Annotation Guidlines for Inline Mathematical Expressions

Body texts in scientific papers (titles, abstracts, captions, section titles, etc.) are divided into paragraphs, each of which is annotated with any spans of mathematical expressions. (Note that independent lines of mathematical expressions, contents of figures or tables, any level of (from pseudo to raw) program codes, algorithms, etc. are excluded from the targets of annotation.)

Fundamental policy:

- Annotate every minimal region which (structurally and semantically) satisfies the following conditions:
  (1) the region can be recognized to play a role for (structurally/semantically) composing a natural language sentence containing the region
  (2) the region has a mathematically-closed (independent) structure
- Cover any structures and notations specific to mathematics, which do not appear in texts of neutral domains

Each maximal text-region of mathematical expressions which satisfies the above policy is annotated as an “inline mathematical expression”. The following Cases A-E provide the guidelines for the way how each span of “inline mathematical expressions is decided”.

Cases:

A. Variables and Operators (mathematical minimal units of expressions)

   - A certain sequence of body text can be detected as mathematical variables or operators if the sequence is:
     (1) defined somewhere in the body text to specify some mathematical entity*, and
     (2) the sequence at the target location can be judged to specify the same entity as the definition detected in (1)

     * Valid cases for (1)
       - The sequence is explicitly defined somewhere in body text to specify some entity
         (Definitions without any intention for mathematical usage (e.g. definitions such as just for abbreviation or labeling)
       - - The sequence is utilized somewhere in a target paper for composing a mathematical expression.
       - (Even if the above two cases are not observed,) the superficial features of the sequence itself can be judged to follow the characteristic features of usual mathematical variables or operators.

B. Mathematically-structured expressions

   (0) Presupposition

- A numerical value (with/without a unit prefix/postfix) itself is not taken as a mathematical expression

(1) Mathematical Operations or Relations Expressed by Operators

- Mathematical expressions are connected by an apparent opearator or one detected in A.
  → The operator and the connected mathematica expressions can be put together as mathematical expressions.
  (e.g.) “x=y+z”
  → variables (mathematical expression) “x”, “y”, and “z” are connected by apparent operator “=” and “+”
  → “x=y+z” can be put together as a mathematical expression
- Assignment of a value to a variable
  → the operator, value, and variable can be put together as a mathematical expression
  (e.g.) “k_max = 50”
  → The value “50” is assigned to the variable “k_max”
  → “k_max = 50” can be put together as a mathematical expression
- One or more (all) slots of operators are Numerical values
  → the operators and any slots of operators are put together as mathematical expression
  (e.g.) “43.8% = (14.6-8.2) / 14.6%”
  → Each side (slot) of operator “=” (and “/”) is a numeric value (with unit)
  → “43.8% = (14.6-8.2) / 14.6%” can be put together as a mathematical expression
- Usage of introduced operators does not satisfy the requirement of mathematical expressions
  → The operators and slots of operators cannot be put together as a mathematical expression
  (e.g.) “P=Precision and R=Recall”
  → “=” is utilized to define the abbreviations of the right hand of the operator like “:”
  → Each of “P=Precision” / “R=Recall” cannot be taken as a mathematical expressions
  (e.g.) “evaluation on randomly polysemous data + significance”
  → “+” is utilized for expressing the pattern like “46% + 2σ”
  → The sequence cannot be taken as a mathematical expression
- Relations expressed using “=” are utilized for reporting observed values such as accuracies
  → The operators and those slots can be put together as mathematical expressions
  (e.g.) “P=95%”
  → The relation expressed by “=” reports that P(precision) was 95%
  → “P=95%” can be taken as a mathematical expression

- Other mentions (such as comments in brackets) are inserted into a mathematical expression with operators
  → The whole span of the mathematical expression including the mentions can be taken as a mathematical expression
  (e.g.) “EW2(Synset number) = CW2”
  → Bracketed comment “(Synset number)” is inserted into mathematical expression “EW2 = CW2”
  → The whole span of “EW2(Synset number) = CW2” can be taken as a mathematical expression
- An operator is introduced as an adverbial role of natural language text (typically by bringing only a right-hand slot of the operator), such as “equals ...”, “more than ...”, “less than ...“, etc.
  (e.g.) “length was ≦3” / “with frequency ≧ 2”
  → Each expression means “the length was shorter than 3” / “with frequency larger than 2”
  → Each of “≦3” / “≧ 2” is taken as a mathematicla expression

(2) Enumeration of Mathematical Expressions

- Enumeration of sequantial mathematical expressions are partially skipped
  → The whole span can be taken as a mathematical expression
  (e.g.) “X_1, …, X_k”
  → Enumeration of sequence {X_n} (n=i … k) is partially skipped
  → The whole span of the expression can be taken as a mathematical expresson
- A set of (relatively various) mathematical expressions are apparently grouped with ellipsis notation, etc.
  → The whole span of the set can be taken as a mathematical expression
  (e.g.) “Labels a , . . . , f , x”
  → A set of various mathematical expressions are grouped using “ . . . ”
  → The whole span of the set can be taken as a mathematical expression
- Individuality of each expression can be interpreted in the mention of enumeration (e.g. using expressions like “respectively”, “each”, etc. using “and” for connecting the end of the enumerated expressions)
  → Each of the expressions is taken as a individual mathematical expression
  (e.g.) “we obtain four SOGs : [G 0 , { x }] , [G 0 , G 1 , { x }] , [G 0 , G 2 , { x }] , and [G 0 , G 1 , G 2 , { x }]）”
  → Ellipsis notation is not utilized for enumerating fixed number (four) of expressions, and “and” is utilized for connecting the end of the expressions
  → Each of “[G 0 , { x }]”, “[G 0 , G 1 , { x }]”, “[G 0 , G 2 , { x }]”, and “[G 0 , G 1 , G 2 , { x }]” is taken as an individual mathematical expression
- Some incoherence is observed among several sets of enumerations which are in almost the same styles

→ Separating mathematical expressions to keep out as much natural language expressions and keeping as coherent interpretations among the enumerations as possible

(e.g.) “the left tags (t −2 , t −1 and t 0) , right tags (t 0 , t + 1 , t + 2) , or centered tags (t −1 , t 0 , t + 1) respectively”

→ “and” is inserted into one of the three sets of enumeration (= natural language expression)

→ Each of "t - 2", "t - 1", "t 0", "t + 1" and “t + 2” is individually taken as a mathematical expression for all of these three sets of enumeration

(3) Pairing / Grouping of Mathematical Expressions

- One or more mathematical expressions are paired or grouped
  → The whole span of the pair/group including brackets utilized for pairing/grouping is taken as a mathematical expression
  (e.g.) “(i, j)”
  → “i” and “j” is paired using “(” and ”)”
  → “(i, j)” is taken as a mathematical expression
- Nested paring/grouping
  → The whole span can be taken as a mathematical expressions
  (e.g.) “[{ a , b , c , d , e , f , x } , { a , b , x } , { x }]”
  → Three groups “{ a , b , c , d , e , f , x }”, “{ a , b , x }” and “{ x }” are further grouped
  → The whole span “[{ a , b , c , d , e , f , x } , { a , b , x } , { x }]” can be taken as a mathematical expression
- A group contains only one mathematical expression
  → The whole span including grouping brackets can be taken as mathematical expression
  (e.g.) “{ x }”
  → Only one mathematical expresson “x” is grouped
  → “{ x }” can be taken as a mathematical expression
- Pairing/Grouping brackets are not given to a paired/grouped expressions
  → If a sequence of expressions can be recognized as a pair/group, the sequence can be taken as a mathematical expression
  (e.g.) “a subgroup of participants P x , P y”
  → “P x , P y” can be taken as a group, and therefore as a mathematical expression
- No mathematical expression belongs to a pair/group
  → The pair/group cannot be taken as a mathematical expression
  (e.g.) “{ bank 1 , bank 2 , bank 3 }”
  → Each of “bank 1”, “bank 2” and “bank 3” is not a mathematical expression
  → The group “{ bank 1 , bank 2 , bank 3 }” cannot be taken as mathematical expression

(4) Mathematical Functions

- The mention of a mathematical function explicitly brings a notation of arguments
  → The mention can be taken as a mathematical expression
  (e.g.) “idf(t_i)”
  → The mention of mathematical function “idf” explicitly brings arguments “(t_i)”
  → The whole mention “idf(t_i)” can be taken as a mathematical expression
- The mention of a mathematical function skips the arguments of the function and thus (can be judged to) represents some concept which the function captures.
  → The mention cannot be taken as mathematical expression
  (e.g.) Function ”idf(t_i)” is defined somewhere, while “idf” in the target sequence does not refer to the function but the concept of idf
  → The target sequence “idf” cannot be taken as a mathematical exppression
- Some or all of the argument variables in the mention of a mathematical function are substituted with other variables / actual values
  → The mention can be taken as a mathematical expression
  (e.g.) “BLEW_{w4}”
  → Variable “x” in “BLEW_{x}” is substituted with value “w4”
  → “BLEW_{w4}” can be taken as a mathematical expression

(5) Interval Expressions
- A span can be interpreted as interval expression (regardless of explicit/inexplicit)
  → (Regardless the boundary values are mathematical expressions or not), the span can be taken as a mathematical expression
  (e.g.) “the interval [0 , 1]”
  → The span represents the interval {x: $0 \leqq x \leqq 1$}
  → “[0 , 1]” can be taken as a mathematical expression

(6) Mathematical Operations/Relations expressed by Unique Notations/Operators
- Special (own) notation in expressing rules or definitions
  → The span utilizing the introduced notation is regarded as expressing relations of mathematical expressions
  → The span can be taken as a mathematical expression
  (e.g.) "a:b→c:?"
  → A unique notation (*:*→*:*) is introduced for expressing the relation between mathematical expressions
  → The whole span “a:b→c:?” can be taken as a mathematical expression

C. Concatenation of Mathematically Structured Expressions According to their Dependency/Complementarity

(1) Bracketed Supplementary Statement for Mathematical Expressions

- Mathematical interpretation is enabled by concatenation of inside/outside of a bracket
  → The whole span can be taken as a mathematical expression
  (Note that if inside of a bracket is much beyond a mathematical expression, such as natural language expressions, the span cannot be taken as a mathematical expression)
  (e.g.) “t_i (∈T)”
  → The sequence “t_i ∈T” obtained by concatenating the inside/outside of the brackets, can be mathematically interpreted
  → “t_i (∈T)” can be taken as a mathematical expression
- A bracketed supplementary comment is delivered to several mathematical expressions which are separated by some natural language expressions
  → As long as the consistency of mathematical expressions is kept, each maximal span of continuous mathematical expressions can be taken as a mathematical expression
  (e.g.) “α = ... and β = ... (k < min(m , n))”
  → Each of “α = …”, ”β= …” and “k < min(m , n)” is individually taken as mathematical expression
- Bracketed supplementary statement is not structurally and semantically essential for interpreting the target mathematical expression.
  → Each of the inside/outside of the brackets are separately taken as a mathematical expression
  (e.g.) "EW2 = CW2 (ii=0)"
  → “ii=0” represents another perspective of the condition “EW2 = CW2”
  → Each of “EW2 = CW2” and “ii=0” is separately taken as a mathematical expression

(2) Concatenation with Natural Language Text

- Mathematical expressions and natural language text are sequentially concatenated
  → The whole span cannot be taken as a mathematical expression
  (e.g.) “(i , j)-th”
  → Mathematical expression “(i , j)” and natural language text“-th” are sequentially concatenated
  → “(i , j)-th” cannot be taken as a mathematical expression, i.e., only the span of “(i , j)” is taken as a mathematical expression

(3) Complementary for Enumerated Various Types of Mathematical Expressions

- Some complementary relations can be observed among enumerated mathematical expressions
  → The whole span of the enumeration can be taken as a mathematical expression
  (e.g.) “(x i , y i) , i = 1 , . . . , l”

→ Pair “(x i , y i)” and Values “i = 1, . . . , I” can be considered to have some complementary relation
→ The whole span of “(x i , y i) , i = 1 , . . . , I” can be taken as a mathematical expression

(Note) On mathematical Expressions Transcribed in Natural Language Expressions

- It is observed that some parts of mathematical expressions seem to be transcribed in natural language text
  → Each span of mathematical expressions are individually taken as a mathematical expression
  (e.g.) “Each sentence s i for i = 1 . . . n”
  → Mathematical expression “s i (i = 1 . . . n)” seems to be transcribed using natural language text
  → Each span of “s i” and “i = 1 . . . n” is individually taken as a mathematical expression

D. Relaxation of Strictness of the Format of Mathematical Expressions

- The span seems to be incomplete as a mathematical expression, while the meaning of the span will be corrupted unless the span is interpreted as a mathematical expression
  → The span can be taken as a mathematical expression
  (e.g.) “δ = 0 . 5 ~ 0 . 95”
  → “~” is not formal, while dividing the span into “δ = 0 . 5” and “0 . 95” will break up the meaning of the span
  → The whole span of “δ = 0 . 5 ~ 0 . 95” can be taken as a mathematical expression
  (e.g.) “features with support ≦ cutoff”
  → “support” and “cutoff” are not mathematical expression, while mathematical operator “≦” does not make sense without these expressions.
  → The span “support ≦ cutoff” can be taken as a mathematical expression
  (e.g.) “K=2 to 60” / “K=25, 35, and 60”
  → Natural language words “to” / “and” are introduced, whle the sequences does not make sense without mathematical interpretation of the words
  → Each of the whole span “K=2 to 60” / “K=25, 35, and 60” is individually taken as a mathematical expression

E. (Supplement) Treatment of Mathematical Expressions Crossing Paragraphs

- A mathematical expression seems to be divided into more than one paragraphs

→ Each fragment of the expression is taken as a mathematical expression
(e.g.) "D = 1 + the num-(boundary of paragraph)ber of NP-markables between anaphor and potential antecedent”
→ According to Case D, the whole span can be taken as a mathematical expression, while a boundary of paragraph divided the span into two fragments
→ Each fragment of the span is taken as a mathematical expression