分野名の粒度に基づくタイトルの自動分類

志村　和也　　福本　文代

山梨大学 工学部
{t13cs023, fukumoto}@yamanashi.ac.jp

1　はじめに

WWW と Web ブラウザの爆発的な普及を背景に, Twitter やブログ, タイトルなど短く情報量が少ないスパースな文書データがオンライン上で利用可能となっている．本研究ではスパースな文書を文書と呼ぶ．文書を自動的に分類する手法として教師付き学習を用いた分類が主流となっている [9]．これは, 人手によりラベル (分野名) が付与された文書の集合を訓練データとして用い, 機械学習を適用することにより分類器を作成し, この分類器を用いテスト文書をラベルが付与された文書集合のいずれかに分類する方法である．教師付き機械学習を用いた文書分類の精度は, 訓練データの量に依存する．しかし, 一般に文書はマルチラベル, すなわち複数の分野名が付与されている．従って, 分野の粒度が細かい場合, 人手により複数の分野名を文書へ付与することは多大なコストと労力を要する．このことから, 大量の訓練データを作成することは難しい．

本研究では新聞記事のタイトルを対象とし, Convolutional Neural Network(CNN) を用いて複数の分野へ分類する手法を提案する．CNN は画像認識分野において優れた成功を収めているニューラルネットワークの一つであり, 写真のタグ付けや自動運転車などのコンピュータビジョンの中心となっている [4]．近年, 文書分類においても CNN が利用されており, 高い分類精度が得られることが報告されている．

Kim は, 単純な構成の CNN を用いて, 感情表現やカテゴリ分類を含む様々なデータセットの精度を評価している．ネットワークは非常に単純な構成であるにもかかわらず, 高い精度が得られることが報告されている [7]．Santos らは, Twitter などの短い文書に対する感情表現を CNN を用いて分類している [11]．彼らは単語だけではなく, 文字に対しても CNN を適用している．Berger[2] や Johnson ら [5] は, マルチラベル問題に着目し, CNN を用いて分類を行っている．しかし, スパースな文書は対象としていないため, CNN が有効であるか否か検証する余地がある．

本研究は文書を粒度の細かい分野へ分類するために, CNN において Fine-tuning を用いた転移学習を利用する．Fine-tuning は, 画像認識において, あらかじめ汎用性の高い大規模な教師付きデータでネットワークを学習し, この学習済みのネットワークを用い, さらに細かい判別に対する学習を行うアプローチである．特に学習に利用できる訓練データが限られている場合, 有効であることが示されている [1]．本研究では, Fine-tuning を用い, 分類が容易である粒度の荒い分野から徐々に粒度の細かい分野へと分類することにより, タイトルの高精度な分類を目指す．

2　提案手法

提案手法は (1) 単語の分散表現の獲得, (2) Fine-tuning を用いた転移学習, (3) マルチラベル分類の 3 つから構成される．提案手法の流れを図 1 に示す．

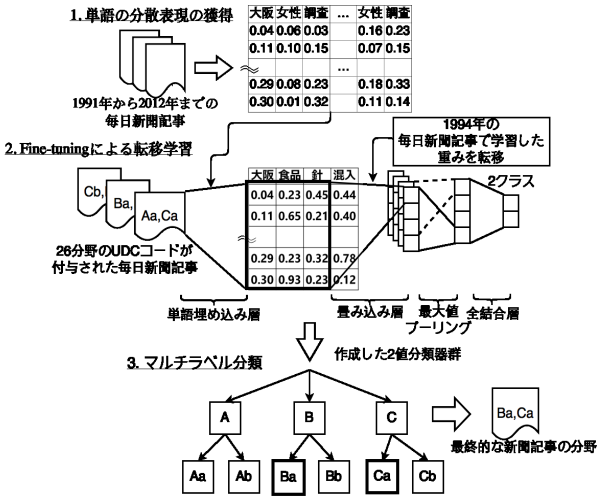

図 1　提案手法の流れ

2.1 単語の分散表現の獲得

画像を対象とした CNN の入力は画素値の 2 次元ベクトルであり, 文書を入力するには文字列をベクトルで表現する必要がある. 文書を数値化する方法として, 文書内の各単語もしくは各文字をベクトルで表現する方法が考えられる. Zhang[13] らは文字レベルの畳み込み計算を行っており, 100 万件ほどの文書データがあるデータでは良い精度が得られることが報告されている. しかし文書データが少ないデータでは CNN による畳込みの精度向上は見られない. 本研究ではスパースな文書を対象とするため, 各単語をベクトルで表現する.

単語のベクトル表現には局所表現と分散表現が用いられる. 局所表現とは, 各次元がある単語であるかどうかを表すベクトル表現である. 例えば「コンピュータ」という単語を表すときには, 「コンピュータ」に対応する次元が 1 でその他が 0 のベクトルとなる. このようなベクトル表現を one-hot 表現という. one-hot 表現では各単語間の関係を捉えることができない上に, 文書中の語彙数が多くなるとベクトルの次元数も増加してしまう. そこで単語の意味を低次元で表現できるベクトル表現として分散表現が考えられた. 単語の分散表現を獲得するツールとして fastText[6][3] がある. fastText は文書を入力し, 中間層と出力層からなる 2 層のニューラルネットで学習を行い, 中間層における各単語の重みを抽出することにより, 各単語の分散表現を獲得することができる. fastText は学習に木構造を利用することにより, 従来の Word2Vec[8] などのツールと比べて高速な学習が可能となっている. 本研究ではこの fastText を利用し, 各単語をベクトルで表現する. 単語の分かち書きには MeCab の解析結果を利用した. 図 1 の 1. では, 大量の分かち書きされた文書を入力とし, 各単語の分散表現が出力される. 200 次元の単語の分散表現を使った 10 単語の文書があった場合には 200×10 の 2 次元ベクトルとなる.

2.2 Fine-tuning を用いた転移学習

転移学習とは, ある分野の問題が知識やデータが不足していることが原因で十分に解決できない場合, 他の関連した分野のデータや学習した結果を利用する学習手法である. 本研究では, 知識を転移するドメインを元ドメインと呼び, 知識が転移されるドメインを目標ドメインと呼ぶ.

提案手法では, 先ずデータ数が目標ドメインよりも多い, 元ドメインの文書データを用い, CNN による学習を行う. CNN は畳込み層とプーリング層という 2 種類の層を含む順伝播型ネットワークである. 本研究で利用するネットワークを図 1 の 2. に示す. 第 1 層目は単語埋め込み層である. この層で文書を 2 次元ベクトルに変換する. 第 2 層では畳込みを行い, 畳込みの結果にプーリング層を適用する. その後, 3 層の識別層が続き, 文書がどの分野に分類されるかという事後確率が出力される.

CNN を用いた転移学習に Fine-tuning を用いたアプローチがある. これは, 元ドメインのデータで学習を行ったネットワークの辺の重みやバイアス (パラメータ) を解きたい分野, つまり目標ドメインを学習するネットワークの初期値として利用する方法である. CNN による学習は初期のネットワークのパラメータに依存する. 学習に利用することができるデータ数が少ない場合は過学習を防ぐためにも, よりよい初期値を獲得する必要がある. また, CNN は入力層から出力層に近づくに連れて段階的にそれぞれの訓練データに特化した特徴が学習されていくことが知られている. 本研究では単語埋め込み層と畳み込み層のパラメータを目標ドメインを学習するネットワークの初期値として再学習を行い, その他の層は目標ドメインを用いて新たに学習を行う. これにより, ネットワークを一から学習する場合よりも, 良い結果を得られることが期待できる.

2.3 マルチラベル分類

図 1 の 3. では分野名の階層構造を利用したマルチラベル分類を行う. 本研究が利用するデータは, 国際十進分類法に基づく UDC コードが付与された 1994 年の毎日新聞記事のタイトルである [14]. UDC コードは世の中の全ての知識を十進の 0 から 9 までの数字を使い, 繰り返し細分化を行っていく分類手法である. これにより分野名の上下の関係に階層性が生じる. また, 1 つの新聞記事には複数の分野名が付与されている. マルチラベル分類は一対多分類の手法を拡張することで実現できる. 一対多分類は, 対象の分野とその他の分野を識別する 2 値分類器を学習することで 2 値以上の分野を分類する手法である. 各分野の 2 値分類器によって新聞記事タイトルの分野を予測し, 予測された分野の和集合をとったものが得られる分野となる. この分類方法で階層の深さが同じ分野ごとに分野を予測し, 次に予測された分野の下位分野で同様に分類を行う. 例えば, あるタイトルに対して「スポーツ」と「経済」の 2 値分類器がそれぞれの分野に属すると判断した場合, このタイ

表 1　実験に利用したデータ

	1994 年の新聞記事	UDC コード付新聞記事
文書数	51,249	18,122
分野数	9	26
階層数	なし	2

トルは「スポーツ」の下位分野と「経済」の下位分野で再度分類が行われる．階層の末端になるまで分類を続け, 最終的な分野を得る．

3　実験

3.1　データ

実験では, 1994 年の毎日新聞記事と RWCP コーパス [10] を用いた．コーパスからタイトルを抽出し, CNN による分類を行った．RWCP コーパスは 1994 年の毎日新聞記事, 30,207 記事からなるコーパスで, 複数の UDC コードが付与されている．毎日新聞記事には UDC コードとは別に各記事が掲載された紙面名が付与されている．その中から「1 面」や「特集」といった分野を除いた 9 分野を, 粒度の荒い分野として利用した．また, この毎日新聞 9 分野と関連する分野を UDC コードから 26 分野選択し, これを粒度の細かい分野として利用した．実験では, 毎日新聞 9 分野に属する新聞記事のタイトルを元ドメインデータとして学習を行い, その知識を転移して UDC26 分野に属するタイトルの分類を行った．実験で用いたデータについて表 1 に示す．表 1 の UDC コードが付与された新聞記事は 2 階層から構成され, 第 1 階層が 5 分野, 第 2 階層が 26 分野に分かれる．毎日新聞 9 分野に属する記事は階層がない．訓練データを全体の 8 割利用し, 残りの 2 割をテストデータとして利用した．単語の分散表現の獲得には, 1991 年から 2012 年までの毎日新聞記事を利用し 200 次元の分散表現を用いた．

3.2　実験結果と評価尺度

提案手法との比較として Support Vector Machines(SVMs) [12] と fastText[6][3] を用いた．SVMs の特徴量には Bag-of-Words 特徴量を利用した．以降, Fine-tuning を用いた CNN による分類を CNN(Fine), Fine-tuning を用いず, 一から学習した CNN による分類を CNN(Full) と表記する．実験は 5 分割交差検証を用いて行った．評価尺度として F 値のマイクロ平均とマクロ平均を求めた．

第 1 階層における実験結果を表 2 に, 第 2 階層における実験結果を表 3 に示す．表 2 より, 第 1 階層では CNN(Fine) がマイクロ平均とマクロ平均共に最も高い精度となっている．一方, CNN(Full) は最も低い値であった．このことから Fine-tuning による効果が得られていることが確認できる．CNN(Full) の精度が最も低くなった原因としてデータ数の少なさによる過学習が考えられる．

表 3 によると, 第 2 階層においては, マイクロ平均では CNN(Fine), マクロ平均では fastText が最も高い値を示した．CNN(Fine) のマクロ平均が fastText よりも低くなってしまった原因として, 分野ごとの文書数の偏りが考えられる．ニューラルネットワークの学習手法としてミニバッチ学習がある．これは訓練データからミニバッチと呼ばれる少量の集合を取り出し, ミニバッチごとに学習を行う方法である．分野ごとに文書数に偏りがある場合, ミニバッチ内にも偏りが生じ, 学習に悪影響を与えることが考えられる．さらに CNN(Fine) を含む全ての手法においてマクロ平均がマイクロ平均よりも低くなっていることから, 文書数が多い分野「スポーツ」といった分野では学習ができているが, 「絵画」のような少ない分野では学習できていないことが分かる．

4　おわりに

本研究では, CNN により新聞記事のタイトルを分類する手法を提案した．粒度の細かい複数の分野へ分類するために, Fine-tuning に適用し, 分類が容易である粒度の洗い分野から徐々に細かい分野へと分類することでタイトルの高精度な分類を目指した．SVMs と fastText を用いた比較実験の結果, Fine-tuning による転移学習の効果が確認できた．今後の課題として, (i) CNN の精度を改善, (ii) ネットワーク構成の調整, (iii) 単語の階層構造の導入, 及び (iv) 文書データの拡張が

表 2　第 1 階層における提案手法と他手法の実験結果

手法	マイクロ平均	マクロ平均
SVMs	0.744	0.709
fastText	0.734	0.696
CNN(Full)	0.725	0.677
CNN(Fine)	**0.756**	**0.714**

表 3　第 2 階層における提案手法と他手法の実験結果

手法	マイクロ平均	マクロ平均
SVMs	0.606	0.353
fastText	0.595	**0.384**
CNN(Full)	0.571	0.324
CNN(Fine)	**0.610**	0.333

挙げられる.

参考文献

[1] Pulkit Agrawal, Ross B. Girshick, and Jitendra Malik. Analyzing the Performance of Multilayer Neural Networks for Object Recognition. *CoRR*, Vol. abs/1407.1610, , 2014.

[2] Mark J. Berger. Large Scale Multi-label Text Classification with Semantic Word Vectors. 2015.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*, 2016.

[4] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to End Learning for Self-Driving Cars. *CoRR*, Vol. abs/1604.07316, , 2016.

[5] Rie Johnson and Tong Zhang. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 919–927. Curran Associates, Inc., 2015.

[6] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*, 2016.

[7] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *The 2014 Conference on Empirical Methods In Natural Language Processing*, pp. 1746–1751, 2014.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[9] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pp. 91–100, New York, NY, USA, 2008. ACM.

[10] RWC. Text Database (Japanese), Real World Computing, 1995.

[11] Cicero Nogueira Dos Santos, MaraGattit. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *COLING 2014 the 25th International Conference on Computational Linguistics*, pp. 69–78, 2014.

[12] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[13] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 649–657. Curran Associates, Inc., 2015.

[14] 社団法人情報科学技術協会. 国際十進分類法 日本語中間版第 3 版 分類表・索引. 丸善株式会社, 1994.