Script-aware embedding を用いた文字表現の獲得

長澤駿太 † 北田俊輔 ‡ 彌冨仁 †‡

† 法政大学 理工学部 応用情報工学科
‡ 法政大学 理工学研究科 応用情報工学専攻
{shunta.nagasawa.2u@stu., shunsuke.kitada.8y@stu., iyatomi@}hosei.ac.jp

概 要

日本語の漢字や中国語などはそれぞれの文字が表意性を持つため，この性質を捉えることはこれら言語の意味理解において重要な手かがりとなる．文字を画像として扱い CNN 等を用いて，形状情報を低次元ベクトルに埋め込むモデルは，こうした特徴を捉えることで文書分類タスクにおいて成果を上げている．しかしながら日本語では平仮名や片仮名等の表音文字が多く使われる文書に対しては適切な文字表現を得ることが難しい．本研究では文字形状を学習した visual feature と文脈情報を学習した context feature の 2 つの文字表現手法を用いることで，表意文字および表音文字を考慮した文字表現の学習手法である script-aware embedding を提案する．本報告では文書分類のタスクにおいて，提案手法の評価を行った．

1 はじめに

日本語は表意文字と表音文字など複数種類の文字体系を用いる言語である．また表意文字の持つ一文字あたりの情報量はアルファベットなどの表音文字と比べて多いことが報告されている [1]．したがって文字の表意性に着目し，文字の形状を考慮することは文書解析の重要な手がかりとなる．

英語の文書解析を行う場合と比較して，日本語や中国語などは形態素解析を行う必要がある．しかし，形態素解析は日々増え続ける未知語や表記ゆれに弱く，正確な単語分割は困難とされている．これらを踏まえると，日本語や中国語では単語単位の処理に比べ文字単位での処理の利点が多い．実際にアジア圏の言語に対する文書分類タスクにおいて，文字単位を入力する手法が単語単位より高い精度であると報告されている [2]．

文字単位の入力を行う深層学習による文書解析手法では，character-level convolutional neural network (CLCNN) が英語の文書解析において高い予測精度を達成している [3]．このモデルは各文字に対する one-hot 表現を一次元 CNN に入力しているが，英語と比べ文字種が圧倒的に多い日本語などでは入力次元数が大きくなるため，過学習を引き起こしてしまう．

文字種の多い日本語や中国語に対しては，漢字の表意性に着目した低次元の文字表現手法や文書解析手法が存在する．具体的には，文字を画像として扱い，convolutional auto encoder (CAE) を用いて明示的に文字形状を保持するような学習を行うことで，低次元表現に埋め込む手法が提案されている [4]．こうした文字形状を考慮した文字表現を入力として，後段の CLCNN で分類を行うことで，文書分類タスクにおいて極めて良好な分類能が実現されている．さらに，文

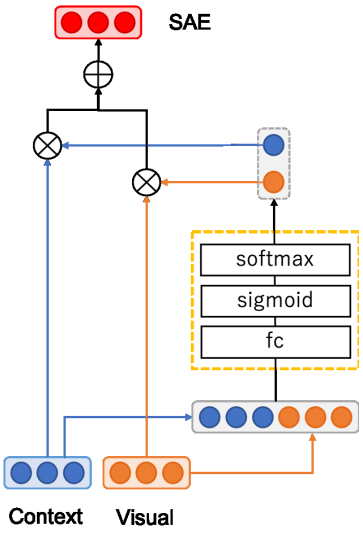

図 1: Script-aware embedding (SAE) の全体図．文脈情報を考慮した文字表現である context feature と，文字形状を考慮した文字表現である visual feature を入力する．点線部分に示す modality attention でこれらの特徴を組み合わせた表現である SAE を得る．

字形状を考慮した文字表現の学習と文書分類モデルを end-to-end で行うモデルが複数提案されている．例えば CNN をベースとした character encoder (CE) で文字画像を低次元表現にエンコードする部分と，文書分類モデルを同時に学習するモデル [5] は，文字の形状的特徴だけではなく，より文書解析に適した文字表現の獲得を実現しており，分類能の向上に大きく寄与している．こうした end-to-end の手法において，特徴空間上の data augmentation に加えて，文字画像空間での data augmentation を導入することで汎化性能や予測精度の向上が報告されている [6].

これらの研究では中国語や日本語の漢字に含まれる表意性の解釈をより効果的に行うことに焦点が当てられており，出現頻度が多いひらがなやカタカナといった表音文字に対しては焦点を当てられていない．また，漢字においても「犬」「大」など意味が違うが形状が似ている文字や,「僕」「私」など形状は違うが意味が似ている文字なども多く存在し，文字の形状に着目する上記の手法には改善の余地が残されている．

一方，画像や言語など異なるモーダルを組み合わせるマルチモーダルの分野では，文中の各単語表現，文字表現，画像情報の異なる表現を attention 機構により取捨選択する modality attention [7] が提案され，SNS データセットに対する固有表現抽出タスクで精度の向上が報告されている．

本研究では表意文字や表音文字などの文字体系を考慮した文字表現手法である script-aware embedding (SAE) を提案する．これは文字形状を学習した文字表現と，文脈情報を学習した文字表現に対して，modality attention を適用し，組み合わせることで文字体系を考慮した文字表現を学習する．

評価実験では誤字脱字などが多く，単語分割が困難とされる楽天市場の商品レビューデータセットを用いて，感情分析の二値分類タスクとレビューの星の数を推定する回帰タスクの 2 つに対して評価を行った．

2 提案手法

提案手法である script-aware embedding は表意文字および表音文字といった文字体系を考慮した文字表現手法である．本研究では文字の形状情報を学習する visual feature と文脈情報を学習する context feature を modality attention によって組み合わせることで，それぞれの長所を考慮した文字表現を獲得する．

表 1: CAE の encoder, decoder のアーキテクチャ(カーネルサイズ k，出力チャンネル数 o)

(a) Encoder

Layer	Encoder
1	Conv(k=(3, 3), o=16) → ReLU
2	Maxpooling(k=(2,2))
3	Conv(k=(3, 3), o=16) → ReLU
4	Maxpooling(k=(2,2))
5	Conv(k=(3, 3), o=16) → ReLU
6	Linear(o=64)

(b) Decoder

Layer	Decoder
1	Linear(o=400) → ReLU
2	Deconv(k=(3, 3), o=16) → ReLU
3	Upsampling(scale=2)
4	Deconv(k=(3, 3), o=16) → ReLU
5	Upsampling(scale=2)
6	Deconv(k=(3, 3), o=16) → ReLU
7	Deconv(k=(3, 3), o=1) → Sigmoid

2.1 Visual feature

Visual feature は文字形状を考慮した学習をすることで得られる文字表現である．日本語常用文字として平仮名，片仮名，漢字　(JIS 第一・二水準)，英数字，記号を含む計約 6,000 字に対し，それぞれの文字を 36×36 pixels のグレースケール画像に変換した．その後 CAE にこれらの文字画像を入力することで，文字形状を考慮した低次元の文字表現を獲得する．実験では，CAE の中間表現を visual feature として用いた．CAE の encoder，decoder のそれぞれのアーキテクチャを表 1 に示す．

2.2 Context feature

Context feature は文脈情報を考慮した学習をすることで得られる文字表現である．周囲の単語を用いて対象の単語の埋め込みを学習する continuous bag of words (CBOW) [8] を文字単位に拡張し，学習することで文脈情報を考慮した文字表現を獲得する．この際，事前学習のコーパスには日本語 Wikipedia のテキストデータを使用した．またパラメータは window size を 5 とした．

表 2: 楽天レビューデータセットに対する評価結果

	二値分類タスク	回帰タスク	
	Accuracy ↑	MSE ↓	R^2 ↑
lookup only (dim=64)	0.926	0.729	0.636
visual only (dim=64)	0.928	0.653	0.674
concatenate (dim=64)	0.932	0.627	0.687
concatenate (dim=128)	**0.933**	**0.618**	**0.691**
(ours) SAE (dim=64)	**0.933**	0.626	0.687

2.3 Script-aware embedding

表意文字や表音文字を考慮するため，事前学習済みの visual feature と context feature を modality attention によって組み合わせる．通常，複数ベクトルを組み合わせる際はそれぞれのベクトルを連結する concatenate の手法が用いられるが，未知文字が出現した場合 context feature の文字表現は零ベクトルとなり，不安定な文字表現となってしまう．そのため，重み付け和を取ることで零ベクトルの影響少なくすることが可能な modality attention を用いた．

具体的には，各文字に 2 つの文字表現に対する重み $a_{\text{visual}}, a_{\text{context}}$ を計算し，重み付け和を出力する．

$$\boldsymbol{x} = [\boldsymbol{x}_{\text{visual}}; \boldsymbol{x}_{\text{context}}] \tag{1}$$

このとき，$\boldsymbol{x}_{\text{visual}}, \boldsymbol{x}_{\text{context}}$ はそれぞれ visual feature と context feature のベクトルを表し，$\mathbf{x}$ は 2 つのベクトルを連結したものとなる．

$$[a_{\text{visual}}, a_{\text{context}}] = \text{softmax}(\sigma(\boldsymbol{W}\boldsymbol{x}^\top + \boldsymbol{b})). \tag{2}$$

このとき，$\boldsymbol{W}, \boldsymbol{b}$ は学習パラメータを，σ は活性化関数である sigmoid 関数を表す．sigmoid 関数の出力を softmax 関数の入力にすることで，片方の feature のみに attention の重みが偏らないようにしている．

$$\boldsymbol{v} = a_{\text{visual}}\boldsymbol{x}_{\text{visual}} + a_{\text{context}}\boldsymbol{x}_{\text{context}} \tag{3}$$

ここで $\boldsymbol{v}$ は最終的なベクトル表現を表す．SAE の全体図を図 1 に示す．

3 実験及び結果

本実験では楽天市場のレビューデータセットを用いて，以下の二値分類タスクおよび回帰タスクを行うことで script-aware embedding の有効性を確認した．

3.1 データセット

評価には楽天市場の商品レビューデータセット[1]を用いた．これはレビュー文とそれに対応する評価値である星の数がラベルとして与えられており，レビュー文には誤字脱字や表記ゆれなどが多く存在している．

二値分類タスク　商品のレビューのテキストデータから評価値である星の数 1,2 をネガティブ，4,5 をポジティブとし，それらを推定する二値分類タスクを行った．データ数は学習用に 80 万件，テスト用 40 万件を使用し, 学習用を 3 : 1 の割合で訓練用と検証用に分割した．評価指標には accuracy を用いた．

回帰タスク　商品のレビューのテキストデータから評価値である星の数を推定する回帰タスクを行った．データ数は学習用に 100 万件，テスト用 50 万件を使用し, 学習用を 3 : 1 の割合で訓練用と検証用に分割した．評価指標には平均二乗誤差 (MSE) と決定係数 (R^2) を用いた．

3.2 比較手法

提案手法である script-aware embedding の有効性を確認するため，従来の埋め込みを用いた文字表現のみを利用する lookup only，文字形状を考慮した文字表現のみを利用する visual only，またこれらを連結する concatnate に対して，2 種類のタスクで比較を行った．

- **Lookup only**：各文字を低次元表現に埋め込み，その表現を文書分類モデルと同時に end-to-end で学習する手法．

[1] https://github.com/zhangxiangxiao/glyph

- **Visual only** [6] :文字画像から文字表現獲得する CE と文書分類モデルを end-to-end で同時に学習する手法．CE のアーキテクチャには表 1a と同じものを用いた．
- **Concatenate** [5] : 上記の 2 つの手法を使用し，それぞれの文字表現を連結する手法．他の手法と比べ文字表現の次元数は 2 倍となる．
- **Script-aware embedding (SAE)** : 事前学習にて得られた，context feature と visual feature を modality attention を用いて組み合わせる提案手法．それぞれの feature は固定する．

3.3 実験設定

本実験の文書分類モデルのアーキテクチャは Bi-LSTM 2 層と全結合層から構成される．Bi-LSTM の中間表現と全結合層の中間表現は 256 次元とした．

入力はレビュー文から 100 文字ランダムに切り出し，それぞれの文字表現の次元数は 64 とした．最適化手法は Adam を用い，学習率 0.001，バッチサイズは 256 で学習を行った．また過学習抑制として wildcard training [4] を用いた．

3.4 実験結果と考察

文書分類の結果 表 2 に二値分類タスクおよび回帰タスクの結果を示す．次元数が同じである他の手法と比べ，提案手法である SAE が最も精度が高いことがわかる．modality attention を用い，文字の形状情報とコンテキスト情報を取捨選択することでより良い文字表現が学習できたと考えられる．また全体を見ると concatenate (dim=128) モデルが一番良い精度を示しているが，これは文字表現の次元数が 2 倍になることで表現力がより強くなったことが考えられる．

文字表現の評価 各文字表現手法のクエリ文字に対する最近傍文字を表 3 に示す．クエリ文字には形状が似ているものに意味の異なるものが多い「犬」と，意味が似ているものに形が異なるものが多い「僕」という字を用いた．lookup と concatenate の最近傍文字では解釈性のあるものが見られなかった．visual では「犬」というクエリに対し，文字形状の近い「大」「太」など意味に関係ないものが見られていることがわかる．提案手法では「犬」に対しては同じ動物である「猫」「猿」といった文字が,「僕」に対しては同じ一人称である「俺」や，人を指す「誰」など解釈性のある結果が得られた．

表 3: それぞれの文字表現手法の最近傍

クエリ文字	lookup	visual	concatnate	(ours)SAE
僕	蔽	償	豊	俺
	陽	俟	燕	…
	鉾	僅	揚	誰
犬	激	大	鼠	猫
	閣	太	8	獣
	販	天	賽	猿

Attention を解析したところ，visual feature 成分を主に使う文字は，使用頻度の低い文字多く見られた．したがって未知文字や使用頻度の低い文字が零ベクトルとなることを回避できていると考えられる．

4 おわりに

本研究では表意文字や表音文字などの文字体系を考慮した文字表現手法である script-aware embedding (SAE) を提案した．実験では文書分類タスクに対して提案手法の有効性が確認できた．また，文字表現の解析では先行研究よりも解釈性のある文字表現が獲得できた．今後は文字体系に対するさらなる解析や文書分類以外でのタスクなどの評価を行っていきたい．

参考文献

[1] G. Neubig and K. Duh, “How much is said in a tweet? a multilingual, information-theoretic perspective,” in *AAAI Spring Symposium on Analyzing Microtext*, 2013.

[2] X. Zhang and Y. LeCun, “Which encoding is the best for text classification in chinese, english, japanese and korean?” *arXiv preprint arXiv:1708.02657*, 2017.

[3] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Proc. of NIPS*, 2015, pp. 649–657.

[4] D. Shimada, R. Kotani, and H. Iyatomi, “Document classification through image-based character embedding and wildcard training,” in *Proc. of IEEE Big Data*. IEEE, 2016, pp. 3922–3927.

[5] F. Liu, H. Lu, C. Lo, and G. Neubig, “Learning character-level compositionality with visual features,” in *Proc. of ACL*, 2017, pp. 2059–2068.

[6] S. Kitada, R. Kotani, and H. Iyatomi, “End-to-end text classification via image-based embedding using character-level networks,” in *Proc. of IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2018, pp. 1–4.

[7] S. Moon, L. Neves, and V. Carvalho, “Multimodal named entity recognition for short social media posts,” in *Proc. of NAACL*, 2018, pp. 852–860.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR preprint arXiv:1301.3781*, 2013.