Towards Grounding of Formulae

Takuto Asakura[1], André Greiner-Petter[2], Akiko Aizawa[3], Yusuke Miyao[1]

[1]The University of Tokyo [2]University of Wuppertal [3]National Institute of Informatics

Introduction

Targets: STEM Documents

Characteristics
- Not fully utilized
- Semi-structured
- Includes math formulae

Example
- Papers
- Textbooks
- Manuals

Long-term Goal: Natural → Formal Conversion

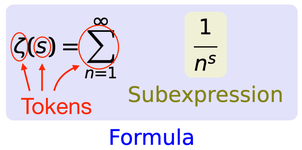
STEM Documents (Natural Language + Formulae)	
Paper, Textbook, etc.	Calculating the mean of n^2 for $n = 1, 2, \ldots, 10$

↓ Conversion

Computational Form (Formal Language)	
Executable code, first-order logic, etc.	`mean([n^2 for n in range(1, 11)])`

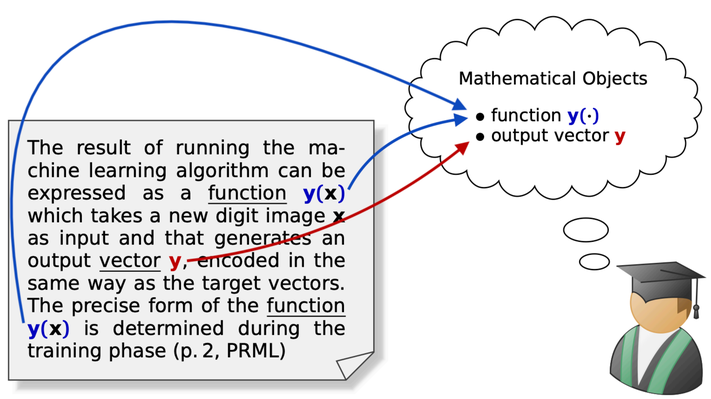
Short-term Goal: Token-level Analysis

- The first step for the conversion
- Still a developing area



Grounding of Formulae

1. Finding math words (groups of tokens) which refer to mathematical concepts
2. Associating a corresponding mathematical concept to each math word



Difficultly of the Grounding

- Various ambiguities in formulae
- Necessity of domain knowledge

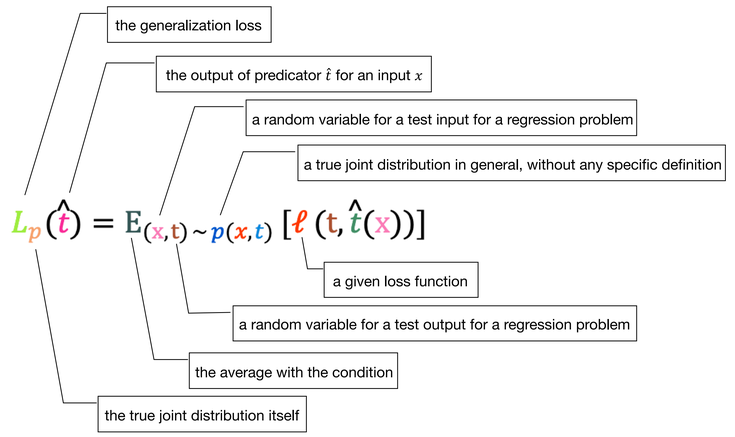
Usage of character y in the first chapter of PRML (except exercises)

Text fragment from PRML Chap. 1	Meaning of **y**
… can be expressed as a function **y**(**x**) …	Function representing an algorithm
… an output vector **y**, encoded in …	Output vector of function **y**(**x**)
… vectors of random variables **x** and **y** …	Vector of random variables
Suppose we have a joint distribution $p(\mathbf{x}, \mathbf{y})$ …	Part of pairs of values

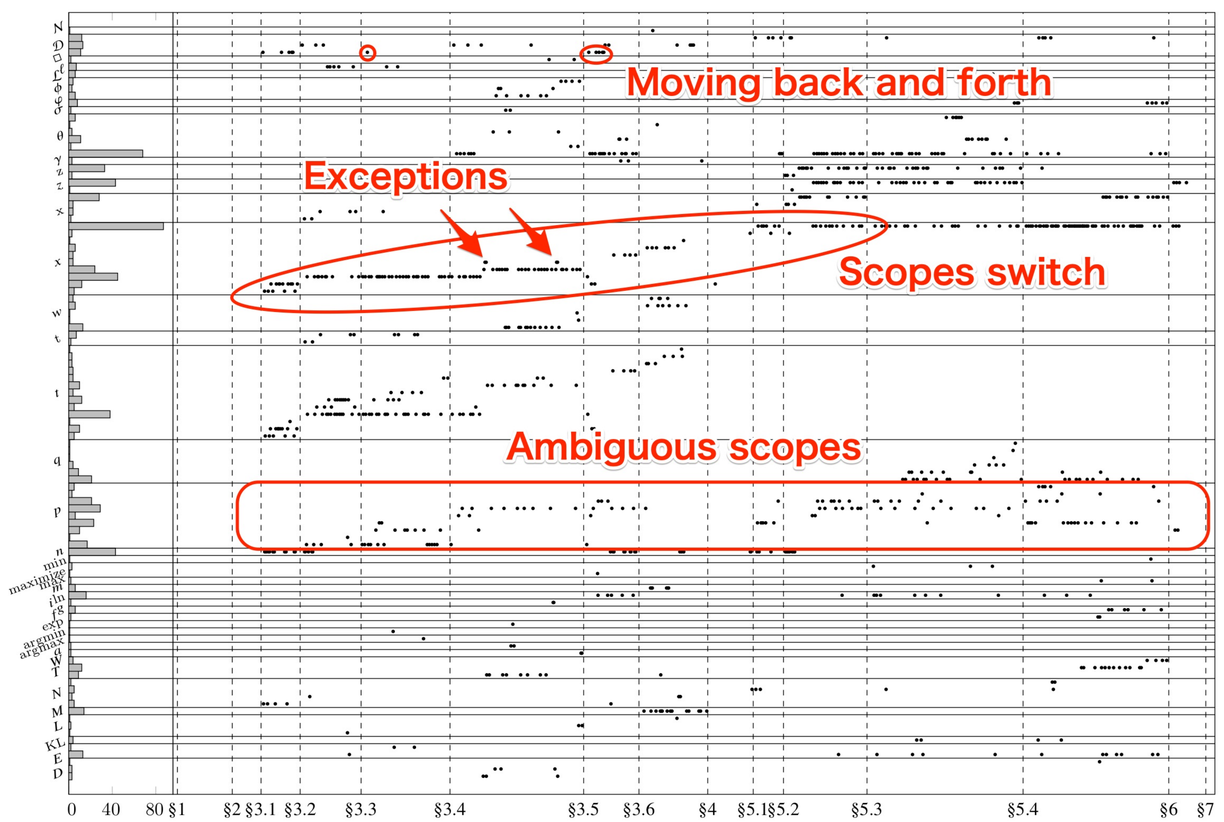
Dataset Creation (Manual Annotation)

Annotating all identifiers in a whole paper
A Very Brief Introduction to Machine Learning With Applications to Communication Systems [Simeone, 2018]

Annotation Example



Math Word Occurrences in the Target Paper



x axis
position in the article

y axis
corresponding mathematical concepts