

提案タスク：数式グラウンディング [Asakura+ 2020]

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



グラウンディング情報源

文書内外の数式グラウンディングの根拠となるもの
文書内 周辺テキスト, 数式 例 同格名詞, $\stackrel{\text{def}}{=}$
文書外 常識, ドメイン知識 例 Wikidata

The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x , the Gaussian distribution can be written in the form

$$N(x | \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

where μ is the mean and σ^2 is the variance. (p. 78, PRML)

Common sense
 ・ π : Archimedes' constant
 ・ exp: real exponential function



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

数式グラウンディング=説明アライメント+共参照解析

説明アライメント

- ▶ 各トークンに説明 (description) を付与するタスク
- ▶ いくつか先行研究あり [Aizawa+ 2013, Alexeeva+ 2020, etc.]
 →ほとんどが**トークンの意味は文書内で一定**と仮定

$$L_p(\hat{t}) = E_{(x,t) \sim p(x,t)} [\ell(t, \hat{t}(x))]$$

Labels: the generalization loss, the output of predictor \hat{t} for an input x , a random variable for a test input for a regression problem, a true joint distribution in general, without any specific definition, a given loss function, a random variable for a test output for a regression problem, the average with the condition, the true joint distribution itself

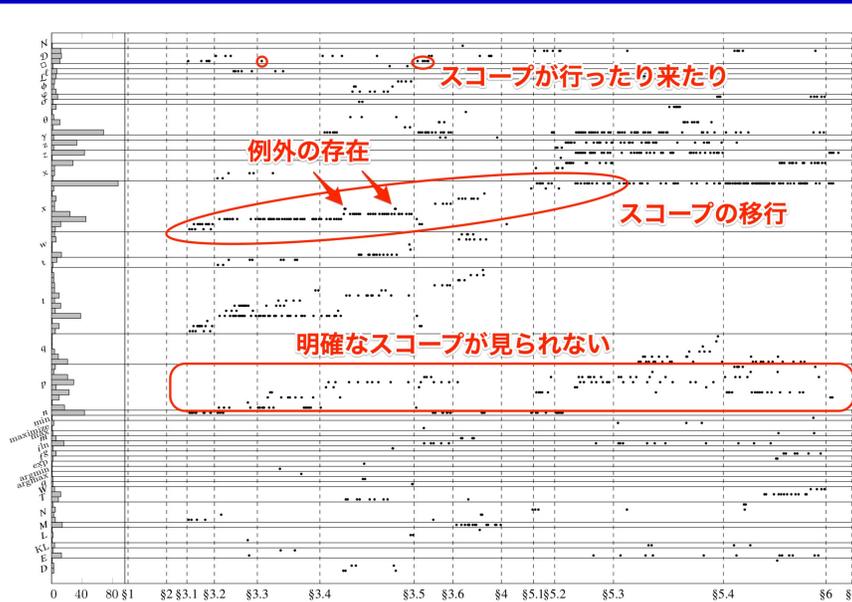
共参照解析

桃太郎は桃から生まれた。
 彼は鬼退治に出かけた。

共参照



数式識別子の出現と数学概念 [Asakura+ 2020]



アノテーションツール MioGatto [Asakura+ 2021]

Math Identifier-Oriented Grounding Annotation Tool

- ▶ 数式グラウンディングデータ構築のための独自ツール
 → Web ベース GUI (Python + TypeScript 実装)
- ▶ アノテーションできるもの
 - ▶ 数学概念 (説明, 付加情報, 共参照情報)
 - ▶ グラウンディング情報源
- ▶ 4月に**オープンソース**として公開! (MIT ライセンス)

<https://github.com/wtsnjp/MioGatto>

アノテーション体制と進捗

- ▶ アノテータ8名 (まだまだ募集中!)
 → NLP 4名・論理学2名・代数学1名・物理学1名
- ▶ 重複のあるグラウンディング情報源を集計

グラウンディング情報源の重複 (対アノテータ1)

	アノテータ1	アノテータ2	アノテータ3
すべて	232	249	257
概念一致	—	159 (63.9%)	183 (71.2%)
概念不一致	—	41 (16.5%)	53 (20.6%)

自動化の方針

1. 文書内グラウンディング情報源の**特定・抽出**
 → パターンマッチ+品詞分解 (同格名詞) 利用
2. 文書内情報源のクラスタリングによる「辞書」生成
 → Short Text Clustering 手法 [Jiaming+, 2017] の適用
3. 文書中の各数式トークンと「辞書」項目の**関連付け**
 → パターンマッチ+品詞分解+分類モデル

