自然言語文中に含まれる数式記号の意味推定の試み

An attempt to infer the meaning of mathematical symbols contained in natural language text

井村　翔[*1]，宮崎　佳典[*2]，田中　省作[*3]
Kakeru IMURA[*1], Yoshinori MIYAZAKI[*2], Shosaku TANAKA[*3]
[*1] 静岡大学 情報学部
[*1]Faculty of Informatics, Shizuoka University
[*2] 静岡大学学術院 情報学領域
[*2]College of Informatics, Shizuoka University
[*3] 立命館大学 文学部
[*3]College of Letters, Ritsumeikan University
Email: imura.kakeru.20@shizuoka.ac.jp

あらまし:数式に含まれる記号や文字は周辺の記号・文字や自然言語文の情報によって意味を判別するが，本研究では自然言語文中に存在する数式内記号の意味推定を試みる．具体的には，形態素解析を行った後，数式を含む自然言語文章用に調整した TF-IDF や N-gram の計算処理を行い，推定対象ならびに比較用データ文章群間の類似度を計測する．これにより数式交じり文章の特徴を考慮した文章比較を実現する．
キーワード：数式，自然言語，形態素解析，TF-IDF，N-gram

1. はじめに

数式には「|」や「{ }」といった記号や「sin」のような文字列が含まれている．人間であれば数式内におけるこれらの意味を理解することは難しくないが，コンピュータにおいては数式から記号を機械的に抜き出した際にその記号の意味を理解できなくなる．本研究は当該課題の解決に向けたものであり，数式内記号の意味がコンピュータに正しく推定されることで，例えば数式検索システム(1)において数式の意味というアプローチから精緻な数式検索が行えることを目指す．

関連研究として，MathML を用いて数式記号の意味推定を行う櫻井ら(2)の研究や，英次略語の類義語の意味推定を行う後藤ら(3)の研究が挙げられる．これらの関連研究と本研究との相違点として，数式の周りの自然言語情報を用いて数式記号を対象とした意味推定を行っている点を挙げる．

2. 意味推定の流れ

2.1 意味推定の概要

意味推定には，入力文と蓄積データ文の 2 種類の文章を扱う．前者は意味を判定したい数式記号を含む文章である．後者は数式を含む自然言語の文章群であり，数式記号の意味別に管理されている．入力文と各意味の蓄積データ文間の類似度を文章比較ツールによって算出し，一番値が大きい蓄積データ文の意味を入力文の示す意味とする．

2.2 文章比較の提案手法

本研究における文章比較の手法を説明する．

入力文と蓄積データ文を形態素解析ツール MeCab(4)によって形態素で分解する．その後，以降に述べるいずれかの手法を用いて 2 文章間の類似度を算出する．

形態素解析のみによる手法

両文章における形態素の出現回数を測定し，それを元に cos 類似度を算出する．

TF-IDF を用いた手法

入力文に含まれる形態素ごとに TF-IDF 値を測定し，文章間の cos 類似度を算出する．ここで TF-IDF とは，文章群中における単語の出現頻度を表す TF 値と文章群中においてどれだけ少ない文章に単語が使用されているかを表す IDF 値の積のことを指す．

3-gram を用いた手法

両文章それぞれの辞書を 3-gram で作成し，その辞書を元に出現回数を測定後 cos 類似度を算出する．ここで 3-gram とは，連続する単語や文字などを 3 個ずつにまとめたものを指す．本研究では形態素を構成要素とする 3-gram の手法を用いる．

3. MeCab・TF-IDF・3-gram への数式適用

本研究では自然言語を対象とする MeCab や TF-ID・3-gram の手法に対して，数式を扱えるように適用を施した．以下でその内容を述べていく．

3.1 MeCab へのユーザ辞書導入

本研究に用いる MeCab にはユーザ辞書を導入しており，記号や数学用語，MathML タグといった本来正しく分解されない形態素への適用処理を施している．例えば，「|a|>1」中の「|>」が 1 つの形態素とされてしまうのを「|」「>」と 1 文字ずつ分解する，「絶対値」が「絶対」と「値」で分けられてしまうところを「絶対値」と数学的に意味のある形に分解する，「<math>」を一つの形態素として認識させるといった処理を可能にしている．本研究にてユーザ辞書に

登録した形態素の種類及び各項目の登録数を表 1 に示す.

表 1　MeCab にユーザ辞書登録した形態素

種類	登録数
記号	305
数学用語	889
MathML タグ	92

3.2 TF-IDF 及び 3-gram への適用

3.1 節にて述べた MeCab へのユーザ辞書導入は, TF-IDF 値の計測対象や 3-gram の構成要素も変化したことを指す．それにより，先に述べた形態素単位での TF-IDF や 3-gram の処理が可能となった.

4.　実験 1

4.1 実験 1 の目的

数式記号の意味推定を行うにあたり，精度の高い推定手法を定めるために実験 1 を行った.

4.2 実験 1 の手法

本実験のために数式を含む日本語文を 220 文用意した. このうち 80 文は「|」, 40 文は「+」, 40 文は「-」, 残り 60 文は「m」の記号または文字に該当する．全ての意味において 20 文ずつ収集した.

これらを記号別に蓄積データ文とし，そこから 1 文を抽出して入力文とする．この状態で入力文と蓄積データ文の比較を行い, 2.2 節に述べた文章比較のいずれかの手法によって類似度を算出する. その後, 類似度を降順に順位付けし入力文の本来の意味における類似度が何番目の順位であるかを数える．入力文に対して正しい意味の蓄積データ文が 1 番高い類似度となった文章数を求め，それを記号全体の文章数で割った値を，当該手法の精度とする.

4.3 実験 1 の測定結果

各推定対象における実験 1 の精度測定結果をまとめたものが以下の表 2 である.

表 2　実験 1 の精度測定結果(単位：%)

推定対象	形態素解析のみ	TF-IDF	3-gram
|	80.00	81.25	82.50
+	85.00	87.50	92.50
-	67.50	67.50	67.50
m	86.67	93.33	86.67

形態素解析のみの手法よりも TF-IDF または 3-gram を用いる手法の方が良い精度になる，または精度が変化しないという結果が得られた．また 2 種類の推定対象において 3-gram の精度が TF-IDF に比べて優れており,「-」の結果のうち誤った判定をしたケースの類似度においても 3-gram の方が TF-IDF よりも優れていた点から，TF-IDF よりも 3-gram の手法の方が優位であると結論付けた.

5.　実験 2

5.1 実験 2 の目的

ベクトルや指数など，プレーンテキストでは正しく表せない数式を正確な形で扱える形式を用いることを考え，実験 2 を行った.

5.2 実験 2 の手法

実験 1 にて扱った 220 文の文章内の数式部分を MathML によって表した．その上で，文章比較の精度はどう変化するのかを測定する．精度は実験 1 と同様の定義で算出する.

本実験では TF-IDF を用いる手法と 3-gram を用いる手法の 2 種類の手法の精度を，MathML を含んだ場合と除いた場合の 2 パターンにおいて算出する.

5.3 実験 2 の測定結果

各推定対象における実験 2 の精度測定結果をまとめたものが表 3 である.

表 3　実験 2 の精度測定結果(単位：%)

推定対象	MathML あり		MathML なし	
	TF-IDF	3-gram	TF-IDF	3-gram
|	67.50	72.50	82.50	100.00
+	75.00	90.00	87.50	100.00
-	75.00	75.00	67.50	100.00
m	81.67	86.67	90.00	76.67

MathML のタグを増やしたことで精度が落ちる傾向が見られた．その裏付けとして，MathML を除いた手法の精度が実験 1 に示した精度と同水準に戻ったことが挙げられる．また「|」「+」「-」においては 3-gram が TF-IDF より優れているという傾向が MathML の有無によって変わらない一方，「m」においては MathML の情報を無くすことで精度の大小に逆転が見られた．このことから，推定対象の文字種によって文章比較に用いる手法を分けることで全体的な精度向上が見込めると考えられる.

6.　まとめと今後の展望

本研究では，MeCab への独自のユーザ辞書導入により本来正しく形態素分解されなかったものへの対応を試みた．それにより，TF-IDF や 3-gram といった自然言語処理を可能とした．現在は MathML のタグごとに対して，文章比較に用いるか否かを分けることで推定精度の向上を目指している．数式の周りの情報と MathML の情報の両者を活かした形で推定手法を定めていく.

参考文献

(1) 渡部孝幸, 宮崎佳典: “正規表現を用いた数式検索手法の提案”, 情報処理学会論文誌, 第 56 巻, 第 5 号, pp.1417-1427　(2015)
(2) 櫻井翼, 宮崎佳典: “意味情報を付与した数式検索システムの検索機能拡張に向けた提案” , 第 86 回全国大会講演論文集, 第 2024 巻, 第 1 号, pp.967-968 (2024)
(3) 後藤和人, 土屋誠司, 渡部広一: “語彙の概念化と Wikipedia を用いた英字略語の意味推定方法”, 自然言語処理, 24 巻, 3 号, pp.351-369　(2017)
(4) MeCab: http://taku910.github.io/mecab/