卒業論文

計算問題の特徴分布に基づく類題選出による自己学習支援

Self Learning Support by Automatic Selection of Calculation Exercises based on Feature Distribution of Exercises

成蹊大学理工学部情報科学科

S152114 宮地 雄也

要旨

ポイント

序論と結論の内容をもとに研究の内容をまとめる.

- 問いは何か??
- 主張は何か??
- 結果はどうだったのか？
- 得られた成果の意義は？

目次

第 1 章

序論

ポイント

問題提起を行う．解く価値があり，簡単には解けず，誰も解いていない問題を扱っていることがわかるようにする．

- どういう問題に取り組んだのか？
- その問題を解くことがなぜ重要なのか？ 社会的意義（有用性）・学術的意義（問題の面白さ）
- その問題はどこが難しいのか？ なぜこれまで解かれていなかったのか？ これまではどうしていたのか？
- その問題をどのようなアプローチで解こうとしたのか？ なぜそうしたのか？

昨今，小・中学生の理系離れが問題視されている．平成 30 年度全国学力・学習状況調査（全国学力テスト）の結果では平均正答率は小学校では算数 B が 51.7%，中学校数学では 47.6% とどちらも最も低く，ついで国語，理科の順で正答率が低い．小中どちらとも理系教科の習熟度が低いことを示している．この要因の一つに，数学は一つの計算方法が様々な分野に横断していくことが一度，苦手を生んでしまったらそこからの分野の理解度が下がり，次の分野での応用がきかないために連鎖的に苦手が蓄積することが原因ではないかと考えた．各単元のちょっとした積み残しが，後々，尾を引いていることが全国学力テストの結果から見て取れる．この状況を打破するには子供一人一人の苦手と向き合い，苦手と感じる前に理解していくしかない．しかしながら，生徒と向き合うべき教師の労働時間は過酷を極めており，ベネッセ教育総合研究所の調査ではし小中高の教員の領導時間は増加の一途を辿っている．[1] 下の表は [1] での調査の結果の抜粋である．(1.1)

	調査年	25 歳以上	26〜30 歳	31〜40 歳	41〜50 歳	51〜60 歳
出勤時間	2010	7:44	7:43	7:44	7:42	7:42
	2016	7:44	7:43	7:44	7:42	7:42
退勤時間	2010	19:30	19:40	19:10	18:57	18:31
	2016	20:00	19:54	19:26	19:05	18:46
学校にいる時間	2010	11 時間 46 分	11 時間 57 分	11 時間 26 分	11 時間 15 分	10 時間 49 分
	2016	12 時間 26 分	12 時間 18 分	11 時間 46 分	11 時間 26 分	11 時間 06 分

表 1.1　出勤時刻・退勤時刻・学校にいる時間（平均時間、経年比較（教員年齢別〔公立全体〕））

表 1.1 によると，教員の労働時間は 2010 年に比べて 2016 年の方が各年次とも増加しており，教員のやることが増えている一方で，主であるはずの教材研究や教務準備に時間が避けていないことを示している．この状況では先生が生徒一人一人二時間をさき，指導することは難しい現状がつついている．

この打開策として，IT 技術駆使した個人別最適化学習に注目が集まっている，しかし，教育の情報は，生徒の情報と結びついている個人情報なためオープン化できず，現在でているサービスでは各サービス利用者の利用状況からデータを取得し，その運用に利用しているため一部の王手企業が情報を独占している．そこで個人の統計データではなく，解く数式の方に着目し，計算式自体の特徴を抽出し，間違えた問題と同様の特徴を持つ問題が復習する類題として最適なのではないかという仮定のもと，本論文では数式の特徴を掴むために自然言語処理の分野で使用される分散表現を適用し，さらに再起ニューラルネットワークを用いて数式ベクトルを作り出すことを目標とし，そのベクトルを用いて実際に復習問題生成を行った．

第 2 章

背景知識

第 3 章

分散表現

様々な手法を紹介する CBow，skipGram，GloVe など計算式を並べて最後にどこが長所で，どこが違うのかを表現

第 4 章

LSTM(Long short-term memory)

通常の RNN では叶わないところを明確に

第 5 章

Attention

これからね

第 6 章

提案手法（章題は変える）

ポイント

自分の提案する解決方法を説明する．

- 章題は適切なものに変えること．章をわけてもよい．
- 必ず具体例を用いること．
- 最初に問題を解く上で最も難しい点とそれを解決するアイデアを示す．
- 詳細については，全体の流れを示した後，各ステップについて説明する．
- 検討時に行った予備評価の結果があれば示す．

6.1 システム全体の流れ

＜図をいれながら＞

6.2 計算式の特徴量抽出

6.2.1 概要

＜ idea ＞（内容充実させる）数式を分布化する際，そのベクトルの中に数式の特徴を入れ込んだベクトルを生成する手法が確立していない．そこで本論文では数式の各文字，記号を単語のようにみなし，onehot ベクトルを作成し，それを埋め込み層で特徴を踏まえた低次元ベクトルに変換したのち，系列変換モデルで読み込むことで低次元で数式の特徴を掴んだベクトルを生成できないかと考えた．

＜手法＞この考えを実現するために数式は我々が目にする $2x + 3 = 5$, $\frac{3x-1}{2} + 4 = \frac{2}{5}$ ではなく，テキスト化かつその特徴を強く受けた形に変換する必要がある．そこで本論文では数式をある一定のルールの中でテキスト化されている TEX 形式の数式を用いる．上記の計算式なら `2x+3=5,\frac{3x-1}{2}+4=\frac{2}{5}`とし，このテキストデータを用いて文字単位の埋め込んだベクトルを作成する．

実験を行った手法は以下の三種法で行い，それぞれ分布を python を用いて確認した．

- CBOW
- SkipGram
- ...

onehot ベクトルの置き方は

- $[0, 1, 2, ..., 9, +, -, =, x...]$ のように各数字，各記号に割り当てる方法
- 出てきた数字，数式の塊を onehot を置く方法 ($[0, 1, 2, ...3.6, 0.11, ... = .+, -]$)
- 3 桁までの数字，数式に現れる記号

6.2.2 文字分布の入手

予備実験として文字の分布を入れる CBOW，SkipGram，（できれば grove も）
できればベクトルの足し引きとかできれば word2vec の論文にもそうので結果を確かめたい．
＜おまけだからさらっと＞

6.3 判定機

今回提案するシステムではといたプリントを読み込みその結果を判別して間違った計算問題を見つけ，その類題を選出する．

6.3.1 概要

『文章で説明』
→ あと誤差も（データ取り直しています）

第 7 章

結果とその検討

ポイント

自分の提案する方法が序論で提起した問題を解決できているかを評価・分析する．

- 目的．何を確認するためのものか
- 方法．そのためにどういう実験を行ったか？ 実験環境・用いたデータとその選定理由・手順を示し，評価の適切性を論証すること．
- 結果．その結果はどうだったか？ 表やグラフを用いてまとめる．表は TeX，グラフは excel でなく python を用いて作成すること．
- 分析．その結果から何が言えるか？ 達成できた点・不足している点を理由と共に述べ，原因を考察する．

7.1 実験

計算式の特徴量抽出

- 目的．数式のベクトル表現は可能なのかどうか
- 方法．埋め込み層，ネットワーク構成を変えながら Encoder-Decoder で復元を試みるその Encoder の出力値を数式ベクトルとしてみなし，pca,T-sne で二次元ベクトルに圧縮し図示する
- 結果．これから
- 分析．これから

計算式の特徴量からの類題選出

- 目的．求めた計算式のベクトルから特徴を捉えた数式を選出できるか
- 方法．特徴量ベクトルから k 近法で選出，ある生徒の間違えた一問から類題を選出し，実際にその問題を間違えていたかを確認，また，逆にあっていた問題からも同様な手法で確認
- 結果．これから
- 分析．これから

7.2　評価

7.3　分析

7.4　検討

第 8 章

関連研究

ポイント

この研究に関連する他の研究を紹介し，この研究との違いを明確にする．

- 文献は「Mnih らは〜という手法を提案している [**one1**].」のように `cite` コマンドを用いて文献番号を示すこと．
- 2 ページ以上書く．

https://code.google.com/archive/p/word2vec/
https://techblog.asahi-net.co.jp/entry/2018/10/05/180310

遺伝的アルゴリズムについては，伊藤 [**伊藤**] に詳しく述べられている．団塊世代がインターネットの利用に抵抗感を持つことが，報告されている．[2]

第 9 章

結論と今後の課題

ポイント

序論で提起した問いとそれに対する答えをまとめる．

- 提案手法のアイデアおよび評価結果を振り返る．
- この研究で得られた知見をまとめる．
- 今後の課題について述べる．

第 10 章

形式上の注意

- 文字コードは UTF-8 に統一する．
- 論文ファイル名は `chishiro-thesis.tex`，文献ファイル名は `chishiro.bib` のように名前`-thesis.tex` とする．
- 句読点は全角のカンマ，ピリオドを用いる．
- 英数字はすべて半角を用いる．ギリシャ文字は TEX の定義を用いる．$\alpha, \beta, ...$
- カンマの前にはスペースを入れず，カンマの後はスペースをひとつ入れる．
- 数式は TEX の数式機能を用いる．例: x^2,

$$f(x) = x^2 + 2x + 1.$$

  .
- プログラムテキストはタイプライターフォントを用いる（例: `hello`）．
- 文章構成（章・節・小節・箇条書き）は TEX の機能を用いて指定する．自分で見出しなどを作らない．
- 題目には研究目的・方法・対象を特徴づける情報を入れる．
- 図のタイトルは図の下，表のタイトルは表の上に書く．
- 図表番号の参照は`\label` および`\ref` を用いる．自分で図表番号を指定しない．
- 表は TEX，グラフはすべて python で作成する．
- 図表番号のない図は用いない．
- 参照の?は必ず取り除く．
- 段落は意味の区切りでわける．意図しない字下げが入った場合`\noindent` を用いて修正する．
- 参考文献は 10 以上あげる．

謝辞

ポイント

本のあとがきに相当する部分．半ページ以上書く．卒業研究に協力者してくれた方々へのお礼を忘れずに述べる．

Bibliography

[1] ベネッセ教育総合研究所. “第 6 回学習指導基本調査 DATA BOOK（高校版）[2016 年]”. In: (2016).

[2] 三木光範. 団塊世代はなぜインターネットが苦手か. ブルーバックス B-1202. 講談社, 1998.