ニューラルネットワークを用いた記述式問題の自動採点

寺田 凛太郎† 久保 顕大† 柴田 知秀†‡ 黒橋 禎夫†‡ 大久保 智哉§
†京都大学 ‡科学技術振興機構 CREST §大学入試センター
†‡{terada, kubo, shibata, kuro}@nlp.ist.i.kyoto-u.ac.jp §okubo@rd.dnc.ac.jp

1 はじめに

現在、中央教育審議会の答申 (2014 年 12 月) を受け、文部科学省を中心に大学入学者選抜改革が議論されている。この改革の柱の一つとして、現行の大学入試センター試験に代わる大学入学希望者学力評価テスト (仮称) に記述式問題を導入することが検討されている。受験者数が数十万人規模の試験において記述式問題を導入する場合、採点の一貫性の確保に加えて、その金銭的・時間的コストが大きな問題となる。

この問題に対して、計算機による採点支援を考えることができる。まず、解答テキストのクラスタリングを行い、各採点者が類似した解答群を採点することで、一貫性向上とコスト削減に寄与することができると考えられる。さらに、解答に含まれるべき重要なキーワードをハイライトするような採点インタフェースの工夫も考えられる。これらに加えて、記述式の採点では各解答に対して複数人が採点を行う必要があることから、計算機による自動採点を 2 人目または 3 人目の採点者とすることもコスト削減となる。この場合、同一問題に対する人間の大量の採点結果があることから、採点を機械学習による分類問題と捉えることができる。本稿ではこのような背景から、文章分類における手法を採点へと応用することを考える。

記述式問題の例 (科目は世界史) とその解答例、採点結果を図 1 に示す。今回は採点を正解・不正解の 2 値としている。素朴な分類方法としては、Bag of Words を素性として SVM などを用いる方法が考えられるが、単語のみに着目すると、句レベルで分類に有効な素性の抽出ができない。そこで本稿では、畳み込みニューラルネットワーク (CNN) により、単語よりも大きな単位での有用な素性の抽出をして解答の分類を行う。数百人規模の解答データを用いて実験したところ、90% 近くの精度を達成できることがわかった。

問題: （ある活動家 (ガンジー) の活動内容についての説明文を受験者に呈示した上で）当時、このような活動家が世界規模で活躍できたのはある理由が考えられる。当時の時代背景を考え、以下の解答欄を埋めて理由を答えなさい。
【30 文字以内】ため、このような世界規模での活躍が可能になった。

正解	20 世紀に入ると、鉄道や蒸気船による交通網が整った
正解	帝国主義政策により汽船・鉄道等の交通網の整備・拡充がなされた
不正解	列強の植民地支配に対し、独立の考え方が主流になってきた

図 1 記述式問題と解答の例

2 関連研究

エッセイの自動採点の研究は盛んに行われており、TOEFL テストの採点にも用いられている e-rater [1] や、日本語における小論文評価システム Jess [5] がある。しかし、どちらも単語のバリエーションや出題文との Topic の関連性などを特徴量に用いており、内容が正しいかどうかについては触れられていない。記述式答案の自動採点としては中島が内容語の Bag of Words を素性とする SVM を用いて検討を行っている [6]。

また、ニューラルネットワークを用いて文章分類問題を解く研究が行われている。ニューラルネットワークのモデルとしては、文の木を基に再帰的に文のベクトルを構成する再帰型ニューラルネット (RNN) [3] や、局所的な情報を畳み込むことでフレーズ単位での判断を下せる畳み込みニューラルネット (CNN) [4] などがある。これらのモデルは映画レビューの極性判定や、質問文の分類などのタスクにおいて用いられており、ベースライン手法である SVM に並ぶ、もしくは、それ

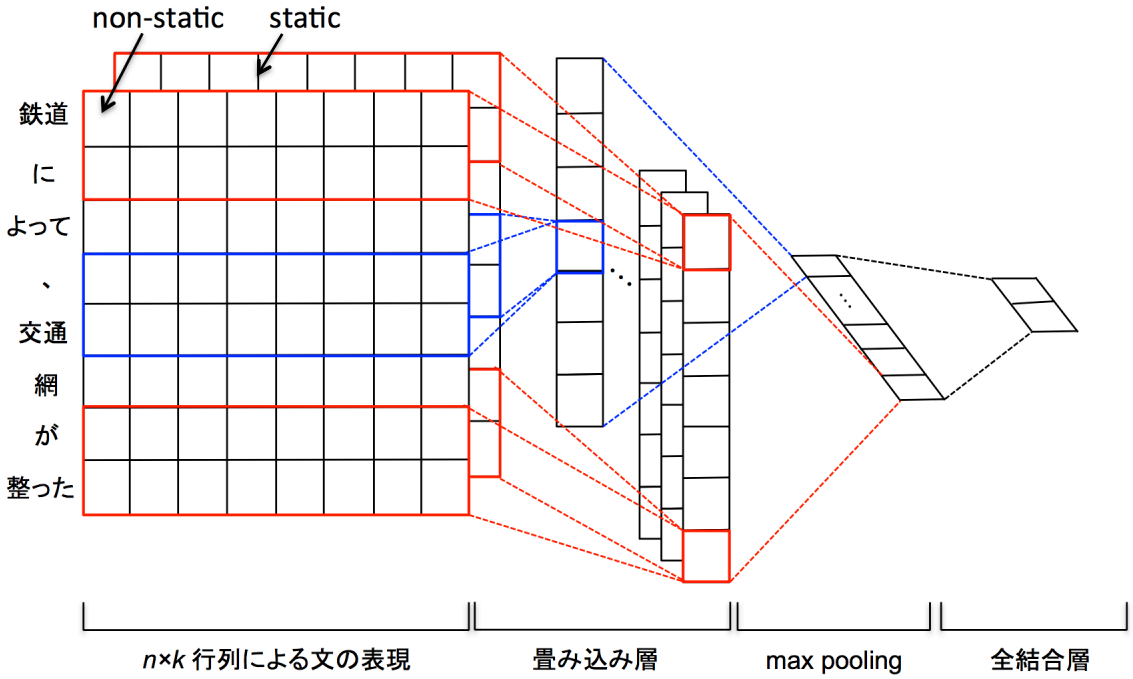

図 2　Multichannel CNN モデルの構造

を超える精度を出している。

3　手法

解答の正解/不正解の分類問題を解く手法として、3.1 節では SVM を、3.2 節では畳み込みニューラルネットワークを用いたものを述べる。

3.1　SVM

ベースラインとして、Bag of Words(BOW) などを素性とした SVM による分類を行う。内容語の BOW に加えて、係り受けを素性とする。例えば、文 “交通網の整備が進んだ” からは以下の素性が抽出される。

- 内容語 BOW: {“交通”, “網”, “整備”, “進む”}
- 係り受け: { “交通網 - 整備”, “整備 - 進む”}

3.2　畳み込みニューラルネットワーク

3.2.1　Multichannel CNN

Kim の提案した Multichannel CNN [4] を用いる。このモデルの構造を図 2 に示す。n 個の単語からなる文を入力とし、各ラベル (本研究では正解と不正解の 2 ラベル) の確率を出力とする。入力文に含まれる i 番目の単語の k 次元 word embedding を $\mathbf{x}_i \in \mathbb{R}^k$ と表わすと、文 $\mathbf{x}_{1:n} \in \mathbb{R}^{n \times k}$ は連結記号 $\oplus$ を用いて以下のように表わされる。

$$\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_n. \tag{1}$$

i 番目の単語から h 単語分のフレーズ $\mathbf{x}_{i:i+h-1}$ に対して、h 単語の窓サイズを持つフィルタ $\mathbf{w} \in \mathbb{R}^{h \times k}$ をかけ、特徴 c_i を計算する。

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b). \tag{2}$$

ここで b はバイアス項、f は非線形関数である。最初の単語から、最後の単語が入るところまで 1 単語ずつフィルタをずらしながら畳み込みをすると、feature map $\mathbf{c} = [c_1, c_2, \cdots, c_{n-h+1}]$ が得られる。こうして得られた feature map $\mathbf{c}$ の中から最大値 $\hat{c} = \max \mathbf{c}$ を抜き出し (max pooling)、それをこのフィルタの取り出した特徴とする。つまり、1 つのフィルタから 1 つの新たな特徴が得られることとなる。本モデルでは複数のフィルタを用意し、複数の特徴を得る。こうして得られた複数の特徴を全結合層へ入力し、各ラベルの確率に相当するソフトマックス出力層へと繋ぎ、システムの出力を得る。そして、正解との誤差を逆伝搬することで重みを更新する。

さらに、このモデルでは 2 つのチャンネルを用いており、片方のチャンネル (non-static と呼ぶ) では word embedding の重みまで誤差を逆伝搬し、もう片方のチャンネル (static と呼ぶ) は word embedding の重みを初期化時から変化させない。この 2 つのチャンネルに同様に複数のフィルタをかけ、その出力の平均をとっ

たものを全結合層への入力としている。これにより、事前学習により得た一般的な単語表現と、今回の問題に特化した単語表現の中間を取ることで、日本語として自然な範囲で問題の文脈に応じた単語の意味を得ることが期待できる。

3.2.2 Dependency-based CNN

前節の手法で局所的なフレーズは捉えることはできるが、離れた単語からなるフレーズは同一のフィルタ内に収められないことがある。そこで、係り受け関係を利用して、意味的に関連がある語を同一のフィルタ内に収めて畳み込みを行う方法を考える。ここでは、Ma ら [2] の研究にならい、各単語についてフィルタの幅分だけ係り受け関係で親にあたる単語を辿り、畳み込み演算を行う方法を用いる (Dependency-based CNN, DCNN)。こうすることにより、例えば解答文「交通網が飛躍的に発達した」において「交通網が」と「発達」が離れているが係り受け関係にあるので同一のフィルタ内に収めることができる。係り受け関係で親にあたる単語がない場合には、全てが 0 の k 次元 word embedding で表わされるような記号を仮に親の単語とすることにする。

4 実験

4.1 記述答案データ

大学生に高校学習範囲の試験問題を解いてもらい、解答データを集めた。生物 1 問、世界史 1 問、化学 1 問、国語 4 問の計 7 問を用い、化学、世界史の解答数は 459、その他は 253 であった。すべての解答は 1 文からなり、平均文字数は 32.4 であった。そして、各解答に正解/不正解の 2 値のラベルを付与し、実験データとして用いた。

4.2 実験設定

全ての実験において形態素解析器 JUMAN[*1] と構文解析器 KNP[*2] を用いた。KNP では “1 つの自立語とそれに続く付属語” からなる基本句という単位で係り受け関係の解析を行うが、Dependency based CNN の実験ではこの基本句内の自立語のみを用いた。

word embedding の重みの初期化には、word2vec[*3] によって Web テキスト 1 億文で学習させた 200 次元の word embedding を用いた。学習済み word embedding が存在しなかった単語に関しては、[-0.25, 0.25] の一様分布を用いてベクトルを初期化した。全結合層の重み $\mathbf{w}$ の L_2 ノルム $||\mathbf{w}||_2$ について、$||\mathbf{w}||_2 > s$ ならば $||\mathbf{w}||_2 = s$ になるように $||\mathbf{w}||_2$ を制限する正則化を行った。本実験では $s = 3$ とした。さらに正則化として、最後の全結合層での dropout(dropout 率 0.5) を行った。

CNN、DCNN ともにフィルタ幅 h について [1, 2, 3] を用いた 1-3gram と [1, 2, 3, 4, 5] を用いた 1-5gram の 2 種類の設定で実験を行った。どちらにおいてもフィルタの数は各幅につき 100 枚ずつ用いた。また、CNN、DCNN いずれにおいてもミニバッチサイズを 20 とし、20 エポック学習を行った。

すべての実験において、leave-one-out 交差検定 (1 文をテストデータとし他の文全てを訓練データとすることを全ての文に対して行う) を用いた。

5 結果

科目ごとの各手法の精度を表 1 に示す。ここで、平均とはマクロ平均を指す。majority baseline とは、全てを多数派のラベルに分類した場合の精度である。どの問題においても、今回実験した機械学習手法を用いて、majority baseline を超える精度を出すことができた。

SVM については、BOW に係り受け関係を追加することで精度が上昇した。CNN を用いた手法は SVM(BOW) とほぼ同じ精度を達成しているが、“SVM(BOW)+ 係り受け” よりも精度が劣っている。CNN に係り受け関係を追加した DCNN とくらべても “SVM(BOW)+ 係り受け” の方が精度が高い。これは、今回の実験ではデータセットが小さく、出現する単語、係り受け関係の種類が少なかったことが理由ではないかと考えられる。データ数が大きくなり、語彙が増えるほど、単語の意味的な類似度を考慮できる word embedding、素性を選択的に利用できる CNN の利点を活かすことができると考えられるので、今後、より大規模な実験で提案手法の有効性を示したいと考えている。1-3gram と 1-5gram を比較すると、CNN では 1-5gram の方が精度が高いが、DCNN では精度が下がっている。これは係り受け関係は 3gram までみれば十分であることを示している。

次に CNN が単語の類似度を考慮しつつ、素性を選択した例を示す。冒頭の図 1 にあげた世界史の問題において、クラス判定に大きく寄与すると判断されたある 2

[*1] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
[*2] http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[*3] https://code.google.com/p/word2vec/

表 1　各手法の精度 (平均はマクロ平均)

手法	化学	生物	世界史	国語 1	国語 2	国語 3	国語 4	平均
majority baseline	71.0	85.4	88.5	68.4	90.5	81.8	92.5	82.6
SVM (BOW)	**92.4**	**87.7**	91.3	88.5	92.9	91.3	**98.4**	91.8
＋ 係り受け	**92.4**	85.8	**92.6**	89.3	**94.5**	**98.4**	97.6	**92.9**
CNN 1-3gram	89.3	85.8	90.6	87.7	93.7	96.4	98.0	91.6
CNN 1-5gram	90.6	85.0	90.8	**91.7**	93.7	96.4	97.6	92.3
DCNN 1-3gram	88.7	**87.7**	91.7	84.2	94.1	97.2	97.6	91.6
DCNN 1-5gram	91.5	87.0	90.4	77.5	94.1	96.8	98.0	90.8

表 2　CNN が抽出した特徴的な句の例

	フィルタ 1	フィルタ 2
正解クラスの重み	0.76	0.07
不正解クラスの重み	−0.82	0.48
特徴的な句の例	交通網 が 整備 交通網 が 発達 蒸気船 が 発達	植民地 内 に 植民地 時代 に 植民地 間 で

つのフィルタの正解、不正解クラスへの重みと各フィルタに特徴的な句の例を表 2 に示す。ここで特徴的な句とは、3.2.1 節の式 (2) をすべての句に適用し、c_i の値が高い句のことを表す。フィルタ 1 は正解クラスへの判定に寄与したフィルタで、「交通網が整備」や「交通網が発達」などのフレーズを含む解答は正解になりやすいことを示し、また、フィルタ 2 は不正解クラスへの判定に寄与したフィルタで、「植民地内に」や「植民地時代に」などのフレーズを含む解答は不正解になりやすいことを示している。word embedding を用いることにより、例えば、「整備」と「発達」が本問題の解答としては同じような意味を表すことを捉えながら、クラス分類に有効な句を抽出することができていることが分かる。

6　まとめ

本稿では、大学入学希望者学力評価テストの記述式問題導入を背景とし、記述式問題の自動採点を分類問題として解く方法について述べた。数百人規模の解答データを用いて実験したところ、SVM、畳み込みニューラルネットワーク (CNN) ともに 90% 近くの精度で分類できることがわかった。

ニューラルネットワークによる他の手法を適用することや、出題の内容など、現在は利用していない情報も用いて更なる精度の改善を目指すこと、より大規模なデータでの評価などが今後の課題である。

参考文献

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 2006.

[2] Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. Dependency-based convolutional neural networks for sentence embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 174–179, Beijing, China, 2015.

[3] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013.

[4] Kim Yoon. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.

[5] 石岡恒憲, 亀田雅之. コンピュータによる小論文の自動採点システム Jess の試作. 計算機統計学, Vol. 16, No. 1, pp. 3–19, dec 2003.

[6] 中島功滋. 機械学習を利用した短答式記述答案の自動識別. 日本教育工学会 第 26 回全国大会, pp. 639–640. 日本教育工学会第 26 回全国大会講演論文集, 2010.