

数学概念への数式グラウンディングのためのデータセット

朝倉卓人^{1, 4}, André Greiner-Petter², 相澤彰子^{1, 3, 4}, 宮尾祐介⁴

総合研究大学院大学¹, ヴッパータール大学², 国立情報学研究所³, 東京大学⁴

イントロダクション

研究対象：科学技術文書（STEM文書）

例

科学技術文書の特徴

- 理工系論文
- 大学の教科書
- マニュアル文書 等
- 人類叡智の真髄
- 定形化されている
- 数式混じりの文章

長期目標：科学技術文書の形式表現への変換

科学技術文書（自然言語＋数式）

論文, 教科書, マニュアル $n = 1, 2, \dots, 10$ について n^2 の平均を計算する

↓ 変換

計算可能な表現（形式表現）

プログラミング言語, 一階述語論理 $\text{mean}([n^2 \text{ for } n \text{ in range}(1, 11)])$

短期目標：トークンレベル解析

まずトークンレベル解析に注力

- 変換の最初のステップ
- 先行研究が少ない

トークンの例

$n, s, \zeta, \infty, \sum, =, \times$, etc.

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

部分式
数式

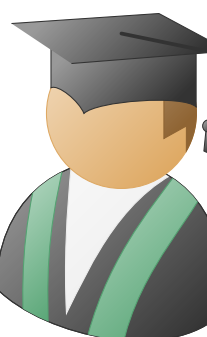
提案タスク：数式グラウンディング

次の2ステップを数式グラウンディングと呼ぶ：

1. 数学概念を指すトークンのまとまり（単語）を認識
2. 各単語に、その語の指し示す数学概念を紐付ける

機械学習のアルゴリズムによって得られるのは関数 $y(x)$ である。この関数に、新たに数字の画像 x を入力すると、目標ベクトルと符号化の仕方が等しい出力ベクトル y が出力される。関数 $y(x)$ の詳細な形は訓練データに基づいて求められる。(PRML, p. 2)

数学概念
・関数 $y(\cdot)$
・出力ベクトル y



読者

数式グラウンディングの困難性

- 数式中の記号や構文の曖昧性 例 $f(a + b)$
- 常識やドメイン知識の必要性

PRML第1章におけるトークン y の多義性

本文のテキスト断片	y の意味
...得られるのは関数 $y(x)$ である...	画像を入力とする関数
...出力ベクトル y が出力される...	関数 $y(x)$ の出力ベクトル
2つの確率変数ベクトル x と y に...	確率変数ベクトル
...同時分布 $p(x, y)$ を考えよう.	x に対応する値

データセットの作成（アノテーション）

1編の論文の中のすべての識別子をアノテーション

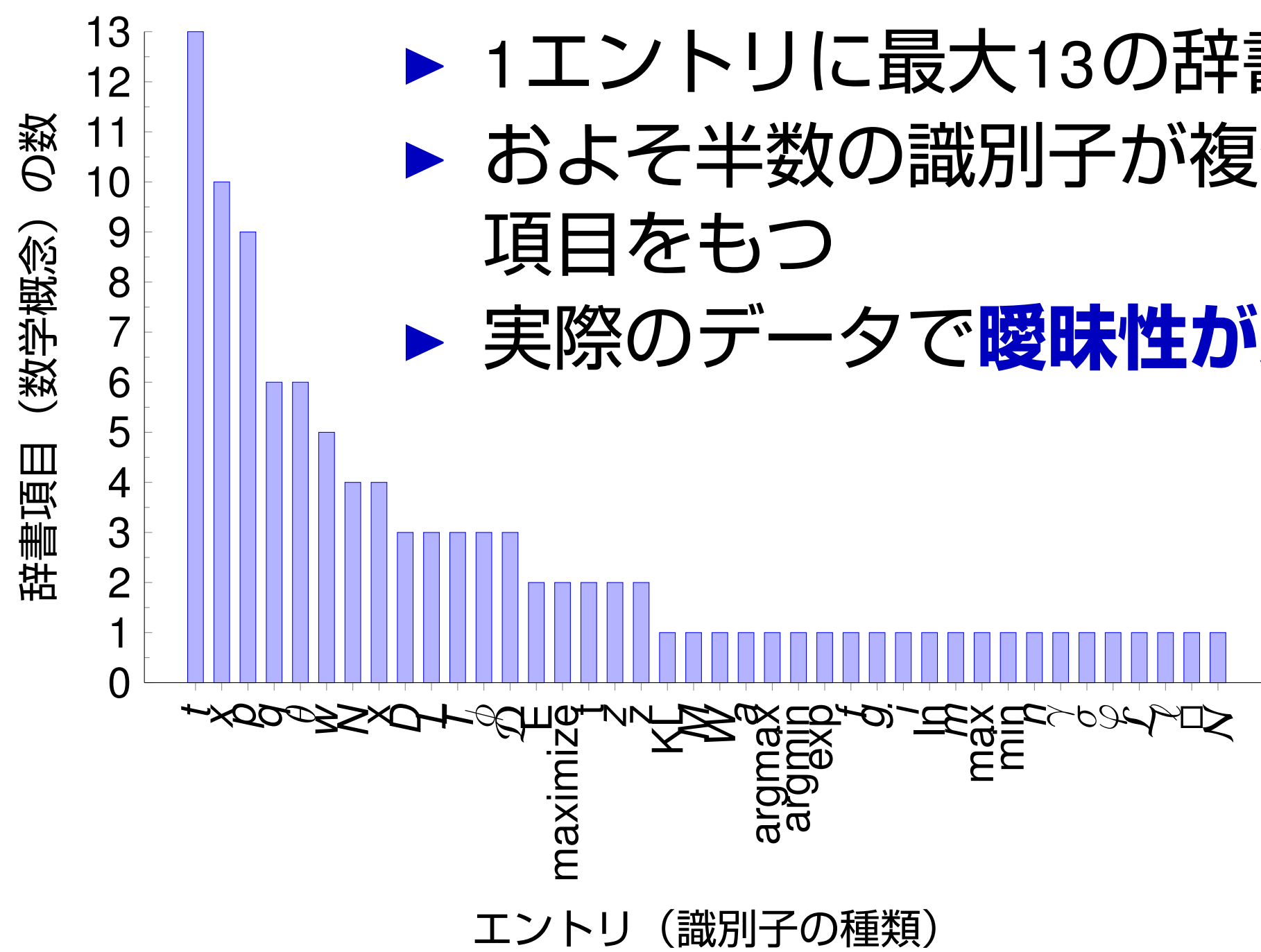
- トークンの1種. プレゼンテーション MathML の $\langle \text{mi} \rangle$
- 文字または短い文字列 例 $x, y, \theta, \sin$, etc.

アノテーション対象の論文

A Very Brief Introduction to Machine Learning With Applications to Communication Systems [Simeone, 2018]

識別子ごとのクラスタ数

- 1 エントリに最大13の辞書項目
- およそ半数の識別子が複数の辞書項目をもつ
- 実際のデータで曖昧性が見られる



対象論文におけるトークンと数学概念の分布

