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

プログラミング言語研究室 S152114 宮地　雄也

1　はじめに

昨今，小・中学生の理系離れが問題視されている．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原因の一つと考える．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させていくしかない．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大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，さらに再帰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推定を行った．

2　解決手法

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．しかしながら自然言語の分野では単語を効率よく埋め込み層に変換する Word2Vec という手法が結果を残している [1]．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 ベクトルを用いて埋め込み層で特徴を踏まえた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．

本論文では日常，私たちが見ている式の形だとベクトル化できないので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．本論文の目的としては系列変換の過程で用いられる Encoder から Decoder に渡す最終出力 h を式ベクトルとしてみなす．図 1 にモデル模式図を掲載する

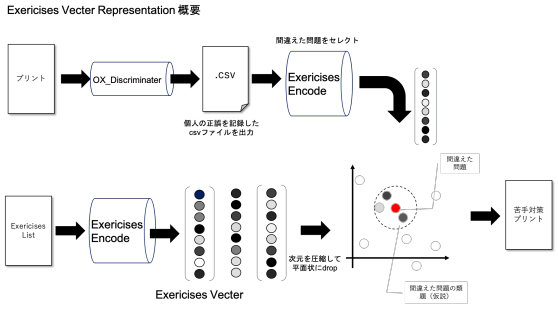

図 1　ExercisesEncoder は入力をベクトルに変換することを担う．

より高い精度の分布を得るためには h の精度を高めていくモデルが必要となる．実験するモデルは通常の LSTM で行う系列変換，双方向 RNN を用いた bi-directionl モデル，通常 Decoder が側で利用される SkipConnection の三種類を Encoder のモデルとした．デコーダ側は通常の LSTM で行う系列変換で行うこととした．Encorder 側のモデル構成図を以下に示す．図 2 は通常の Encoder，Decoder からなる系列変換モデルである．LSTM 層は縦に特徴を抽出するために層を増やすことは可能だが，層を深くするればするほど過学習を起こしやすい性質がある．よって，この点を改善するために図 3 に示す双方向 LSTM を実装した Bi-DirectionalEncoder モデルと層を深くすることで発生する勾配爆発，あるいは消失を防ぐために図 4 に示す SkipConnection モデルを実装した．

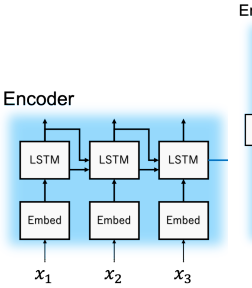

図 2　基本となる NormalEncoder

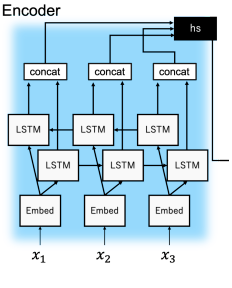

図 3　bi-directionl Encoder

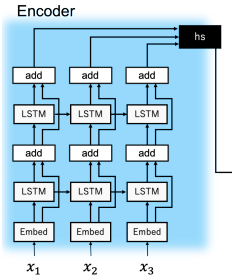

図 4　Skip Encoder

入力は　$f, \{, \}, (,), x, =, +, -, ;0, 1, 2, 3, 4, 5, 6, 7, 8, 9, < EOS >, < UNK >$ の 22 種類で $< EOS >$ は数式の始まりを表し，$< UNK >$ は予想される出力が入力の 22 次元内に存在しない時，または未知の入力に対して対応するものとする．

実験にて確認を行ったネットワーク構成を 1 に示す．

表 1　ネットワーク構成

モデル	入力サイズ	Encoder の出力サイズ	LSTM 層の数
Normal	22	200	1
		200	4
		200	8
		500	1
		500	4
		500	8
SkipConnect	22	200	1
		200	4
		200	8
		500	1
		500	4
		500	8
BiDirectional	22	200	1
		500	1

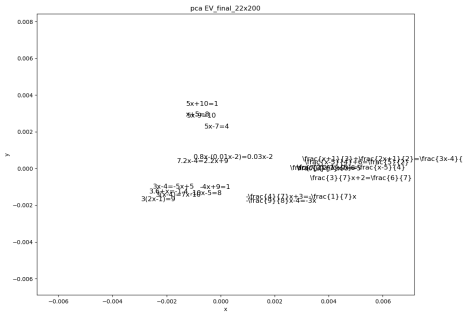

図 5　Normal　 LSTM 1 層 200 次元　学習データ

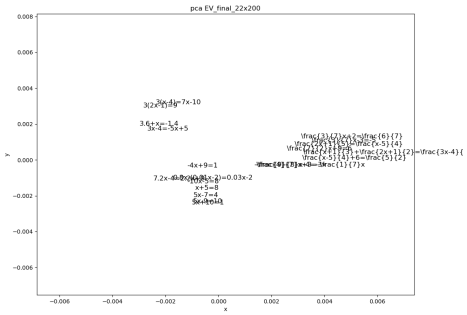

図 6　Skipconnect : LSTM 4 層 200 次元　学習データ

3　実験結果

以下，実験で得た式のベクトルを PCA を用いて二次元に次元削減したものである．学習データとして用いたのは以下の 20 個の式である．

1. x+5=8
2. 5x-7=4
3. 3.6+x=-1.4
4. 3x-4=-5x+5
5. \frac{7}{2}x+9=6
6. 3(2x-1)=9
7. \frac{3}{7}x+2=\frac{6}{7}
8. -\frac{4}{7}x+3=-\frac{1}{7}x
9. \frac{x+1}{3}+\frac{2x+1}{2}=\frac{3x-4}{2}
10. -4x+9=1
11. 5x-9=10
12. 7.2x-4=2.2x+9
13. 3(x-4)=7x-10
14. \frac{3}{7}x-3=-5
15. -\frac{9}{8}x-4=-3x
16. \frac{x-5}{4}+6=\frac{5}{2}
17. \frac{2x+1}{5}=\frac{x-5}{4}
18. 0.8x-(0.01x-2)=0.03x-2
19. -10x-5=8
20. 5x+10=1

以下，実験の結果の抜粋である．

図 5, 図 6，図 7 のどれもが同系列そうな問題が近い分布となった．特に図 6 は分数の中でも符号でも分かれており，高精度でベクトル化に成功しているように見える．このことより系列変換を用いて数式の分類は可能であり応用の余地がある事が確認できた．ただし，層を深くすることや，次元をあげるのでは結果が大きく向上する事がない事が判明しており，より精

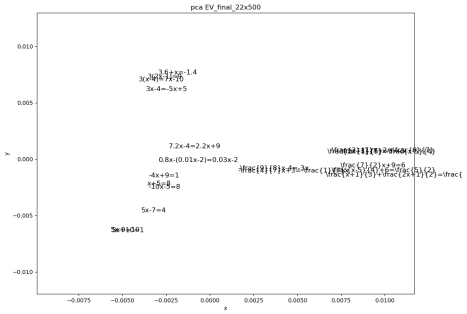

図 7　BiDirection : LSTM 4 層 500 次元　学習データ

度の高いベクトル取得には様々なチューニングが必要だと思われる．

4　まとめ

実験結果より数式をベクトル表現し類似問題に分類する事ができた．これは系列変換の可能性を広げる結果となり，自然言語処理の分野で発展してきた技術は人工言語でも利用することができることが示せた．これを用いて類題の選出に取り組んでいきたい．

しかしながら，未だ問題点はあり，計算問題には簡約な問題ほどパターンがなくなり，式から有用な情報が取り出せなくなる恐れがある．また，多くの子供が苦手とする文章題は本研究では扱っていないが，Dec2Vec[2] など文章をベクトル化する手法は提案されており，それを応用することにより可能性は広がるだろう．また，本研究で扱った一次方程式の特徴量を使って別の分野を分類し，個人のまちがえた問題の情報を元に転移学習を行うとより，包括的なアダプティブラーニングシステムの構築が行えるようになり自分で学ぶ際の道しるべとなることを期待している．

参考文献

[1] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013).

[2] Mohler, M., Rink, B., Bracewell, D. B. and Tomlinson, M. T.: A Novel Distributional Approach to Multilingual Conceptual Metaphor Recognition, *COLING* (2014).