly peer-reviewed journals worldwide [1,5]. This trend is projected to move at a burgeoning rate of 4–5% per year. It is, therefore, clear that effective management of big scientific data must accompany its growth to ensure that meaningful insights can still be drawn from magnanimous available data without compromising the researcher's time and resources [6,7].

It is most fundamental that some sort of structure be given to data for it to be effectively managed. A solution that researchers are looking into is automated information extraction (IE) and streamlined knowledge management systems (KMS) [7–9]. As proposed by Remolona [7], cluttered data can be given structure by comprehensively storing them in Ontologies. For a management system like this to be done, data nodes and edges must be identified. In other words, specific mentions constituting an instance must be determined and related to its class or each other. This is currently done manually by researchers reading through the entire text page and then pointing out which are Variable, Definition, Equation, Magnitude, Units, and the like. It is also fundamental to relate these nodes to detail their ontological property and refine their details in the ontology tree. This becomes practical only when there are few pages considered by

* Corresponding author.
*E-mail addresses:* crrey@up.edu.ph (C.A. Rey), jadanguilan@up.edu.ph (J.L. Danguilan), kmmendoza3@up.edu.ph (K.P. Mendoza), mmremolona@up.edu.ph (M.F. Remolona).

https://doi.org/10.1016/j.heliyon.2023.e20505
Received 14 November 2022; Received in revised form 19 September 2023; Accepted 27 September 2023
Available online 29 September 2023

the researcher; however, this task becomes taxing, time-consuming, and resource-draining once a much larger corpus is marshaled. While seemingly menial, the task of separating the mentions in one column and their definition in another requires that one understands the context of the sentence and how the mentions were used therein.

One of the approaches to automatically arrange mathematical knowledge in this fashion is called Variable Typing. This is a relatively novel Natural Language Processing (NLP) task that delve into a document span to identify and relate the variables together with their semantic meaning [10]. Suppose a document below (Fig. 1) is entered into the program, the system should be able to identify that examples $m$, $z_(1)$, $z_2$, and $g$ are variables, also referred to as nodes or entities, and that the phrases “mass”, “initial elevation”, “final elevation”, “local acceleration of gravity” are contextually related to them as their respective description or mathematical type [10].

There is very limited literature available entirely delving into Variable Typing which calls for a wide room for contribution in this field; additionally, the surrounding methods can be fully compounded with the aid of existing NLP techniques which, as with other tasks, can be done either using comprehensive rules derived from semantic and lexical patterns [11,12] or through machine learning (ML) models [10].

Current methods in literature treat variable typing akin to sentiment analysis or a binary classification that terminates after determining whether there is a positive or negative edge between a variable and its type [10,13]. They are also focused on a single edge type alone which effectively excludes other mathematical types possibly present in the document [10]. These presented a gap this work was able to bridge. The contribution of this work focused on three fronts which include:

a) the information extraction and annotation of a gold-standard dataset from a Chemical Engineering textbook;
b) a primordial attempt to finetune the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) model [14] to implement a Variable Typing task; and
c) the task’s extension into a comprehensive, end-to-end entity recognition (ER) and relation extraction (RE) system.

As a primary limitation, the gold-standard annotation produced is only a micro-dataset since the annotation was only done to the first 10 chapters of the book. It is also noteworthy that manual annotation, as in this case, usually comes with a certain degree of bias on the part of the annotator. This is due to the annotator’s inherent and subjective understanding of the text. Attempts to mitigate this inherent bias were made by laying down standard annotation guidelines described in the succeeding sections, as well as reiterative consultation among annotators.

Additionally, the micro-dataset was also used only, among other models, on the BERT-based transformer because of the expansive pre-training done to it. It was also applied to the Tok2Vec model because it is the default model used by Spacy, and therefore, is a good model for comparison.

Variable Typing is a component task for a more complex information extraction (IE) system. This system effectively reallocates research resources to other more complex and insight-requiring tasks. It may also be used to automatically populate ontologies [7,15], forward the advancement of topic modeling and database enrichment [16], that consequently improve current systems in question answering, semantic web, and mathematical information retrieval (MIR).

## 2. Background

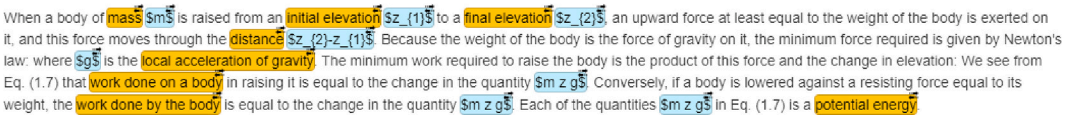
### 2.1. Information extraction framework

The “Hybrid Ontology-Learning Materials Engineering System (HOLMES)” framework of Remolona [7] lays out the foundation of an end-to-end IE for a KMS specifically ontology population for the scientific domain. It is composed of four major steps as displayed in Fig. 2.

In this research, the same steps were taken except that they were done individually and manually. In his work, he posits that the population of ontologies on “Value and Dimension, Mathematical Models, Experimental and Unit Operations, and Physical, Chemical, and Biological Reactions” requires mathematical knowledge to some degree. Remolona (2018) pointed out that for mathematical knowledge to be comprehensively stored in the ontologies, the nodes must be treated as entities falling under tags Variable, Definition, Equation, Magnitude, Units, and others. It is also fundamental to execute a relation extraction task between these entities to elaborate their ontological property and refine their details in the ontology tree.

### 2.2. The variable typing task

Grigore et al. [17] pioneered the work in Variable Typing in an attempt to disambiguate symbolic expressions in mathematical



When a body of mass $m$ is raised from an initial elevation $z_{1}$ to a final elevation $z_{2}$, an upward force at least equal to the weight of the body is exerted on it, and this force moves through the distance $z_{2}-z_{1}$. Because the weight of the body is the force of gravity on it, the minimum force required is given by Newton's law: where $g$ is the local acceleration of gravity. The minimum work required to raise the body is the product of this force and the change in elevation: We see from Eq. (1.7) that work done on a body in raising it is equal to the change in the quantity $m z g$. Conversely, if a body is lowered against a resisting force equal to its weight, the work done by the body is equal to the change in the quantity $m z g$. Each of the quantities $m z g$ in Eq. (1.7) is a potential energy.

**Fig. 1.** Example of Snipped Paragraph from the corpus with annotated entities and relations.

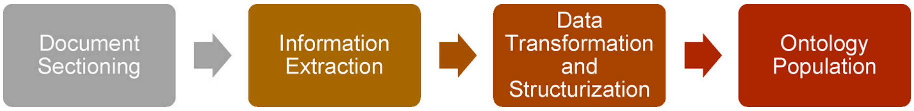


**Fig. 2.** Simplified HOLMES Framework for the unassisted information extraction and ontology population from a scientific corpus.

documents. In their study, they estimated the resemblance between the linguistic context of the section where the given ME is located, and the set of terms from the OpenMath Content Dictionaries, Cambridge Mathematics Thesaurus, and MathWorld Lexicon of Mathematical Terms. They did so by manually formulating taxonomies to assign meaning to MEs. Grigore et al. were able to satisfactorily affirm that the lexical information located immediately surrounding the expression may describe the features leading to the identification of the entities.

The instant task also includes a rules-based system first implemented by Quoc, Yokoi, Matsubayashi, and Aizawa [11]. The Concept Description Formula they developed uses this technique to template coreference relations between formulas, and their accompanying concepts from the Wikipedia Mathematics Portal. They followed a four-aspect framework which includes text processing, text matching, pattern generation, and pattern matching. Their findings demonstrate a fair accuracy performance while also offering a promising first step toward resolving coreference relationships between formulas and the surrounding words. A similar rules-based approach was coursed by an unpublished work of an internal team in the laboratory to attempt variable typing in various scientific journals. They posited that an intermediate level of linguistic skill was necessary to formulate rules that would broadly describe lexical and semantic patterns of natural language expressed in scientific discourse. These rules were underpinnings to identify which spans can ultimately build the most robust dataset.

As in this study, the researchers in the laboratory manually extracted the mathematical equations and the paragraphs from the references. They used the MathPix ® OCR system [18] by manually drawing a boundary around the text or equations under consideration. The software generated a LaTeX code with an ensuing confidence level, the LaTeX code was then fed to the self-made variable typing program in MatLab. The final version of the program was able to obtain a Precision of 75% for definitions and 83.33% for equations; meanwhile, it obtained a Recall of 75%. A rules-based approach seems to be constrained only to the extent to which the rules can be applied. It is also heavily dependent on the skill of the researchers to translate syntax into rules and codes. A remedy to this is to utilize the rules to extract features and use the same in machine learning (ML) based systems.

### 2.3. *State-of-the-art approaches to variable typing*

The state-of-the-art method of variable typing is the ML-based approach initiated by Stathopoulos et al. [10] where they annotated a mathematical dataset and fed the same to train three machine learning models namely: Nearest Type (NT), Convolutional Neural Network (CNN), and Support Vector Machine (SVM). After Stathopoulos et al. (2018), there has been a scarce number of literature directly tackling variable typing. This provides the opportunity to push the boundaries of Variable Typing further by using more advanced tools like Transformers and observing whether more advanced methods can affect significant improvements as compared to the data available in the literature. Additionally, this paves the way to extend the capability of the task to a finer granularity. This work, as an attempt to implement variable typing through the BERT Model [14], optimized bidirectional self-attention layers which were heavily derived from transfer learning architectures. The power of self-attention was first illustrated by the influential work of Vaswani et al. [19]. Eventually, this was extended by Devlin et al. [14] to remedy unidirectional language representation and brief back-propagation capabilities in the self-attention layers of Transformers. This was done by using a Masked Language Model (MLM) pre-training which fuses contexts along the left and right side of the masked token; and Next Sentence Prediction (NSP) which causes the model to understand the relationship between two sentences. As a result, BERT performs excellently on 11 NLP tasks at the token and sentential levels – this includes the RE and ER tasks. The BERT model is not limited to the mathematical domain and can be finetuned into a wide range of domains [14].

The latest study on Variable Typing using the BERT model was done by Ferreira et al. [13]. While their study utilized the same transformer model used in this work, there are fundamental differences such that their approach was an extension of the primordial work of Stathopoulos et al. [10]. In their work, given a sentence with a pre-defined set of variables V and types T, their task is a binary classification all of the edges $V \times T,$ where a positive edge denotes that the variable is assigned that type and a negative edge denotes the opposite. In this research, Variable Typing goes beyond binary classification, as the variables are not pre-defined or pre-identified because it is the model itself that shall identify the same in a given mathematical sentence. Further, the relations were not only tested as positive or negative edges but were specified.

### 2.4. *The entity recognition and relation extraction tasks*

This work's approach to Variable typing is anchored on two streamlined NLP tasks called Entity Recognition (ER) and Relation Extraction (RE). Both tasks are independent of each other in the training and evaluation phase but were concocted as if they comprise a seamless single system.

In the ER task, a span of string generally referred to as a document is fed into the system where mentions of entities are located and classified according to a predefined set of labels [20,21]. More generally, this task is called Named Entity Recognition (NER) in literature since much of the previous works dealt with noun mentions such as a person, organization, location, or similar classes. This is not a trivial task because the syntax and context of the adjacent tokens affect how well the ER model performs [21]. There have been many studies done to explore and optimize the complexity of the ER task – there were approaches based on Conditional Random Field (CRF), Temporal Convolutional Network [22], or Long Short-Term Memory (LSTM) [23]. BERT can also be fine-tuned to perform entity recognition [24].

Among the many variations of the BERT model, the BERT-BASE-CASED model is best for NER because it preserves word-shape features during Word-piece tokenization. In the paper of Devlin et al. (2019), a stunning 92.4 NER F1-score was obtained when it was tested on the CoNLL dataset. In the mathematical domain, a sole NER task was attempted by Y. Zhang et al. (2022) where they used the BERT vectors to extract features from documents using Bidirectional LSTM and Iterated Dilated Convolutional Neural Network (IDCNN). The results were then merged and corrected using CRF. The final model was then called the BERT-BiLSTM-IDCNN-CRF model.

Another crucial NLP task is Relation Extraction, whose main goal is to ascertain whether there is a relationship between two previously identified entities in a document [25,26,26,27,27–29]. Given a sentence $S = w_1, w_2, w_3, \ldots e_1, w_j \ldots e_2, \ldots w_k$, the task tests whether there is a semantic relation between the entities $e_1$ and $e_2$ mapped from a predefined set of categories.

The result of the RE task is generally called a relation triple composed of two entities and their relations tags. This, therefore, presupposes that the entities are already known, they are non-overlapping, and they belong to the same document entry.

Methods such as Perceptron, Support Vector Machines (SVM) [30], Kernel Methods [31], and log-linear models [32] were used as some of the previous approaches.

## 3. Methodology

This study follows the general framework as displayed in Fig. 3.

### 3.1. Document selection

Both the named-entity identification task and the relation extraction task have access to many datasets. For the former, there is the CoNLL [33], BioCreative [34], and Wiki Gold [35] datasets; while for the latter, the BioRED [36], CodRED [37], and DocRED [37] datasets. However, the research objective of building a micro-dataset arises from the need for a math-rich dataset in training and evaluating a variable typing task. Hence, a corpus related to pure or applied science is best suited for this task. The 8th Edition of the book Introduction to Chemical Engineering Thermodynamics by J.M. Smith, H. C. Van Ness, M. M. Abbott, and M. T. Swihart was used as a reference corpus to build the dataset used in the training and testing of the models. This textbook was utilized for two primary reasons. First, this reference is not a fundamental Mathematics textbook which gives us a broader overview of model performance on the mathematical discussion which is biased towards scientific undertones; and second, the level of difficulty or complexity of discourse is intermediate, which means that sentence spans are combinations various simple to complicated topics.

### 3.2. Document sectioning

Document Sectioning aims to identify the different regions or blocks of the document at hand, e.g., text blocks, images, graphs, and so on. This step, despite preliminary, is important because it determines the robustness of available entities present in the dataset. In this study, document sectioning was done manually. Unlike others, the scientific domain contains additional heterogeneity since on top of the usual contents like photos and text blocks, it may also contain data visualizations, complex figures, and equations.

From the textbook of interest, the text blocks were identified irrespective of their relative position in the document (Fig. 4). Certain assumptions and guidelines were followed to template the locations of entity and relations-rich sections to build the most robust dataset possible.

a. Entities and relations appear exclusively from sentence to sentence.

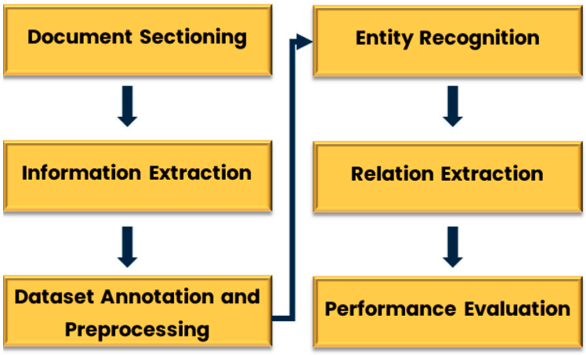
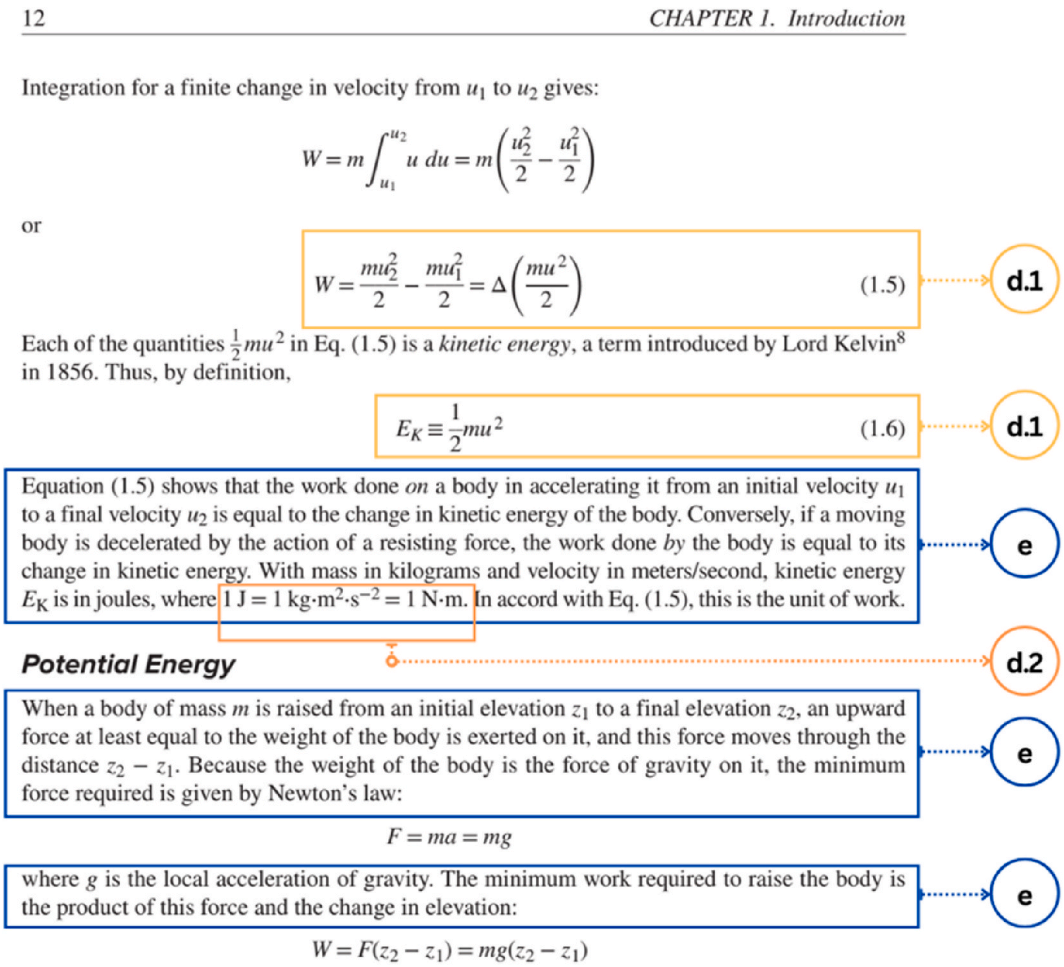


**Fig. 3.** General Framework of this study.



12 CHAPTER 1. Introduction

Integration for a finite change in velocity from $u_1$ to $u_2$ gives:

$$W = m\int_{u_1}^{u_2} u\,du = m\left(\frac{u_2^2}{2} - \frac{u_1^2}{2}\right)$$

or

$$W = \frac{mu_2^2}{2} - \frac{mu_1^2}{2} = \Delta\left(\frac{mu^2}{2}\right) \qquad (1.5)$$

Each of the quantities $\frac{1}{2}mu^2$ in Eq. (1.5) is a *kinetic energy*, a term introduced by Lord Kelvin[8] in 1856. Thus, by definition,

$$E_K \equiv \frac{1}{2}mu^2 \qquad (1.6)$$

Equation (1.5) shows that the work done *on* a body in accelerating it from an initial velocity $u_1$ to a final velocity $u_2$ is equal to the change in kinetic energy of the body. Conversely, if a moving body is decelerated by the action of a resisting force, the work done *by* the body is equal to its change in kinetic energy. With mass in kilograms and velocity in meters/second, kinetic energy $E_K$ is in joules, where $1\text{ J} = 1\text{ kg}\cdot\text{m}^2\cdot\text{s}^{-2} = 1\text{ N}\cdot\text{m}$. In accord with Eq. (1.5), this is the unit of work.

**Potential Energy**

When a body of mass $m$ is raised from an initial elevation $z_1$ to a final elevation $z_2$, an upward force at least equal to the weight of the body is exerted on it, and this force moves through the distance $z_2 - z_1$. Because the weight of the body is the force of gravity on it, the minimum force required is given by Newton's law:

$$F = ma = mg$$

where $g$ is the local acceleration of gravity. The minimum work required to raise the body is the product of this force and the change in elevation:

$$W = F(z_2 - z_1) = mg(z_2 - z_1)$$

**Fig. 4.** Shows a sample document image where some of the preceding guidelines were applied.

b. Relations are identified between entities irrespective of their order.
c. Priority sentences are those that are superfluous in entities Variables and Definition.
d. Equations are presented in two ways.
   1. As a displayed equation whereby characters involving an equal sign with texts on its left and right side, are surrounded by a relatively large area of white spaces. Then it usually comes with an equation number.
   2. As an in-line equation where the entity is in the middle of the sentence. In contrast, there are no wide white spaces surrounding the equation of this type, and they are not marked using an equation number. The former is treated akin to an independent sentence, while the latter is similar to other entities.
e. Sentences preceding or succeeding a displayed equation commonly present multiple entities, and consequently the relations among these entities.
f. Sentences found in between sections describing how an equation was derived may also contain entities and relations.
g. Explicit or structured presentation of the definition of terms is identified to contain the entities Variable and Definition.
h. Sentences that present constants contain entities belonging to classes Magnitude, Definition, or Units.

### 3.3. Information extraction

Once relevant sections were determined, information extraction was done manually by snipping out the identified regions using the MathPix © Optical Character Recognition (OCR) system [18]. This tool converts the selected region into a document image and then extracts the text into its digitized equivalent. The conversion is ensued by a confidence level, which allows us to only accept entries crossing the 90% extraction accuracy.

No two succeeding paragraphs were snipped simultaneously as this may cause conversion errors where the machine may erroneously identify a character. The image should have a good resolution to avoid this error. This is also the same reason why no entire page was snipped altogether. When two target paragraphs are consecutively positioned on the page, for example, two snipping instances are made. Additionally, since the OCR is akin to a snipping tool that requires distinct bounding regions, and the regions on the image are not necessarily complete sentences, truncated paragraphs and phrases were included in the sampling as long as their position fits the guidelines presented in this paper. Given this premise, the succeeding sections shall refer to them as **documents** as a more general term instead of using the term **sentence**.

### *3.4. Dataset annotation*

The snipped documents were pasted on a web-based open-source annotating tool called TagTog (tagtog.com) and were manually labeled. The type of annotation made in this fashion produces the Gold-Standard Dataset.

All words or phrases that were identified as entities fall under this set of tags: S = {“Variable”, “Definition”, “Units”, “Magnitude”, “Equation”}; while the annotation scheme for relations is predefined using these relation labels: {‘Definition_Variable’, ‘Units_Magnitude’, ‘Definition_Units’, ‘Variable_Units’, ‘Variable_Magnitude’, ‘Equation_Definition’}. These tags and labels were adopted from the study of Remolona (2018) [7] as they are key terms in organizing mathematical knowledge. Since the goal of the annotations is to train a model to conduct variable typing, a few natural language deviations were made. For example, a variable may also include items that are not necessarily a single letter or symbol. A word can be identified as a Variable when it is the one being defined in a sentence. Definitions can also include a single word, aside from phrases – as long as it “defines” or describes another entity.

Entities were first identified by highlighting the words or phrases in the documents using the appropriate color equivalent to the entity label. Relations, on the other hand, were identified by clicking “Add Relations” on the dropdown menu on top of one entity and then clicking another entity to automatically relate them with one another.

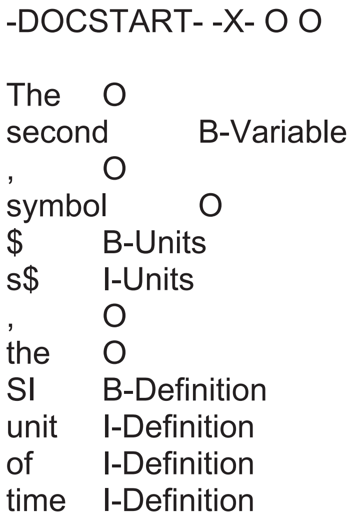
Note that the labels may not necessarily correspond to their semantic meaning in the English Language; rather, they must be contextually consistent with the instant NLP objective – Variable Typing – with respect to its surrounding tokens. For example, a token tagged as a Variable in one instance may be taken as a definition in another. Consequently, the relations will also change as it is dependent on the label of the entities. Two deductions can be inferred from this observation. One is that there is inherent bias embedded in the annotation process precisely because the labeling is dependent on the understanding and skills of the annotator. The second is that the quality and quantity of relations annotations heavily rely on entity annotations, and it also follows that the relation annotation biases are minimized since much of the impact of bias is encumbered on the entities.

### *3.5. Dataset preprocessing*

Two separate files were created in the preprocessing step. One file is for the Entity Recognition task and the other is for the Relation Extraction task. The former assumes a CoNLL dataset format as shown in Fig. 5. The latter, on the other hand, was preprocessed to form a JSON file comprised of a dictionary of dictionaries containing the document itself, a list of labeled entities, their indices, ordinal positions, and the relations as shown in Fig. 6. The specific steps for the preprocessing steps were not shown for brevity.

A few instances of conflicts may arise due to the unavoidable inconsistency in the annotation. In such a case, a guiding thought during the preprocessing step is to ensure model optimization. This means that the researchers prioritize the consistency of entity-label pairs or entity-relations triple across the entire corpus. On the other hand, the researchers try to minimize the number of annotations deleted from the dataset. For example, some strings annotated as a Definition, are nested with at least one Variable or Magnitude entity, in such a case, the researchers decide to delete the annotation Variable.

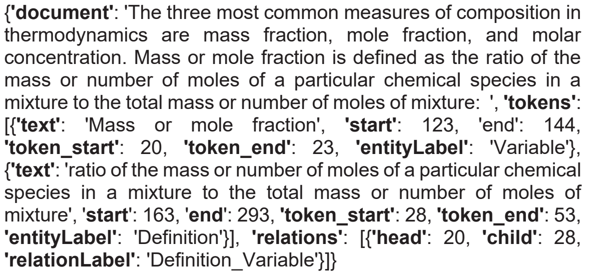
In this step, neither was the dataset extensively cleaned nor were “stop words” removed. Instead, the dataset was only preprocessed



```
-DOCSTART- -X- O O

The     O
second          B-Variable
,       O
symbol          O
$       B-Units
s$      I-Units
,       O
the     O
SI      B-Definition
unit    I-Definition
of      I-Definition
time    I-Definition
```

**Fig. 5.** Sample CoNLL dataset format.



{**'document'**: 'The three most common measures of composition in thermodynamics are mass fraction, mole fraction, and molar concentration. Mass or mole fraction is defined as the ratio of the mass or number of moles of a particular chemical species in a mixture to the total mass or number of moles of mixture: ', **'tokens'**: [{**'text'**: 'Mass or mole fraction', **'start'**: 123, 'end': 144, **'token_start'**: 20, **'token_end'**: 23, **'entityLabel'**: 'Variable'}, {**'text'**: 'ratio of the mass or number of moles of a particular chemical species in a mixture to the total mass or number of moles of mixture', **'start'**: 163, **'end'**: 293, **'token_start'**: 28, **'token_end'**: 53, **'entityLabel'**: 'Definition'}], **'relations'**: [{**'head'**: 20, **'child'**: 28, **'relationLabel'**: 'Definition_Variable'}]}

**Fig. 6.** Example of the Resulting Format of for the RE Dataset (emphasis on keys supplied).

into a CoNLL format.

{**'document'**: 'The three most common measures of composition in thermodynamics are mass fraction, mole fraction, and molar concentration. Mass or mole fraction is defined as the ratio of the mass or number of moles of a particular chemical species in a mixture to the total mass or number of moles of mixture: ', **'tokens'**: [{**'text'**: 'Mass or mole fraction', **'start'**: 123, 'end': 144, **'token_start'**: 20, **'token_end'**: 23, **'entityLabel'**: 'Variable'}, {**'text'**: 'ratio of the mass or number of moles of a particular chemical species in a mixture to the total mass or number of moles of mixture', '**start**': 163, '**end**': 293, **'token_start'**: 28, **'token_end'**: 53, **'entityLabel'**: 'Definition'}], **'relations'**: [{**'head'**: 20, **'child'**: 28, **'relationLabel'**: 'Definition_Variable'}]}

### 3.6. *The variable typing task*

Two NLP tasks were streamlined to build the end-to-end variable typing system, namely Entity Recognition and Relation Extraction. Both models were finetuned from the BERT-BASE-CASED Transformer model using the Spacy 3.0 pipelines. The same models were also implemented through the Tok2Vec Model in Spacy to compare the results of the BERT-based transformer. Spacy is a high-powered industry-preferred library due to its streamlined and simplistic approach to common NLP tasks while ensuring flexibility in model selection and parameter tuning. The models were fine-tuned separately using the separately preprocessed dataset for ER and RE.

As a starting point, the ER dataset was split into a training and test set with a 70%–30% division whereas the RE dataset was separated into a 60% training, 20% test, and 20% validation set. The validation dataset was set up for RE but not for ER because the latter is a built-in pipeline in Spacy, while the former is a custom pipeline and therefore requires more iteration in tuning the hyperparameters used during the fine-tuning process. On the other hand, a built-in pipeline in Spacy is already governed by default finetuning parameters included in the package. Some modifications were applied to the dataset divide of the latter for optimization. This will be discussed in the succeeding section of this paper. The datasets were converted to Spacy objects and were consequently fed to finetune the model. Spacy has a built-in pipeline in training a NER model; thus, this step was a straightforward application of the same (More here: https://spacy.io/usage). The RE pipeline is also available as a customized pipeline in Spacy repositories (https://github.com/p123hx/rel_component). This, together with its dependencies, must be installed through the command line before training the RE task.

Performance evaluation for this work, as in many similar NLP tasks in literature [7,11,12], was derived from Precision (Equation (1)), Recall (Equation (2)), and F1 score (Equation (3)). A recall is the fraction of target (or relevant) data retrieved, while precision is the fraction of retrieved data that are intended to be extracted. F1-score is the harmonic mean between precision and recall.

$$\mathrm{P} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \qquad \text{[Equation 1]}$$

$$\mathrm{R} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \qquad \text{[Equation 2]}$$

$$\mathrm{F1 - Score} = \frac{2\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \qquad \text{[Equation 3]}$$

To benchmark the results on a non-BERT model, the same gold-standard dataset was used for ER and RE implementations through the Tok2Vec model, which is the default model used by Spacy in ER and RE implementations. These implementations were separately evaluated using Precision, Recall, and F1-scores calculations.

## 4. Results and discussion

### 4.1. *The gold standard dataset*

A significant portion of the dataset is found in the earlier chapters of the corpus precisely because much of the foundational knowledge is introduced in that part of the textbook. Introductory paragraphs are easier to annotate than those found in the middle of the discussion because entities are usually presented more explicitly and straightforwardly in the former. Similarly, the distribution of the number of entities and relations follows the same trend; because for entities to be identified, entity-rich documents must be identified first. In the same breath, the identification of the relations requires that the entities are already known. This observation is important in case future researchers plan on expanding data sources. By then, researchers may opt to annotate introductory chapters from multiple books.

After pre-processing, the full gold-standard dataset is composed of 334 documents. Within the dataset, 2047 entity-label pairs and 863 relation triples were annotated. Table 1 and Table 2 show the statistics of entities across the dataset taken from the count of TagTog Annotating Tool. On the other hand, due to sparseness and limited instances of triples belonging to relation categories Definition_Units and Variable_Units, they were ultimately dropped and were not included in the fine-tuning process. Eventually, only 793 relations were retained. The tokenized and processed dataset assumes the form found in Fig. 5 for the ER dataset, and Fig. 6 for the RE dataset.

### 4.2. *Optimized model performance*

#### 4.2.1. *Entity recognition component*

Table 3 shows the overall ER performance of the BERT and Tok2Vec Models. Meanwhile, Tables 4 and 5 display the models' performance per entity label.

The overall model performance (Table 3) reveals strongly satisfactory evaluation results of the NLP task. The overall F1 score of 0.7975 is at par with Entity Recognition results done to BERT Models which are not tailor-fitted to a specific domain. For instance, the BERT Model applied to entity recognition in German texts was only able to show an F1 score of 0.7751 [38], while the same model applied to the pharmaceutical corpus resulted in a 0.5575 F1 score [20].

The model was able to show impressive performance despite having to overcome challenges that root back to the nature of the documents themselves. First, the BERT-BASE-CASED model, which was pretrained using a large corpus across several domains and presented in a standard English syntax, is the model that was employed in the fine-tuning process. In this case, however, the LaTeX strings contain extraneous punctuations and symbols that affect the word-shape features of the entities. This is called the Textual Genre or Domain Factor [39]. It must be noted that the fine-tuning process will revert a much higher evaluation metrics when the finetuning dataset closely follows the syntax and format of the pre-training corpus. This is the same challenge faced by various domain-specific NER adaptations in such as biomedical, scientific journals, or geographical datasets [40–42].

An impressive Precision, Recall, and F1 score obtained by entity category Magnitude infers that consistent word-shape characteristic is a performance booster. As recalled, almost all entities belonging to the category Magnitude were numerical figures and were presented with the '$' symbol to their right and left. This means that once the model identifies a possible entity, it will measure its distance from the nearest dollar sign to test whether it is a Magnitude or another entity. The same inference can be drawn from the results elucidated by entities belonging to the category Variable and Equations. Similarly, Variables and Equations were accompanied by dollar signs '$' at its left and right; at the same time, the latter must contain an equal sign ' = ' in between. The reason why there is a slight relative drop in its metrics is because of its ensuing length. Comparing Magnitude to Variables and Equations, the former is simpler as it is only usually broken down into a shorter list of tokens.

The relative performance of Definition as compared to the rest of the entity class is very telling of the effect of entity length on metrics. The latter classes are usually short or contiguous strings, while the former is usually a phrase in the document. Contrasting them all, Definition showed the biggest difference between Precision and Recall signifying that the model does not readily identify a Definition in a collection of possible entities. A fair result of 0.6045 Recall suggests that longer sequences of texts cause the model to miss identifying the correctly tagged entities. The feature of a definition is also complex as there is no explicit boundary (e.g., dollar sign) that separates it from the rest of the tokens, nor is there a consistent grammatical pattern in the presentation of each entity in the corpus. A definition may explicitly describe another entity or a group of entities enumerated in the document. A remedy to this performance is to increase the share of Definition entities in the gold-standard dataset or to essentially frame a more targeted annotation scheme so that span lengths are reduced to a minimal number of tokens.

**Table 1**
Number of positive instances of each entity across the corpus.

| Chapter | Units | Definition | Magnitude | Variable | Equation | Total |
|---|---|---|---|---|---|---|
| 1 | 51 | 68 | 19 | 74 | 11 | 223 |
| 2 | 11 | 76 | 2 | 102 | 8 | 199 |
| 3 | 7 | 105 | 16 | 306 | 67 | 501 |
| 4 | 19 | 51 | 20 | 130 | 10 | 230 |
| 5 | 11 | 72 | 16 | 116 | 18 | 233 |
| 6 | 4 | 51 | 2 | 122 | 8 | 187 |
| 7 | 7 | 37 | 13 | 85 | 20 | 162 |
| 8 | 2 | 15 | 2 | 21 | 1 | 41 |
| 9 | 19 | 30 | 24 | 60 | 10 | 143 |
| 10 | 2 | 51 | 1 | 71 | 3 | 128 |
| TOTAL | 133 | 556 | 115 | 1087 | 156 | 2047 |
| % Composition | 6.50% | 27.16% | 5.62% | 53.10% | 7.62% | |

**Table 2**
Number of positive instances of each Relation across the corpus.

| Chapter | 'Definition_Variable' | 'Units_Magnitude' | 'Variable_Magnitude' | 'Equation_Definition' | Total |
|---|---|---|---|---|---|
| 1 | 48 | 17 | 32 | 2 | 99 |
| 2 | 87 | 2 | 16 | 2 | 107 |
| 3 | 127 | 5 | 1 | 12 | 145 |
| 4 | 67 | 15 | 4 | 6 | 92 |
| 5 | 77 | 10 | 1 | 7 | 95 |
| 6 | 76 | 2 | 6 | 0 | 84 |
| 7 | 27 | 7 | 0 | 8 | 42 |
| 8 | 15 | 2 | 0 | 0 | 17 |
| 9 | 26 | 20 | 0 | 11 | 57 |
| 10 | 52 | 0 | 1 | 2 | 55 |
| TOTAL | 602 | 80 | 61 | 50 | 793 |
| % Composition | 75.91% | 10.09% | 7.69% | 6.31% | |

**Table 3**
Overall model performance for Entity Recognition Task.

| Model | Precision | Recall | F1 Score |
|---|---|---|---|
| **Tok2Vec** | 0.7500 | 0.6192 | 0.6784 |
| **BERT** | 0.8142 | 0.7816 | 0.7975 |

**Table 4**
BERT Model performance for Entity Recognition task per class.

| | Precision | Recall | F1 Score |
|---|---|---|---|
| **Variable** | 0.8929 | 0.8789 | 0.8858 |
| **Definition** | 0.6231 | 0.6045 | 0.6163 |
| **Units** | 0.9583 | 0.6970 | 0.8070 |
| **Magnitude** | 0.9024 | 0.8810 | 0.8916 |
| **Equation** | 0.7500 | 0.7059 | 0.7273 |

**Table 5**
Label-wise evaluation results for Tok2Vec-based Entity Recognition.

| | Precision | Recall | F1 Score |
|---|---|---|---|
| **Variable** | 0.8776 | 0.8398 | 0.8583 |
| **Definition** | 0.5122 | 0.3134 | 0.3889 |
| **Units** | 0.3636 | 0.3529 | 0.3582 |
| **Magnitude** | 0.8462 | 0.5238 | 0.6471 |
| **Equation** | 0.6923 | 0.5455 | 0.6102 |

#### *4.2.2. Relation extraction component*

The first attempt for the RE task is to feed the preprocessed dataset directly to the model without further manipulation. This is a litmus test that gave insights into how the model, pretrained on standard English sentence structure, will perform given a novel dataset of differing syntax. A promising F1 score of 0.5043 was noted (Table 6), which is within the range of results obtained from similar tasks in the literature. The error in the evaluation matrix stems from the number of relation labels. In the study of Remolona [7], he observed that the complexity of an NLP classification problem increases linearly with the number of categories involved while the model accuracy on the other hand decreases. Therefore, it is logical to reduce the number of categories by identifying the least and most sparse classes in the dataset. That is primarily why the researchers have identified them as Magnitude_Variable and Variable_Units categories.

There is an increase of almost 3% in the F1 score which shows the positive impact of reducing the number of categories. Given the limited number of instances, this effect will be more significant in a more robust dataset. In the same lens, it elucidates that another cause of error is the complexity of the relations between entities – as more relations are made, the more error is made in prediction. Hence, it is important to keep annotations as fundamental as possible. Additionally, there was a noticeable decrease in the variance between the number of positive instances versus the total number of all instances in this approach. This shows that evaluation metrics can be additionally improved by decreasing this difference between the number of positive instances. This is the premise of the third attempt.

To increase the evaluation metrics, the difference between the total instances versus positive instances must be reduced. This difference can be minimized by manipulating the data so that each entry should only contain one relation, forming only four instances. However, this will increase errors since the model is not trained to detect multiple relations in a document entry. Therefore, to

**Table 6**
Overall evaluation results for Relation Extraction Using the BERT Model.

| Approach | No. of Positive Instances | Total No. of Instances | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| **Chapters 1–10** (All) | 807 | 62, 424 | 0.4643 | 0.5519 | 0.5043 |
| **Chapters 1–10** (Reduced) | 759 | 43, 984 | 0.4798 | 0.5980 | 0.5324 |
| **Chapters 1–10*** (Reduced and capped at 4 relations per document) | 759 | 25, 632 | 0.4919 | 0.6030 | 0.5418 |

maintain the model's ability to determine the context of multiple relations in a document, the same must be used in the training set. A cap of four relations per document was employed, and if a document has an excessive number of relations, the same will be duplicated as another entry and the excess relations will be transferred as part of the new entry. This step has significantly reduced the variance by 41.06% of the original number of total instances. Similarly, it has driven a sustained increase in the F1 score and has consistently impacted model precision and recall as displayed in Table 6.

The last approach, which produced the best evaluation results for the BERT model was replicated using the Tok2Vec Model (Table 7). For both entity recognition and relation extraction tasks, the BERT-base implementation performed better than Tok2Vec. In the ER task, BERT performs better by 13.67% in the F1-score as compared to Tok2Vec implementation. While it performed by 26.03% in the RE task. It also consistently performed better across the three metrics showing that the BERT model predicts more precisely and accurately for a new micro-dataset than its Tok2Vec counterpart. This is attributable to BERT's pretrained character, and the transfer learning approach for its fine-tuning. Hence, this is best used for the micro-dataset. In literature, the work of Zhang et al. [27] on BERT-based relation extraction using a Biomedical corpus, is comparable where they obtained a Precision of 0.6990, Recall of 0.6333, and F1-Score. The slight difference in the results is attributable to the more robust dataset used in their study.

## 5. Conclusion

Variable Typing is seemingly simplistic at face value, but it must not be trivialized. It cannot be treated as a mere automation task because it requires a model that can understand the context; hence it necessitates rigorous scientific steps or an intermediate understanding of the target domain to frame up rules or train a model. This has been achieved in this work.

A total of 334 spans from the first 10 chapters of the textbook Introduction to Chemical Engineering Thermodynamics, were extracted and annotated to build the gold-standard dataset. The same was then used as a novel attempt to finetune the BERT Model into two downstream tasks namely ER and RE. A composite of these two tasks comprises the end-to-end implementation of the Variable Typing system that can recognize other nodes, such as Equation, Magnitude, and Units, aside from Variable and Definition. What's more, is that the RE task allows the prediction of the name of the edges between nodes and not just affirms its existence. This novel treatment to variable typing has given us new insights on how BERT impacts the accuracy of automated data transformation and structurization, and accordingly, what steps can be taken so that the results of future work will be improved significantly. Our results consistently show that the BERT-based approach produced better results for both tasks and across all evaluation criteria.

It is recommended that future researchers explore the suitability of other annotation tools that output a raw file that is easier to process. An annotation tool like UBIAI (https://ubiai.tools/) is recommended as it was truly developed for Spacy dataset preparation. It does not only have a better user interface, but it also produces a raw file that is already closer to the desired format for Spacy training. It is also imperative for researchers to attempt using a different model such as Roberta-base, Bio-BERT, or ALBERT on the micro-dataset which performs an end-to-end ER and RE task. Once a very superfluous dataset is available, pretraining a BERT variation on a mathematical dataset following a LaTeX syntax can also help improve the Domain Factor. Lastly, A postprocessing step may be done after the finetuning process. Future researchers may apply coreference relation to avoid repetition of the same word in tagging outputs, or they may apply a rules-based postprocessing overlay to work around with the vicissitudes of threshold requirement.

## Funding

This research was funded by the Department of Science and Technology Engineering Research and Development for Technology (DOST-ERDT) Program of the Republic of the Philippines and the PhD Incentive Awards of the University of the Philippines Diliman.

## Author contribution statement

Charles Arthel Rey, Miguel Francisco Remolona: Conceived and designed the experiments; performed the experiments, analyzed and interpreted the data; contributed reagents, materials, analysis tools or data; wrote the paper. Jose Lorenzo Danguilan, Karl Patrick Mendoza: Performed the experiments; analyzed and interpreted the data; contributed reagents, materials, analysis tools or data.

**Table 7**
Full dataset training metrics for RE using Tok2Vec Model.

| | No. of Positive Instances | Total No. of Instances | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| **Chapters 1–10*** (reduced and capped at 4 relations per document) | 759 | 25, 632 | 0.4144 | 0.4466 | 0.4299 |

## Data availability statement

Data associated with this study has been deposited at https://github.com/chrlsrey/variable-typing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

C.A. Rey is grateful to the Philippine Department of Science and Technology - Engineering Research and Development for Technology (DOST-ERDT) Program for its support. M.F. Remolona is thankful to the PhD. Incentive Awards of the University of the Philippines Diliman (UPD). J. L. Danguilan and K. P. Mendoza acknowledge the UPD Department of Chemical Engineering for their guidance and help.

## References

[1] S. Boon, “21st Century Science Overload,”, Canadian Science Publishing, 2017. http://blog.cdnsciencepub.com/21st-century-science-overload/. (Accessed 7 April 2022).

[2] Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, Addressing big data issues in scientific data infrastructure, in: 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013, pp. 48–55, https://doi.org/10.1109/CTS.2013.6567203. May.

[3] S. Leonelli, Scientific Research and Big Data, May 2020. Accessed: Apr. 07, 2022. [Online]. Available: https://plato.stanford.edu/entries/science-big-data/?ref=hackernoon.com.

[4] T. Reller, “Elsevier Publishing – a Look at the Numbers, and More,”, Elsevier Connect, 2016. https://www.elsevier.com/connect/elsevier-publishing-a-look-at-the-numbers-and-more (accessed Feb. 08, 2021).

[5] P.G. Altbach, H. de Wit, Too much academic research is being published, Internet High Educ. (2018), https://doi.org/10.6017/ihe.2019.96.10767.

[6] D. Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety,”, META Group, 2001.

[7] M.F. Remolona, Holmes, A Hybrid Ontology-Learning Materials Engineering System, Columbia University Academic Commons, 2018 [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D8WH46P7.

[8] M.A. Andrade, Bork Peer, Automated Extraction of Information in Molecular Biology, Elsevier Sci., 2000, https://doi.org/10.1016/S0014-5793(00)01661-6.

[9] T. Charnois, Nicolas Durand, and Jiri Klima, “Automated Information Extraction from Gene Summaries,”, presented at the Workshop on Data and Text Mining for Integrative Biology, Berlin, Germany, 2006, pp. 4–15.

[10] Y. Stathopoulos, S. Baker, M. Rei, S. Teufel, Variable typing: assigning meaning to variables in mathematical text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies vol. 1, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 303–312, https://doi.org/10.18653/v1/N18-1028.

[11] M. Nghiem Quoc, K. Yokoi, Y. Matsubayashi, A. Aizawa, Mining coreference relations between formulas and text using Wikipedia, in: Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010), Coling 2010 Organizing Committee, Beijing, China, Aug. 2010, pp. 69–74. Accessed: Apr. 14, 2022. [Online]. Available: https://aclanthology.org/W10-3910.

[12] G.Y. Kristianto, G. Topić, A. Aizawa, Extracting textual descriptions of mathematical expressions in scientific papers, -Lib Mag. 20 (11/12) (Nov. 2014), https://doi.org/10.1045/november14-kristianto.

[13] D. Ferreira, M. Thayaparan, M. Valentino, J. Rozanova, A. Freitas, To be or not to be an integer? Encoding variables for mathematical text, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, May 2022, pp. 938–948, https://doi.org/10.18653/v1/2022.findings-acl.76.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv181004805 Cs, May 2019. Accessed: Apr. 19, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805.

[15] P. Suresh, et al., Onto MODEL: ontological mathematical modeling knowledge management, in: B. Braunschweig, X. Joulia (Eds.), Computer Aided Chemical Engineering, In 18 European Symposium on Computer Aided Process Engineering, vol. 25, Elsevier, 2008, pp. 985–990, https://doi.org/10.1016/S1570-7946(08)80170-8.

[16] P. Sojka, M. Růžička, V. Novotný, MIaS: math-aware retrieval in digital mathematical libraries, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, Oct. 2018, pp. 1923–1926, https://doi.org/10.1145/3269206.3269233. CIKM ’18.

[17] M. Grigore, M. Wolska, M. Kohlhase, Towards context-based disambiguation of mathematical expressions, in: Asian Symposium on Computer Mathematics and Mathematical Aspects of, Computer and Information Sciences, 2009.

[18] N. Jimenez, K. Cunningham, MathPix, 2016. https://mathpix.com/blog/snip-notes-beta.

[19] A. Vaswani, et al., Attention Is All You Need,” *ArXiv170603762 Cs*, Dec. 2017. Accessed: Apr. 19, 2022. [Online]. Available: http://arxiv.org/abs/1706.03762.

[20] K. Hakala, S. Pyysalo, Biomedical named entity recognition with multilingual BERT, in: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 56–61, https://doi.org/10.18653/v1/D19-5709. Nov.

[21] F. Souza, R. Nogueira, R. Lotufo, Portuguese Named Entity Recognition Using BERT-CRF, ArXiv190910649 Cs, Feb. 2020. Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1909.10649.

[22] C. Che, et al., Fast and effective biomedical named entity recognition using temporal convolutional network with conditional random field, Math. Biosci. Eng. 17 (4) (2020) 3553–3566, https://doi.org/10.3934/mbe.2020200.

[23] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, ArXiv150801991 Cs, Aug. 2015. Accessed: May 03, 2022. [Online]. Available: http://arxiv.org/abs/1508.01991.

[24] M.V. Koroteev, BERT: A Review of Applications in Natural Language Processing and Understanding, Mar, ArXiv210311943 Cs, 2021. Accessed: Sep. 29, 2021. [Online]. Available: http://arxiv.org/abs/2103.11943.

[25] Y. Zhang, S. Wang, B. He, P. Ye, K. Li, A BERT-based named entity recognition method for elementary mathematical text, 计算机应用 42 (2) (2022) 433–439, https://doi.org/10.11772/j.issn.1001-9081.2021020334.

[26] P. Shi, J. Lin, Simple BERT Models for Relation Extraction and Semantic Role Labeling, ArXiv190405255 Cs, Apr. 2019. Accessed: May 04, 2022. [Online]. Available: http://arxiv.org/abs/1904.05255.

[27] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, P. He, Fine-tuning BERT for joint entity and relation extraction in Chinese medical text, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov., 2019, pp. 892–897, https://doi.org/10.1109/BIBM47256.2019.8983370.

[28] X. Han, L. Wang, A novel document-level relation extraction method based on BERT and entity information, IEEE Access 8 (2020) 96912–96919, https://doi.org/10.1109/ACCESS.2020.2996642.

[29] C. Lin, T. Miller, D. Dligach, S. Bethard, G. Savova, A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, Jun. 2019, pp. 65–71, https://doi.org/10.18653/v1/W19-1908.

[30] G. Zhou, J. Su, J. Zhang, M. Zhang, Exploring various knowledge in relation extraction, in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, Jun. 2005, pp. 427–434, https://doi.org/10.3115/1219840.1219893.

[31] D. Zelenko, C. Aone, Richardella, Kernel methods for relation extraction, J. Mach. Learn. Res. (2003) 1083–1106.

[32] N. Kambhatla, Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, Jul. 2004, pp. 178–181. Accessed: May 04, 2022. [Online]. Available: https://aclanthology.org/P04-3022.

[33] E.F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. Accessed: Jul. 23, 2023. [Online]. Available: https://aclanthology.org/W03-0419.

[34] C.-H. Wei, et al., Assessing the State of the Art in Biomedical Relation Extraction: Overview of the BioCreative V Chemical-Disease Relation (CDR) Task, Patent, 2016, https://doi.org/10.1093/database/baw032 baw032, Jan. 2016.

[35] J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Learning multilingual named entity recognition from Wikipedia, Artif. Intell. 194 (Jan. 2013) 151–175, https://doi.org/10.1016/j.artint.2012.03.006.

[36] L. Luo, P.-T. Lai, C.-H. Wei, C.N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction dataset, Briefings Bioinf. 23 (5) (2022) bbac282, https://doi.org/10.1093/bib/bbac282. Sep.

[37] Papers with Code - DocRED Dataset." https://paperswithcode.com/dataset/docred (accessed July. 23, 2023).

[38] S. Schweter, J. Baiter, Towards robust named entity recognition for historic German, in: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Association for Computational Linguistics, Florence, Italy, Aug. 2019, pp. 96–103, https://doi.org/10.18653/v1/W19-4312.

[39] G. Krasteva, V. Georgiev, A. Pavlov, Recent applications of plant cell culture technology in cosmetics and foods, Eng. Life Sci. 21 (3–4) (2021) 68–76, https://doi.org/10.1002/elsc.202000078.

[40] L. Li, R. Zhou, D. Huang, Two-phase biomedical named entity recognition using CRFs, Comput. Biol. Chem. 33 (4) (2009) 334–338, https://doi.org/10.1016/j.compbiolchem.2009.07.004. Aug.

[41] M.J. Silva, B. Martins, M. Chaves, A.P. Afonso, N. Cardoso, Adding geographic scopes to web resources, Comput. Environ. Urban Syst. 30 (4) (2006) 378–399, https://doi.org/10.1016/j.compenvurbsys.2005.08.003. Jul.

[42] E. Yan, Y. Zhu, Identifying entities from scientific publications: a comparison of vocabulary- and model-based methods, J. Informetr. 9 (3) (2015) 455–465, https://doi.org/10.1016/j.joi.2015.04.003. Jul.