Extracting context of math formulae contained inside scientific documents

Amarnath Pathak[1], Ranjita Das[1], Partha Pakray[2], and Alexander Gelbukh[3]

[1] Department of Computer Science & Engineering
National Institute of Technology Mizoram
Aizawl, India
[2] Department of Computer Science & Engineering
National Institute of Technology Silchar
Silchar, India
[3] Instituto Politécnico Nacional, Mexico City

Abstract. A math formula present inside a scientific document is often preceded by its textual description, which is commonly referred to as the context of formula. Annotating context to the formula enriches its semantics, and consequently impacts the retrieval of mathematical contents from scientific documents. Also, with a considerable surety, a context can be assumed to be one of the Noun Phrases (NPs) of the sentence in which formula occurs. However, the presence of several different misleading NPs in the sentence necessitates extraction of an NP, which is more precise to the formula than the rest. Although a fair number of methods are developed for precise context extraction, it can be fascinating to prospect other competent techniques which can further their performances. To this end, this paper discusses implementation of an automated context extraction system, which follows certain heuristics in assigning weights to different candidate NPs, and tune those weights using a development set comprising annotated formulae. The implemented system significantly outperforms nearest noun and sentence–pattern based methods on the ground of F–score.

Keywords: Context Extraction · Math Information Retrieval · NTCIR · Parser · Noun Phrase.

1 Introduction

Increased research in Science, Technology, Engineering and Mathematics (STEM) disciplines has boosted the count of scientific documents, which are majorly constituted of math formulae. As a consequence, a number of Math Information Retrieval (MIR) systems, which can retrieve mathematical contents alongside plain text, have come into being. Math tasks [1, 2, 21] of NII Testbeds and Community for Information access Research (NTCIR) conferences [7, 6, 8] have also triggered widespread development of competent MIR systems.

Current MIR systems are either adapted versions of conventional text-search engines [11] or the systems developed from scratch [14, 15, 17]. While the text-search engines adapted for MIR perform linearization and plain text matching to retrieve formulae, the MIR systems developed from scratch employ novel formula indexing and search techniques. Although the challenges in designing such math-aware systems are enormous, the inability to account for ambiguity of formula, whereby the same formula may have different alternative interpretations, can cause severe performance degradation. Consider, for example, the ambiguous formula ($c = \sqrt{a^2 + b^2}$), which may exhibit following different meanings in two different documents:

(a) Using Pythagorean theorem to compute hypotenuse (c) of a right-angled triangle, whose base is 'b' and perpendicular is 'a'.

(b) Computing linear eccentricity (c) of hyperbola, with 'a' being distance from the center to the vertex and 'b' being the half distance between the asymptotes.

Given the two above-mentioned documents and a user query ($c = \sqrt{a^2 + b^2}$) intending to retrieve search results for *"linear eccentricity"*, a math-aware search engine [11], which only considers formula matching and discards the underlying semantics of the formulae, will also retrieve the irrelevant document containing *"Pythagorean theorem"*. This diminishes precision score, hence reduces retrieval performance of MIR systems. Therefore, it becomes essential to extract most appropriate context from the surrounding text and perform semantification of formula by associating it with the extracted context. Also, the semantification of formula eliminates the need for querying a formula using only formula. Instead, the end-users relish flexibility to specify a text query (say, *"Kinetic Energy"*) for searching a formula (say, $\frac{1}{2}mv^2$) inside document. Moreover, semantification facilitates improvement in comprehensibility of formula.

The work described in this paper is based on a reasonable assumption that the context of formula is one of the Noun Phrases (NPs) of the sentence containing formula (henceforth called target sentence). Therefore, the context extraction task reduces to parsing the target sentence, extracting all the candidate NPs, and devising an algorithm to select the most appropriate NP from among the diverse pool of candidate NPs. However, as the most appropriate NP does not adhere to a strict pattern, the task of context extraction turns out to be challenging. The following three example situations elaborate on this particular insight:

Example 1.1:
Often, momentum transfer is given in wavenumber units in reciprocal length $Q = k_f - k_i$.

In example 1.1 above, the three candidate NPs for the context of formula ($Q = k_f - k_i$) are: *"momentum transfer"*, *"wavenumber units"* and *"reciprocal length"*. Also, the most appropriate context is the NP (*"momentum transfer"*) which occurs farthest from the formula.

Example 1.2:
This is the simplest example of scattering of two colliding particles with initial momenta $\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}$.

In example 1.2 above, the candidate NPs for the context of formula $(\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2})$ are: *"the simplest example of scattering of two colliding particles with initial momenta"*, *"the simplest example"*, *"scattering of two colliding particles with initial momenta"*, *"scattering"*, *"two colliding particles with initial momenta"*, *"two colliding particles"* and *"initial momenta"*. Out of all the seven candidate NPs, the most appropriate one (i.e. *"initial momenta"*) occurs closest to the formula.

Example 1.3:
Using this free-body diagram the torque required to lift or lower a load can be calculated: $T_{raise} = \frac{Fd_m}{2}\left(\frac{l+\pi\mu d_m}{\pi d_m - \mu l}\right) = \frac{Fd_m}{2}\tan(\phi+\lambda)$.

In example 1.3 above, the candidate NPs for the context of formula are: *"this free-body diagram"*, *"the torque required to lift or lower a load"*, *"the torque"* and *"a load"*. Also, the most appropriate context (i.e. *"the torque required to lift or lower a load"*) is the longest of all NPs and appears somewhere in the middle of the parse tree generated for target line.

To summarize, the uncertainty in position of occurrence of the context in the target sentence poses challenge to the design of context extraction system. Nevertheless, owing to the advantages of formula semantification, recent years have witnessed a surge in the research activities concerned with context extraction. Some such activities include use of nearest noun method [13, 10], sentence–pattern based method [10, 20] and machine learning approach [10, 20]. After having identified all the NPs in target sentence, the nearest noun method considers nearest NP to be the context of formula. The sentence–pattern based method works under the assumption that the formulae are often linked to their contexts through specific words or group of words, such as *"denotes"*, *"describes"*, *"means"*, *"is given by"* and so on. For instance, in example 1.3 above, the formula and context are linked through the pattern, namely *"can be calculated"*. However, the method incurs failure in retrieving context–formula pairs which do not adhere to such fixed patterns. The machine learning based method [10] pairs math formula with "all NP" and "minimal NP" in the target sentence and extracts features for each pair. Thereafter, each pair is fed to a binary classifier to decide if the NP is most appropriate context for the formula.

The main contribution of this paper lies in devising a context extraction system, which extracts target sentences from scientific documents, parses all such sentences using Stanford Shift-Reduce Constituency Parser[4][22, 4], extracts all

[4] https://nlp.stanford.edu/software/srparser.html

NPs from the parse trees of target sentences, assigns weight to different candidate NPs of a formula using certain heuristics, tunes the weights using a development set containing formulae and their respective gold contexts, and eventually extracts most appropriate contexts from target sentences of test formulae. The implemented system performs reasonably well in comparison to other competent systems.

Rest of the paper is structured as follows: Section 2 reviews past works related to extraction of context of formula and identifier definition. Section 3 comprehensively describes working of different constituents of the implemented system. Section 4 describes experimental setup used to develop and evaluate the system. Section 5 presents experimental results and in-depth analysis of results to comprehend strengths and weaknesses of the implemented system. Section 6 concludes the paper and points directions for future research.

2 Related Works

In past, the works related to extraction of contexts of formulae and definitions of their constituent identifiers have been prominent. As the two categories of works closely resemble, the following subsections elaborate on past developments related to both the categories.

2.1 Extraction of formula context

It is a usual practice to compare performance of any context extraction system with the performances of nearest noun and sentence–pattern based methods, which were introduced in the previous section. Sentence–pattern based method often uses the seven distinct patterns, described in [10, 19] and shown in Table 1, for discovery of context.

Table 1. Patterns used in sentence–pattern based method. MATH: math formula; DEF: definition of math formula i.e. context; OTHERMATH: other math formula

Sl.No.	**Patterns**
1	... denoted (as | by) MATH DEF
2	(let | set) MATH (denote | denotes | be) DEF
3	DEF (is | are)? (denoted | defined | given) (as | by) MATH
4	MATH (denotes | denote | (stand | stands) for | mean | means) DEF
5	MATH (is | are) DEF
6	DEF (is | are) MATH
7	DEF (OTHERMATH)* MATH

The work described in [9] views context extraction as a binary classification problem, wherein the description candidates associated with formulae are classified as correct or incorrect. Performances of nearest noun, sentence–pattern and machine learning methods for context extraction are compared and analyzed. The model using “All NP” approach and all possible feature augmented to machine learning approach depicts better performance than the model using minimal NP approach.

Work described in [20] focuses on connecting mathematical mentions, namely names, definitions and explanations, with corresponding mathematical expressions contained inside Japanese scientific papers. A Support Vector Machine (SVM) trained using features, such as basic patterns and linguistic information, helps select correct description for an expression and outperforms conventional pattern based method.

The guideline to annotate mathematical expressions with their respective definitions is described in [10]. The annotated data is used to examine performance of proposed context–extraction method. The proposed machine learning method extracts a set of 10 features (such as distance of candidate NP from the formula, Parts Of Speech (POS) tags of the text surrounding candidate NP, and so on) for candidate NPs and compares them with gold context. Under the constraint of strict matching, the machine learning method significantly outperforms nearest noun and pattern matching based methods.

The MARACHNA system [12] exploits Natural Language Processing (NLP) methods for extracting information from mathematical texts. More specifically, the MARACHNA generates ontologies for mathematical information extracted from different sources, and later stores them in a Knowledge Base (KB). The KB also stores different keywords and texts associated with a formula.

Concept Description Formula (CDF) approach [16] prospects coreference relation, if any, between the formula and context. The claim is made that extracting keywords using CDF and associating them with formulae will ease the task of MIR. Text preprocessing, text matching, pattern generation and pattern matching constitute key steps of CDF approach. Experimented using Wikipedia articles, the system depicts competence in finding coreference relation between text and formula.

An approach [5] to disambiguate mathematical expressions computes similarity between the words extracted from surrounding text of formula and a collection of term clusters derived from Content Dictionaries of OpenMath [3]. Subsequently, the cluster which shares highest similarity with the words in surrounding text is considered to be the most accurate textual interpretation of the formula.

2.2 Extraction of identifier definition

Similar to formula, the meaning of an identifier may differ across documents or even across the formulae inside same document. For instance, the symbol ‘E’ in a formula may designate electric field or Young’s modulus. As the extractions of formula context and identifier definition share same underlying concerns, this subsection reviews some of the past works related to identifier definition extraction.

Mathematical Language Processing (MLP) project [13] computes probabilities of identifier–definition pairs using POS based distance and sentence positions. Identifier definitions are discovered using pattern–based and statistical approaches. While the pattern–based approach uses 6 static patterns, the statistical approach extracts candidate definitions and ranks them using a weighted sum. Concretely, the statistical approach of MLP is a five–step process: (i) Detecting formulae from the documents (ii) Extracting identifiers from the formula (iii) Finding identifiers in surrounding text (iv) Finding candidate phrases/tokens for identifiers, and (v) Ranking candidate phrases/tokens using weighted sum. The recall measure for statistical approach is found to be greater than that of pattern–based approach. Moreover, the statistical approach is least affected by the change in sentence structure.

Semantification of identifiers in formula is further improved through discovery of namespaces [18]. Although the concept of namespaces primarily applies to software development, the idea is extended to mathematical identifiers. Using NLP techniques, namespaces are discovered from the surrounding text of a formula. In summary, the identifier–definition extraction suing namespace approach is a four–step process: (i) Automatic discovery of namespaces (ii) Clustering of documents (iii) Building namespaces, and (iv) Building namespace hierarchy.

3 System Description

Key constituents and working principle of the system are explained in subsequent subsections.

3.1 Corpus Description

The system is experimented using open source Wikipedia corpus[5] of NTCIR-12 MathIR task [21]. Unlike arXiv corpus, the Wikipedia corpus is intended for non-technical users. Each document in the corpus contains scientific text alongside the formulae encoded using Presentation MathML, Content MathML and TeX.

3.2 Preprocessor

Owing to the presence of redundant HTML tags and spaces in the target sentences extracted from Wikipedia documents, the preprocessing of sentences was felt necessary prior to parsing. The job of preprocessor, therefore, is to remove all such redundant tags, spaces, links to footnotes and references, and so on. Table 2 shows sample examples of the preprocessings done by preprocessor.

3.3 Stanford Shift-Reduce Constituency Parser

The Stanford Shift-Reduce Constituency parser maintains the sentence on queue and the parse tree on stack. A set of transitions, namely shift, unary reduce,

[5] www.cs.rit.edu/ rlaz/NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2

binary reduce, finalize and idle, are applied on the current state of the parse tree, unless the queue gets empty and the stack contains complete parse tree. Further, details related to the parser can be seen here[6]. Some sample processed target sentences and their respective parse trees as generated by the parser are shown in Table 4.

3.4 Noun Phrase Extractor

After parsing, the parsed target sentences are fed to Noun Phrase Extractor (NPE), which extracts all different NPs from all the parsed sentences. Some examples of the candidate NPs extracted by the system are shown in Table 3. Up to this stage, a total of 4,919 instances (i.e. formulae, their target sentences, their parsed target sentences and all the candidate NPs present in parsed target sentences) are generated from 500 Wikipedia documents. Next, out of all such 4,919 instances, a set of 100 instances is selected as development set, and another different set of 100 instances is selected as test set. A development set is required to tune weights assigned to different candidate NPs of a target sentence. Moreover, for each instance in development set and test set, a gold context is manually selected from the candidate NPs. While the purpose of selecting gold contexts for development set is tuning of weights, the purpose of selecting gold contexts for test set is testing efficacy of the system for predicting correct context. Table 3 shows some sample entries of the development set. It also shows different candidate NPs from which the gold context is selected.

Table 2. Sample examples of preprocessing done by preprocessor

Original target sentence	**Processed target sentence**
$\langle /dl\rangle$ so Gaussian measure is a Radon measure; is not translation-invariant, but does satisfy the relation $\langle dl\rangle$ $\langle dd\rangle\langle /dd\rangle\langle dt\rangle$	so Gaussian measure is a Radon measure; is not translation-invariant, but does satisfy the relation
The magnetic diffusivity is defined as:$\langle sup\rangle 1\langle /sup\rangle$	The magnetic diffusivity is defined as:
⟨ h3 id=“equation”⟩Equation⟨/h3⟩ The mathematical equation for Boyle’s law is	The mathematical equation for Boyle’s law is

.

[6] https://nlp.stanford.edu/software/srparser.html

Table 3. Sample entries of the development set

Target sentence	Formula	Candidate NPs	Gold context
The units of specific contact resistivity are typically therefore in	$\Omega.cm^2$	• The units of specific contact resistivity • The units • specific contact resistivity	• The units of specific contact resistivity
The level of interaction can be measured by the Gravity model of trade	$I_{i,j} = \frac{p_i.p_j}{d_{i,j}^{\beta}}$	• The level of interaction • The level • interaction • the Gravity model of trade • the Gravity model • trade	• the Gravity model of trade
Often, momentum transfer is given in wavenumber units in reciprocal length	$Q = k_f - k_i$	• momentum transfer • wavenumber units • reciprocal length	• momentum transfer
Assuming infinite planes, the magnitude of the electric field E is	$E = -\frac{\Delta\Phi}{d}$	• infinite planes • the magnitude of the electric field E • the magnitude • the electric field E	• the magnitude of the electric field E
The total Hamiltonian of an atom in a magnetic field is	$H = \mathrm{H}_O + V + M$	• The total Hamiltonian of an atom in a magnetic field • The total Hamiltonian • an atom • a magnetic field	• The total Hamiltonian of an atom in a magnetic field
The following formula approximates the Earth's gravity variation with altitude	$g_h = \mathrm{g}_0 \left(\frac{r_\mathrm{e}}{r_\mathrm{e}+h}\right)^2$	• The following formula • the Earth 's gravity variation with altitude • the Earth 's gravity variation • the Earth 's • altitude	• the Earth 's gravity variation with altitude

Table 4. Target sentences and their respective parse trees

Target sentence	**Parse tree**
The RMSD of an estimator	(ROOT (NP (NP (DT The) (NN RMSD)) (PP (IN of) (NP (DT an) (NN estimator)))))
The level of interaction can be measured by the Gravity model of trade	(ROOT (S (NP (NP (DT The) (NN level)) (PP (IN of) (NP (NN interaction)))) (VP (MD can) (VP (VB be) (VP (VBN measured) (PP (IN by) (NP (NP (DT the) (NN Gravity) (NN model)) (PP (IN of) (NP (NN trade))))))))))
The units of specific contact resistivity are typically therefore in	(ROOT (S (NP (NP (DT The) (NNS units)) (PP (IN of) (NP (JJ specific) (NN contact) (NN resistivity))))(VP (VBP are) (ADVP (RB typically))(ADVP (RB therefore)) (X (IN in)))))
The wave number k is the absolute of the wave vector	(ROOT (S (NP (DT The) (NN wave) (NN number) (NN k)) (VP (VBZ is) (NP (NP (DT the) (JJ absolute)) (PP (IN of) (NP (DT the) (NP (NN wave) (NN vector))))))))

3.5 Weight Assigner

Weight Assigner (WA) uses certain heuristics in assigning weights to different NPs as extracted by the NPE. Specifically, the following heuristics govern weight assignment:

(a) The WA only considers isolated NP (an NP which neither subsumes any NP nor is subsumed by any NP), maximal NP (an NP which subsumes one or more NPs, but is not subsumed by any other NP) and nearest NP (an NP which is nearest to the formula and may or may not be isolated and/or maximal NP) for weight assignment, and the rest candidate NPs are discarded. The isolated or maximal NP, which occurs farthest from the formula, is assigned a weight W_{begin} ($W_{begin} \in \mathbb{R}$). Weight assigned to a subsequent maximal or isolated NP differs from its antecedent by an addend value f ($f \in \mathbb{R}$). More specifically, the weight assigned to second farthest will be $W_{begin} + f$, third farthest will be $W_{begin} + 2f$, and so on.

(b) In some cases, it is observed that the nearest NP itself is the most appropriate context for formula. Therefore, an additional weight, equal to $W_{nearest}$ ($W_{nearest} \in \mathbb{R}$ and $W_{nearest} \geq 0$), is added to the existing weight of nearest NP.

(c) Yet another heuristic is based on the observation that in most cases, if the farthest NP in original set of candidate NPs is an isolated NP, and the second farthest NP is a maximal NP, then the isolated NP is often trivial and the maximal NP is most appropriate context for formula. Consider the below given example to elucidate this point.

Example 3.1:

Target sentence: *The following formula approximates the Earth's gravity variation with altitude*

Formula: $g_h = g_0 \left(\frac{r_\mathrm{e}}{r_\mathrm{e}+h}\right)^2$

Candidate NPs: •The following formula •the Earth 's gravity variation with altitude •the Earth 's gravity variation •the Earth 's •altitude

Gold context: the Earth 's gravity variation with altitude

Here, the farthest NP is *"The following formula"*, which is also an isolated and trivial NP. Furthermore, the second farthest NP is *"the Earth 's gravity variation with altitude"*, which is also a maximal NP and the gold context for formula.
Therefore, under such situation, an additional weight equal to W_{second} ($W_{second} \in \mathbb{R}$ and $W_{second} \geq 0$) is added to existing weight of the maximal and second farthest NP.

Table 5 describes weights assigned to different candidate NPs, using abovementioned heuristics, under different example situations. After the weights are assigned to different candidate NPs of a target sentence, and the weights are tuned using Weight and Addend Tuner (see subsection 3.6), the one having a maximum weight is predicted as the context of corresponding formula.

3.6 Weight and Addend Tuner

As the initial values of three different weight measures (W_{begin}, $W_{nearest}$ and W_{second}) and the addend f may not be optimal, these values need to be tuned to ensure maximum F–score and, hence, optimal context prediction ability. The values of weight measures and addend are tuned using Weight and Addend Tuner (WAT), which compares the system predicted contexts (for development set) against the gold contexts to tune the weights and addend in trial and error fashion. Eventually, the WAT discovers best configuration (best set of values for W_{begin}, $W_{nearest}$, W_{second} and f), which exhibits optimal performance (in terms of F–score) in context prediction.

3.7 System Testing

The efficacy of best configuration selected by WAT is examined using a test set comprising 100 previously unseen test instances. System predicted contexts are compared with the gold contexts of test instances, and the F–score is computed.

Figure 1 shows working principle and different constituents of the implemented system.

Table 5. Weights assigned to different candidate NPs of development set by WA

Target sentence	**Candidate NPs**	**NPs selected by WA for weight assignment**	**Weights assigned**
The level of interaction can be measured by the Gravity model of trade	● The level of interaction ● The level ● interaction ● the Gravity model of trade ● the Gravity model ● trade	● The level of interaction ● the Gravity model of trade ● trade	W_{begin} $W_{begin} + f$ $W_{nearest}$
Often, momentum transfer is given in wavenumber units in reciprocal length	● momentum transfer ● wavenumber units ● reciprocal length	● momentum transfer ● wavenumber units ● reciprocal length	W_{begin} $W_{begin} + f$ $W_{begin} + 2f + W_{nearest}$
Assuming infinite planes, the magnitude of the electric field E is	● infinite planes ● the magnitude of the electric field E ● the magnitude ● the electric field E	● infinite planes ● the magnitude of the electric field E ● the electric field E	W_{begin} $W_{begin} + f + W_{second}$ $W_{nearest}$
The total Hamiltonian of an atom in a magnetic field is	● The total Hamiltonian of an atom in a magnetic field ● The total Hamiltonian ● an atom ● a magnetic field	● The total Hamiltonian of an atom in a magnetic field ● a magnetic field	W_{begin} $W_{nearest}$
Rectangular hyperbolas have eccentricity	● Rectangular hyperbolas ● eccentricity	● Rectangular hyperbolas ● eccentricity	W_{begin} $W_{begin} + f + W_{nearest}$
The Rydberg constant is seen to be equal to	● The Rydberg constant	● The Rydberg constant	$W_{begin} + W_{nearest}$

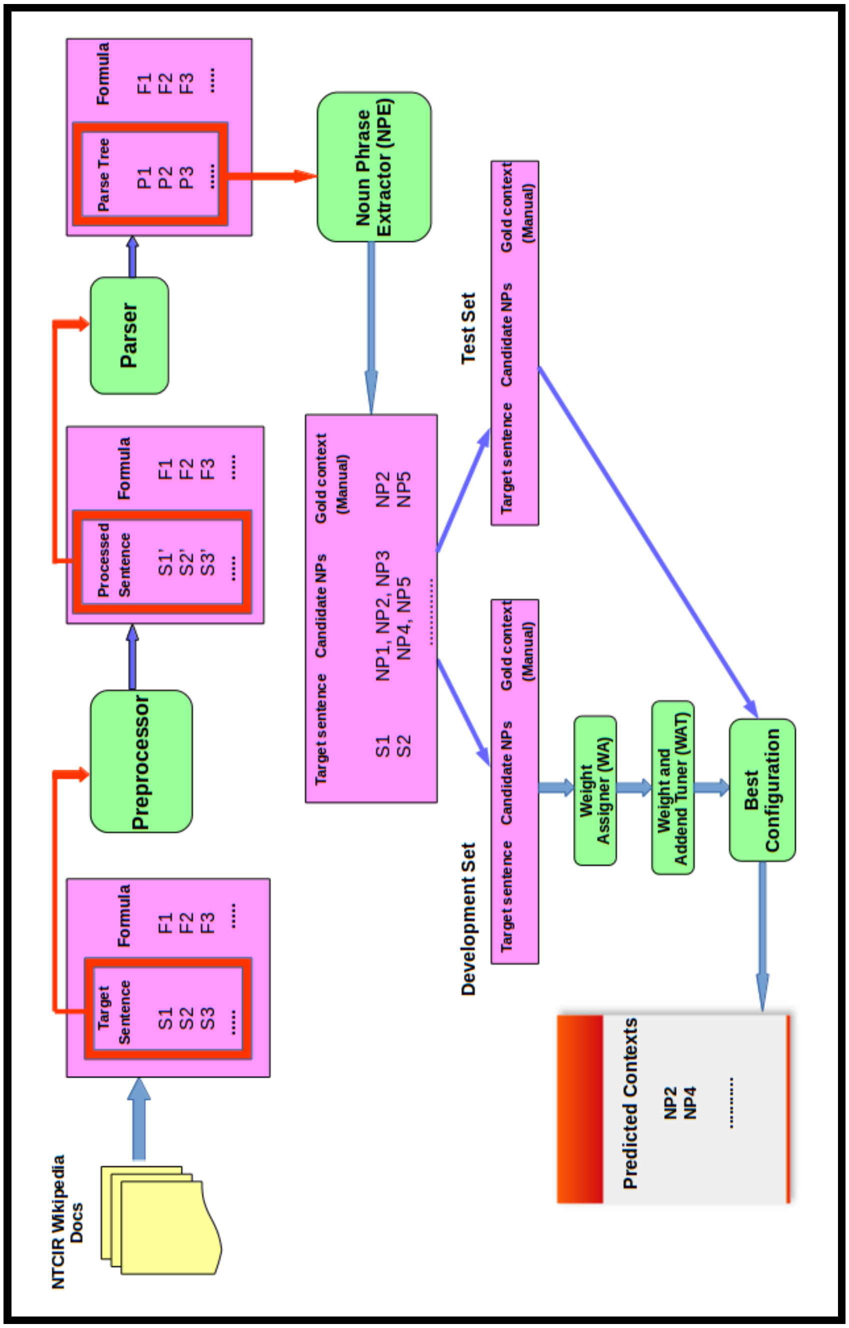

Fig. 1. Overall architecture of the implemented system

4 Experimental Design

To develop and test the system, following experimental setups are employed:

(a) As mentioned in previous section, the system predicted contexts are evaluated on the ground of F–score. However, to compute F–score (see Equation 3), precision and recall measures need to be computed beforehand. While precision (see Equation 1) gives a measure of the correct context predictions out of the total contexts predicted, recall (see Equation 2) gives a measure of correct context predictions out of the total gold contexts.

$$\text{Precision (P)} = \frac{\text{Count of correct context predictions}}{\text{Count of predicted contexts}} \tag{1}$$

$$\text{Recall (R)} = \frac{\text{Count of correct context predictions}}{\text{Count of gold contexts}} \tag{2}$$

$$\text{F–score} = \frac{2 * P * R}{P + R} \tag{3}$$

(b) Also, as defined in previous section, the weight W_{begin} and the addend f may attain any real value during tuning, whereas the weights W_{second} and $W_{nearest}$ only attain either 0 or positive real values.

5 Results and Analysis

System results and their comprehensive analysis are presented in this section. Following points regarding the implemented system and the predicted contexts are worth noting:

(a) System attains a maximum development set F–score of 67% for the following values of weights and addend: $W_{begin} = 0.6$, $W_{second} = 0.4$, $W_{nearest} = 0.1$ and $f = -0.2$. A negative value of addend f is indicative of the fact that the value of weight W_{begin} is decremented by 0.2, every time, on going from farthest maximal/isolated NP to the nearest maximal/isolated NP.

(b) F–scores of the implemented system for development set and test set are 67% and 65%, respectively. The system predicts contexts (either correct or incorrect) for all the instances in development and test sets. Some sample examples of gold contexts, predicted contexts and their associated weights are shown in Table 6. For some of the test target sentences, there are more than one probable gold contexts (see example 5 of Table 6). In all such cases, the predicted context is considered relevant, if it is any one of the gold contexts.

(c) To assess the comparative strengths, the performance of proposed system is compared with those of nearest NP and sentence–pattern based methods (see Table 1 of Related Works section for different sentence patterns). Figure 2 shows F–scores of the three methods for development and test sets, and

Table 6. Sample examples of gold contexts and predicted contexts

Target sentence	Gold context	Predicted context (weight)
Using this free-body diagram the torque required to lift or lower a load can be calculated:	the torque required to lift or lower a load	the torque required to lift or lower a load (0.8)
In its general form, the steering law can be expressed as	the steering law	the steering law (0.9)
The gravity depends only on the mass inside the sphere of radius	radius	the mass inside the sphere of radius (0.8)
The Young Equation relates the contact angle to interfacial energy	The Young Equation	The Young Equation (0.6)
The BEST theorem states that the number $ec(G)$ of Eulerian circuits in a connected Eulerian graph G is given by the formula	(a) The BEST theorem (b) the number $ec(G)$ of Eulerian circuits in a connected Eulerian graph G	the number $ec(G)$ of Eulerian circuits in a connected Eulerian graph G (0.8)

the scores are indicative of the fact that the proposed method significantly outperforms the two conventional and naive methods in predicting most appropriate context for a given formula. While the nearest noun method assumes nearest NP to be the context of formula, the sentence-pattern based method assumes presence of certain patterns between context and formula. Such naive assumptions need not always be correct, and hence the poor performance.

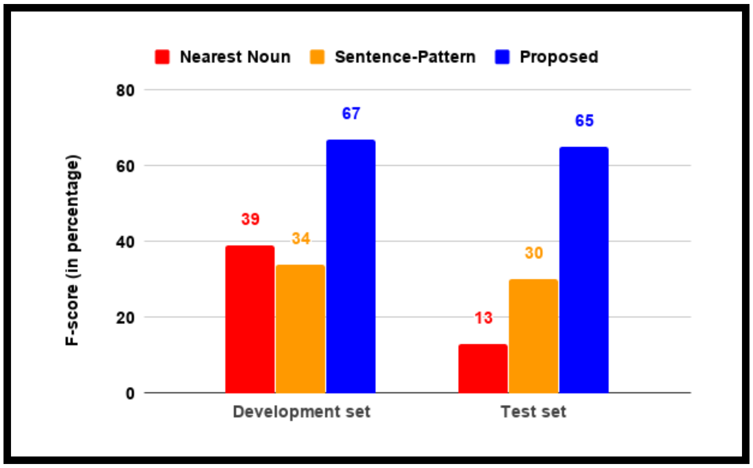

Fig. 2. Performace comparison of proposed method

(d) Different number of candidate NPs are extracted by the implemented system for different target sentences in development and test sets. The graph shown in Figure 3 shows statistical distribution of number of candidate NPs in the two sets, which can be interpreted as follows: *"3 candidate NPs are extracted for 15 target sentences in development set and 32 target sentences in test set. Similarly, 2 candidate NPs are extracted for 7 target sentences in development set and 4 target sentences in test set."* The plot is indicative of the fact that the two sets vary in terms of the instances corresponding to different number of candidate NPs. To further ascertain correlation, if any, between development and test sets, Pearson correlation coefficient (r) is computed between number of instances in two sets corresponding to different number of candidate NPs. The value of r equal to 0.387 (more close to 0 than 1) confirms that the two sets are almost uncorrelated. Also, even though the development and test sets are almost uncorrelated, the implemented system delivers comparable performance. This confirms that the performance of system is independent of the nature of target sentences and, hence, the system is neither overfit nor underfit.

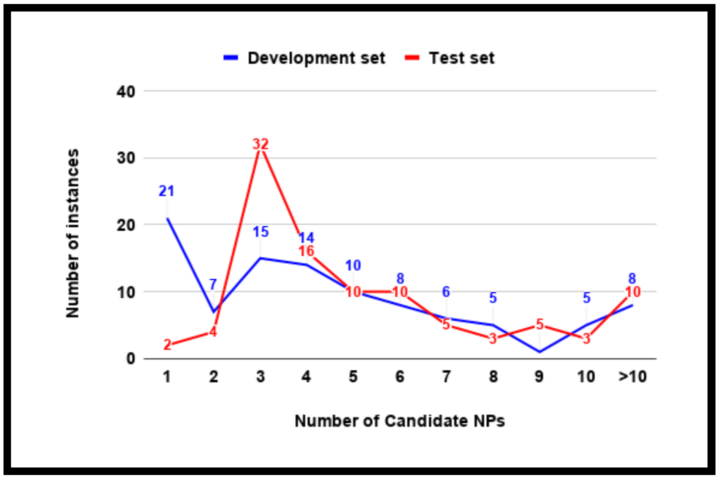

Fig. 3. Statistical distribution of number of candidate NPs in development and test sets

5.1 Error Analysis

Even though the proposed approach works effectively for substantial variety of test instances, the following shortcomings are worth considering:

(a) The system incurs failure in situations where the gold context is absent in the sentence containing formula. In example 5.1 below, the system incorrectly infers one of the NPs of target sentence to be the correct context, as the gold context (i.e *"Rayleigh criterion"*) is present in some previous sentence and not in the sentence containing formula. Although it may be accounted by extending the window size from a single sentence to multiple sentences or even complete passage, such an attempt will lead to significant increase in number of NPs which are unrelated to the formula.

Example 5.1:

Target sentence: *If one considers diffraction through a circular aperture, this translates into:*

Formula: $\theta = 1.220\frac{\lambda}{D}$

Gold context: Rayleigh criterion

Predicted context: a circular aperture

(b) System lacks ability to combine different NPs in a systematic or meaningful way. More specifically, the system fails to account for situations wherein the gold context is intricate combination of two or more candidate NPs. For instance, in example 5.2 below, the gold context is combination of two candidate NPs, namely *"the pressure force"* and *"an isothermal fluid"*, but the system incorrectly predicts only partial gold context.

Example 5.2:

Target sentence: *For an isothermal fluid, the pressure force takes the form*

Formula: $F_{fluid} = -k_B T_e \Delta n_e$

Gold context: the pressure force for an isothermal fluid

Predicted context: isothermal fluid

(c) The system also incurs failure if the gold context does not adhere to the heuristics as mentioned in subsection 3.5. For instance, WA assigns weights only to maximal NP, isolated NP and nearest NP, and discards all other candidate NPs. However, in some cases, the gold context may be one of the minimal NPs, instead of being maximal NP, isolated NP or nearest NP. The gold context *"initial momenta"* is a minimal NP in example 5.3 below.

Example 5.3:

Target sentence: *In the simplest example of scattering of two colliding particles with initial momenta of the form*

Formula: $\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}$

Candidate NPs: •the simplest example of scattering of two colliding particles with initial momenta of the form •the simplest example •scattering of two colliding particles with initial momenta of the form •scattering •two colliding particles with initial momenta of the form •two colliding particles •initial momenta •the form

Gold context: initial momenta

Predicted context: the simplest example of scattering of two colliding particles with initial momenta of the form

6 Conclusion and Future Directions

This paper proposes and implements a system, which can extract textual description (also called *"context"*) of math formula present inside scientific document. Preprocessor, Shift–Reduce Constituency Parser, Noun Phrase Extractor (NPE), Weight Assigner (WA), and Weight and Addend Tuner (WAT) form key constituents of the implemented system, which work in a sequential fashion to predict most appropriate context for a given formula. After the sentences containing formula (target sentences) are processed and parsed, different Noun Phrases (NPs) are extracted from the parse trees using NPE. Eventually, the WA assigns weights to different NPs using certain heuristics, and the WAT tunes values of those weights using a development set containing gold contexts for target sentences. Thereafter, the best set of weight values are used to predict context for test target sentences. The proposed method achieves a test set F–score of 65% and significantly outperforms the conventional nearest noun and sentence–pattern based methods of context extraction.

Followings are some of the future research directions worth exploring:

(a) The WAT, as of now, uses trial and error to discover the best set of weight and addend values. Instead, in the future, multiple linear regression will be used to express F–score in terms of weights and addend. Subsequently, the constrained multivariable optimization, with constraints being $W_{nearest} \geq 0$ and $W_{second} \geq 0$, will be used to discover the optimal values of weights and addend for which the F–score will be maximum.

(b) Furthermore, it will be interesting to prospect the impacts of followings over performance of the system: (i) increase in development set size (ii) enabling support for judiciously combining different candidate NPs, and (iii) extending the context window size from a single sentence to multiple sentences.

Acknowledgement

The work presented here falls under the Research Project Grant No. YSS/2015/000988 and supported by the Department of Science & Technology (DST) and Science and Engineering Research Board (SERB), Govt. of India. The authors would like to acknowledge the Department of Computer Science & Engineering, National Institute of Technology Mizoram, India for providing infrastructural facilities and support.

References

1. Aizawa, A., Kohlhase, M., Ounis, I.: NTCIR-10 Math Pilot Task Overview. In: Proceedings of the 10th NTCIR Conference. pp. 654–661. Tokyo, Japan (2013)
2. Aizawa, A., Kohlhase, M., Ounis, I., Schubotz, M.: NTCIR-11 Math-2 Task Overview. In: Proceedings of the 11th NTCIR Conference. pp. 88–98. Tokyo, Japan (2014)
3. Buswell, S., Caprotti, O., Carlisle, D.P., Dewar, M.C., Gaetano, M., Kohlhase, M.: The Open Math standard. Tech. rep., version 2.0. Technical report, The Open Math Society (2004)
4. Goldberg, Y., Nivre, J.: A Dynamic Oracle for Arc-Eager Dependency Parsing. Proceedings of COLING 2012 pp. 959–976 (2012)
5. Grigore, M., Wolska, M., Kohlhase, M.: Towards context–based disambiguation of mathematical expressions. In: The joint conference of ASCM. pp. 262–271. Fukuoka, Japan (2009)
6. Joho, H., Kishida, K.: Overview of NTCIR-11. In: Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. pp. 1–7. Tokyo, Japan (2014)
7. Joho, H., Sakai, T.: Overview of NTCIR-10. In: Proceedings of the 10th NTCIR Conference. pp. 1–7. Tokyo, Japan (2014)
8. Kishida, K., Kato, M.P.: Overview of NTCIR-12. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. pp. 1–7. Tokyo, Japan (2016)
9. Kristianto, G.Y., Aizawa, A., et al.: Extracting Textual Descriptions of Mathematical Expressions in Scientific Papers. D-Lib Magazine **20**(11), 1–9 (2014)
10. Kristianto, G.Y., Nghiem, M.Q., Matsubayashi, Y., Aizawa, A.: Extracting Definitions of Mathematical Expressions in Scientific Papers. Proceedings of the Annual Conference of JSAI **JSAI2012**, 1–7 (2012)
11. Líška, M., Sojka, P., Ružicka, M.: Similarity Search for Mathematics:Masaryk University team at the NTCIR-10 Math Task. In: Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies. pp. 686–691. Tokyo, Japan (2013)

12. Natho, N., Jeschke, S., Pfeiffer, O., Wilke, M.: Natural language processing methods for extracting information from mathematical texts. In: Advances in Communication Systems and Electrical Engineering, pp. 297–308. Springer (2008)
13. Pagael, R., Schubotz, M.: Mathematical Language Processing Project. arXiv preprint arXiv:1407.0167 (2014)
14. Pathak, A., Pakray, P., Gelbukh, A.: A Formula Embedding Approach to Math Information Retrieval. Computación y Sistemas **22**(3), 819–833 (2018)
15. Pathak, A., Pakray, P., Sarkar, S., Das, D., Gelbukh, A.: MathIRs: Retrieval System for Scientific Documents. Computación y Sistemas **21**(2), 253–265 (2017)
16. Quoc, M.N., Yokoi, K., Matsubayashi, Y., Aizawa, A.: 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). pp. 69–74. Beijing, China (2010)
17. Ruzicka, M., Sojka, P., Líska, M.: Math Indexer and Searcher under the Hood: Fine-tuning Query Expansion and Unification Strategies. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. pp. 331–337. Tokyo, Japan (2016)
18. Schubotz, M., Grigorev, A., Leich, M., Cohl, H.S., Meuschke, N., Gipp, B., Youssef, A.S., Markl, V.: Semantification of Identifiers in Mathematics for Better Math Information Retrieval. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 135–144. ACM, Pisa, Italy (2016)
19. Trzeciak, J.: Writing Mathematical Papers in English: A Practical Guide. European Mathematical Society (1995)
20. Yokoi, K., Nghiem, M.Q., Matsubayashi, Y., Aizawa, A.: Contextual Analysis of Mathematical Expressions for Advanced Mathematical Search. Polibits (43), 81–86 (2011)
21. Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., Davila, K.: NTCIR-12 MathIR Task Overview. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. pp. 299–308. Tokyo, Japan (2016)
22. Zhu, M., Zhang, Y., Chen, W., Zhang, M., Zhu, J.: Fast and Accurate Shift-Reduce Constituent Parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 434–443. Sofia, Bulgaria (2013)