Prediction of Mathematical Expression Constraints (ME-Con)

Jason Lin, Xing Wang, Jyh-Charn Liu
Department of Computer Science and Engineering
Texas A&M University, College Station, TX 77843, USA
{senyalin, wxadb, jcliu}@tamu.edu

ABSTRACT

Mathematical Expressions (MEs) are routinely blended into plaintext sentences, or ME-plaintext (MEP) sentence, in technical papers. This paper presents a novel algorithm for predictive labelling of *mathematical constraints* (ME-Con) in publications, based on analyses of mathematical symbols F_S and the part-of-speech (POS) tags in the neighboring plaintext F_C. An example of the ME-Con is the range/scope of a variable or a function in an ME. An optimal subset of F_S and F_C for ME-Con labeling are derived from heuristic rules to achieve the F1 scores of 68.8% and 68.5%, respectively. Then, based on the naïve Bayesian classifier, the posterior prediction performance of ME-Con based on F_S is boosted to 82.1% for its F1 score, and 18.3% of error rate when the algorithm is used to analyze a public dataset of ME-plaintext mixed sentences provided by Elsevier. There is no noticeable performance gain for F_C for the Bayesian classifier due to the high flexibility of ME within the same context.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Software and its engineering** → LaTeX

KEYWORDS

Constraint, mathematical expression, prediction

1 INTRODUCTION

Automated discourse analysis of technical publications is useful for developing machine reading ability of technical papers for extraction of crucial scientific information at large scale. Most existing discourse analysis of scientific papers relies on natural language processing (NLP) tools, which are mainly focused on analysis of plaintext, with much less analysis on MEs and their semantic ties to plaintexts. With some basic training, human readers in technical fields can reliably distinguish discourse types such as definition, assumption, constraint, and condition; but the same cannot be said for machine readers. In this work, we develop a learning based predictive model for labelling of ME related constraints or conditions, ME-Con.

ME-Con represents the quantitative boundaries or conditions of variables in an ME for the subject topic. Every ME has its constraints, but they are often in the implicit form after they are declared explicitly first time. Explicit declaration needs to be made each time when the constraints are adjusted. Our goal is to detect those explicit declaration, so that the results can aid readers in capturing the mathematical framework more easily.

For our purpose, ME-Con may include but not limited to the following five types: (1) condition type (e.g., "$x \geq 1$", "Let $x = 1$, the base case holds"), (2) index (range) type (e.g., e_i for $i \in [lhs, rhs]$), (3) set type ($S' \subseteq S$), (4) enumerative type (e.g., the number of permuta- tions of n elements is $n!$), and (5) performance bound type (e.g., $O(n)$).

Labelling of ME-Con is involved mathematical symbols (e.g., "|" within "{ }") that commonly depict a constraint statement, or the part-of-speech (POS) tag that indicates a ME-Con, e.g., "where" [Adverb], "greater than" [Adjective Pronoun]), and their combinations (e.g., "where … {…|…}").

For the design of pattern learning and matching algorithms, these patterns can be trained from annotated samples to form regular expression (regex) based rules. We proposed a heuristic approach and a naïve Bayesian classification model based upon the frequency of mathematical symbols (ME-level features, F_S) and the POS of the local neighbor plaintext (Contextual-level features, F_C) in a sentence.

The statistical patterns of F_S and F_C are trained from annotated dataset [12] for design of the prediction rules. As an observation, F_S is much more indicative of the ME-Con than F_C is. Though has a few indicative features for ME-Con, but they will also introduce many false positives, leading to lower precision. Empirical results suggest that ME-Con could be freely expressed in different syntactical forms, while certain mathematical symbols convey the constraint semantics. The detection outcomes can be used for initialization of most likely ME role in discourse analysis.

2 RELATED WORKS

Several works [1][2][3][4] have defined the word "constraint" as properties, attributes and attribute values of a subject, and relational properties between subjects. Closely related to information extraction, constraint analysis has been long studied for event extraction [5], for name entity extraction [6], within the biomedical domain [7], and in automated information extraction rules generation [8][10]. Resources, structure, hierarchy, and dependency are consider the main constraints in system development documents [9]. Existing constraint extraction techniques are largely based on a mix of keyword matching, NLP tools. For example, the work in [4] has patented a NLP constraint extraction system for the generation of testing data. They proposed template matching approach to identify words from their POS tag

to capture the three elements: subject, object, and condition. Since it is patented, no further experimental results or specific details are provided. To date, no known work has been done for ME related constraint analysis.

3 ME-CON LABELING

Labeling of ME-Con consists of the preprocessing step for feature extraction (subsection 3.1.), labeling heuristic (subsection 3.2), followed by a naïve Bayesian (subsection 3.3) to produce the final result.

3.1 Feature Extraction

We divide the features into ME-level and contextual-level extraction. Features that are indicative of ME-Con can be further refined as being context-dependent or context-independent. Context-dependent constraints are essentially based on cue words or symbols, so called the *constrainator* in an analysis window. On the contrary, context-independent constraints contains strong indicator such as constraint operator, comparative adjective words, etc., that indicate presence of an ME-Con.

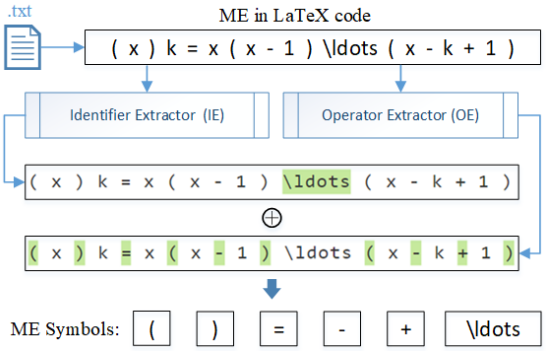

Figure 1: Mathematical symbol extraction

For ME-level features, we employed a regular expression (regex) parser [11] to parse the LaTeX annotated MEs to extract mathematical symbols. Noted that LaTeX encodes special mathematical symbols and characters by backslash plus some reserved letters. For example, "\geq" represents the symbol "≥". The input math part of LaTeX string will go through the identifier extractor (IE) by regex "(\\[A-Za-z]+)" and the operator extractor (OE) by regex "([-!%&*()+|~=\[\]\':;<>?,.\/]|\\[{}])" to retrieve the non-typable and typable symbols, respectively. Then, we aggregate their results to obtain all possible mathematical symbols in the given string, respectively, as the example shown in Fig. 1.

The contextual-level analysis is based on the assumption that ME-Con appears in certain location with respect to the syntactic roles of its surrounding plaintexts. Instead of trying to enumerate the massive number of possible word choices, we extracted the POS tags of the two adjacent words for each ME as features, and took the POS tag of words in the annotated Elsevier dataset [12] for model training. The co-occurrence and the order of the POS tag pair are also used as two additional features. An ME has only one POS tag based feature when it is adjacent to another ME, or is located at the beginning or ending position of a sentence. The type of adjacent punctuation is also used as a tag. Fig. 2 shows an example for "Let $n \geq 2$.". Noted that we take each tag with substring "FRM" in the dataset as the ME chunk for our analysis.

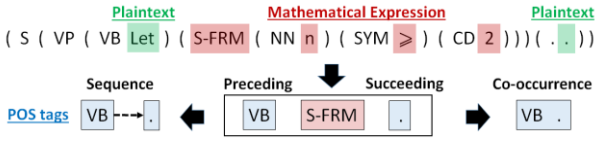

Figure 2: An example of the contextual POS features extraction.

3.2 Heuristic Rule for ME-Con Labeling

A human reader uses the combination of symbols and adjacent words as cues to assert if an ME represents a mathematical constraint. However, such cues are not deterministic due to the conditional, multifaceted semantics of these ME and plaintexts. Based on this observation, we believe there must exist an optimal subset X^* of symbols and POS patterns that can assist us to detect the ME-Con. Hence, we proposed a heuristic rule-based approach to automatic iteratively learned from the four metrics: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) of the confusion matrix.

For convenience in the following discussion, we denote the MEs and the plaintexts of a MEP as set $M = \{m_1, \dots, m_{|M|}\}$ and set $P = \{p_1, \dots, p_{|P|}\}$, respectively, where the cardinal number of $|M|$ and $|P|$ represents the amount of MEs and plaintext in the given MEP. We further define the contextual-level feature F_C as a binary vector $\vec{w} = \{b_t\}$ where $b_t = 1$ if the contextual POS pattern t exist in the given ME. The ME-level feature F_S of a given ME is defined as another n-dimentional binary vector $\vec{u} = \{b_1, \dots, b_n\}$ where n is the number of all possible symbols used in MEs. Finally, let the binary variable $\mathcal{L}(m_i)$ represents the constraint semantic label of an input ME $m_i \in M$. That is, m_i is ME-Con if and only if $\mathcal{L}(m_i) = 1$.

During the process of the approach, each feature vector $\vec{v}$ and feature set T of any ME will be jointly assigned to one of the four classification sets (X_{TP}, X_{FP}, X_{TN}, and X_{FN}) based on the labeling result of ground truth and prediction. The classification set is defined as a binary vector $\vec{x}_j = \bigcup_{j=k} \vec{v}_k$ where $\vec{v} \in \{\vec{u}, \vec{w}\}$ and $j, k \in \{TP, FP, TN, FN\}$. The optimal subset is learn by the heuristic of taking the all possible features that appeared in either TP or FN instances, then subtracts the common features appeared in the instances of FP and TN. More specifically, $X^r = (X_{TP}^{r-1} \cup X_{FN}^{r-1}) - (X_{TN}^{r-1} \cap X_{FP}^{r-1})$ where r is the round number of iteration and $S^0 = \emptyset$. The positive predictions of ME-Con are made upon $\vec{x} \wedge \vec{v} \neq \vec{0}$. Note that in the initial stage, all predictions are set to 0 due to $X^* = \emptyset$ ($\vec{x} = \vec{0}$).

3.3 Bayesian Model for ME-Con Labeling

Following some definitions in subsection 3.2, we also construct a naïve Bayesian model that considers the likelihood of individual feature with the assumption of conditional independence when doing the ME-Con assessment upon features F_S and F_C, respectively.

Let π_i be a ME in a MEP and its preceding element π_{i-1} and the succeeding element π_{i+1} are used for calculates the contextual-level features of π_i. We would like to estimate the posterior probability of ME-Con $\theta = \mathcal{L}(\pi_i)$ condition on the likelihood of mathematical symbols in π_i, and the contextual features derived by the POS tag of π_{i-1} and π_{i+1}, respectively. The estimated function can be expressed as $Pr(\theta|e(\pi_i)) = \frac{Pr(e(\pi_i)|\theta)Pr(\theta)}{Pr(e(\pi_i))}$ where $\theta \in \{0,1\}$ and $e(\pi_i)$ is an evidence set of features for ME π_i in either one of the type F_S or F_C. With assumption of conditional independence, we can obtain the conditional probability $Pr(e(\pi_i)|\theta) = \prod_j Pr(e(\pi_i)_j|\theta)$. Note that the posterior probability $(\theta|e(\pi_i)) \approx Pr(e(\pi_i)|\theta)Pr(\theta)$, and $Pr(\theta)$ is the prior probability of ME-Con as positive instances in the training data. We can derive the labeling result by the likelihood ratio $\rho = \frac{Pr(e(\pi_i)|\theta=1)}{Pr(e(\pi_i)|\theta=0)}$ where the predicted $\theta = 1$ (ME is a ME-Con) if $\rho > 1$; Otherwise, $\theta = 0$ (ME is not a ME-Con).

4 EXPERIMENT

In this section, we will first introduce the data set and the evaluation criteria used in our experimental process. We will then explore the relation between mathematical symbols and ME-Con. Third, the POS tag of the contextual neighboring words are explored to further discuss its usefulness. Finally, we show the performance results on the two proposed model by the widely used indices precision, recall, and F1 score for classification performance. Then, we further discuss the possible reasons why naïve Bayesian outperforms than the heuristic approach.

4.1 Dataset and Evaluation Criteria

OS-STM-Elsevier [12] provides 10 papers with all 2757 sentences annotated in Penn tree format [13]. Each paper is from a different STEM (Science, Technology, Engineering and Math) field. Each ME is labeled as "*-FRM", where the "*" could be replaced with a more specific label based on its syntactic role, i.e., noun-phrase (NP), sentence or subordinate clause (S), and noun modifier (NML). Among all sentences, there are 346 sentences contain both MEs and plaintexts. About 600 MEs are in these 346 sentences. All MEs are manually annotated with 'yes', 'no', or 'uncertain' that whether they are expressing a constraint semantic. Both the mathematical symbols and the local POS tag features were apply to all 600 MEs out of 346 sentences.

4.2 ME-Con Properties and Distribution

The likelihood of each mathematical symbol that appeared in each ME which supports the positive instance of ME-Con is calculated, and we found that some symbols as shown in Fig. 3 tends to be strong indicator in determining whether a ME is ME-Con, which is consistent to human cognitive process of interpreting the semantic of ME. These symbols include, for instances, the asymptotic notions (e.g., ω, Ω, O) used for time complexity in algorithm design, comparative operators such as $\geq$ and $\leq$ describe the value boundary of a variable.

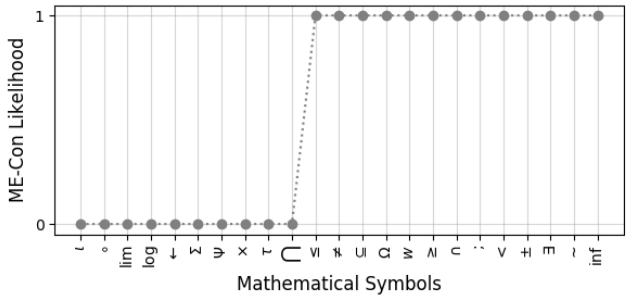

Figure 3: Strong indicators for ME-Con.

On the other hand, some mathematical symbols easily cause ambiguity of ME semantics. These symbols are considered softer indicators that exists likelihood with the appearance of such symbol that contributes the positive side of ME-Con as in Fig. 4. One of the common operator such as the equals sign (=) usually used as comparative operator to express equivalent between two mathematical entities, but it also uses as assignment or definition to introduce a newly arrival mathematical entity. Another example would be the braces "{}". When accompanied by the bar "|" or colon ":", they typically imply that there are constrains or condition defined inside the ME. These indicators adjust the likelihood of our Bayesian model even when strong indicators appear in the instance. Therefore, a more precise decision can be made according to the likelihood.

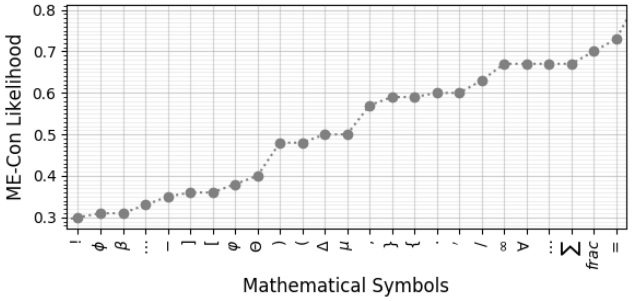

Figure 4: Soft indicators for ME-Con (Likelihood 30% ~ 80%).

4.3 Local Neighbor Plaintext Features

For the study of contextual features, we took both the preceding and the succeeding POS tags of each ME, and computed the differences between their positive and negative sample rates of ME-Con as shown in Fig. 5. Note that for those POS tags with rate

differences closed to zero imply less discriminative power for ME-Con, and the others are either positive or negative in supporting the ME-Con. For the POS of the preceding word, we observed that an ME following a left bracket "LRB" and preposition word with POS "IN" has a relatively high likelihood to be a ME-Con. This is because technical writing very often places a condition inside the brackets as a supplement to the description. MEs following preposition "IN" are also more likely to be ME-Cons as we can often see elaboration such as "with " or "for ". However, an ME following noun (with POS "NN" and "NNS") is more likely not ME-Con because the MEs is usually the apposition of the previous noun phrases. On the other hand, from view of the POS of the succeeding word, we find that ME followed by punctuations such as "." and ",", or right bracket "RRB" are more likely to be ME-Con. This is because any constraint or condition statement is usually made at the end of the statement to enlist other conditions. Moreover, a noun "NN" or a verb "VBZ" in a MEP does not usually follow a ME-Con, where the ME usually play the role of noun modifier or subject.

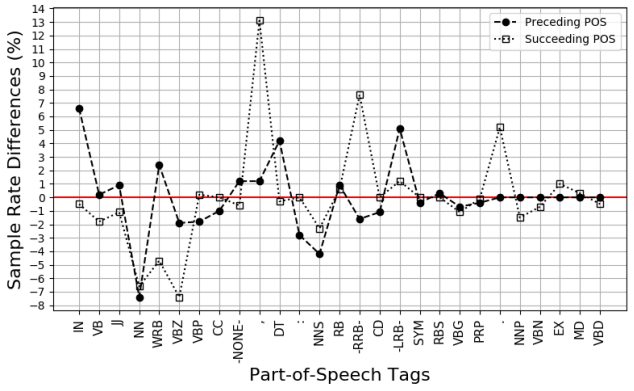

Figure 5: Sample Rate Differences of ME-Con and POS.

Those indicators mentioned above are reasonable in identifying the ME-Con, which could explain the good recall rate in the Table 1. However, they could also contribute to false positives, leading to low precision. For the preceding POS as preposition "IN", we could retrieval 38% of the ME-Con, but we also make wrong prediction of 31.4% of none ME-Con. Similar arguments could be made for the preceding POS as "DT" and "WRB", and succeeding POS as "." and ",". Besides simply examining the POS tag on the left-hand-side and right-hand-side, we also examine the co-occurrence of both the preceding POS and succeeding POS. We found the relatively more cases in the ME-Con are pairing the POS with preposition "IN" and punctuation or both as prepositions, but no significant high frequency cases in the study among all instances. Hence, no relatively strong indicators are found for ME-Con in the contextual words.

4.4 ME-Con Detection Performance

We adopted the 10-fold cross validation to evaluate the performance of our models. The performance of the classifier is assess based on the average of three metrics: precision (P), recall (R), and F1 score.

From Table 1, we found that the heuristic model has a recall of 100%, but nearly 54% and 53.5% precision on both features. This indicates that the heuristic approach is sensitive to all possible combinations of symbols or contextual POS tag patterns. However, it has a difficult time precisely locating the real samples for ME-Con. Fortunately, the proposed naïve Bayesian model outperforms the heuristic approach with a precision of 80.2% and boost up the F1 score to nearly 81.4%. Since the naïve Bayesian model assumes a conditional independence on the features, it suggests us that the determination of ME as a ME-Con is actually consistent to human cognitive process, which relies on some strong indicators to make decision on what the ME actually mean. Such strong indicator are more on the mathematical symbols than the contextual syntactic rule since the features in local neighbor POS have a relatively low performance of 68.5% and 69.8% on both approaches, respectively.

Table 1: Average Performance Summary of Classifier

	Mathematical Symbols			**Local Neighbor POS tags**		
	P	**R**	**F1**	**P**	**R**	**F1**
Heuristic Rule	54.0%	100%	68.8%	53.5%	100%	68.5%
Naïve Bayesian	**80.4%**	85.9%	**82.1%**	57.9%	91%	69.8%

5 CONCLUSION

In this work, we propose a heuristic model and a naïve Bayesian model for ME to determine whether its semantic belong to a ME-Con. Results show effectiveness of the mathematical symbols as feature along with independence assumption in identifying the ME-Con. The lowest error rate of ME-Con prediction thus far is under naïve Bayesian with 18.3%, which is reliable enough to integrate with the existing semantic parsing tools as a filter.

Since this is a preliminary study, more experiments are required to validate the conclusion of this work in a larger data set. The future study will consider on more high-level syntactic role (e.g., Clauses and Phrases) of plaintexts, and high-resolution syntactic role and structural properties of MEs to discover new feature pattern in constraint semantic.

REFERENCES

[1] Lok, Simon, and Steven Feiner, "A survey of automated layout techniques for information presentations," In Proceedings of SmartGraphics, pp. 61-68 (2001)

[2] Ailomaa, Marita, and Martin Rajman, "Natural language techniques for model-driven semantic constraint extraction," No. LIA-REPORT-2007-001, (2007)

[3] Wei, Dengping, Ting Wang, Ji Wang, and Yaodong Chen, "Extracting semantic constraint from description text for semantic web service discovery," The Semantic Web-ISWC 2008, 146-161 (2008)

[4] Misra, Janardan, Milind Savagaonkar, Neville Dubash, Sanjay Podder, and Sachin Hanumantappa Waddar, "Constraint extraction from natural language text for test data generation," U.S. Patent Application No. 14/990,051. APA, (2016)

[5] Grishman, Ralph, and Beth Sundheim, "Message understanding conference-6: A brief history," COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, Vol. 1, (1996)

[6] Nadeau, David, and Satoshi Sekine, “A survey of named entity recognition and classification,” Lingvisticae Investigationes 30(1), 3-26, (2007)
[7] Miwa, Makoto, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou, “Extracting semantically enriched events from biomedical literature,” BMC bioinformatics 13(1), 108 (2012)
[8] Califf, Mary Elaine, and Raymond J. Mooney, “Relational learning of pattern-match rules for information extraction,” In AAAI/IAAI, pp. 328-334 (1999)
[9] Campbell, Robert L., and Mark H. Bickhard, “Types of constraints on development: An interactivist approach,” Developmental Review 12(3), 311-338 (1992)
[10] Soderland, Stephen, “Learning information extraction rules for semi-structured and free text,” Machine learning 34(1), 233-272 (1999)
[11] Wang, Xing, Jason Lin, Ryan Vrecenar, and Jyh-Charn Liu. "Syntactic role identification of mathematical expressions." In Digital Information Management (ICDIM), 2017 Twelfth International Conference on, pp. 179-184. IEEE, 2017.
[12] Jason Lin, Xing Wang, “Manual Constraint Labeling for ME in [12], web posting: http://rtds.cse.tamu.edu/resources/, will be posted in Apr. 2018. Elsevier OA STM Corpus (http://elsevierlabs.github.io/OA-STM-Corpus/), January 11, 2015.
[13] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19, no. 2 (1993): 313-330.