機械学習を用いたQAサイト質問文のカテゴリの類推

加藤 玲大† 馬 青† 村田 真樹‡
†龍谷大学大学院理工学研究科 ‡鳥取大学大学院工学研究科
t15m002@mail.ryukoku.ac.jp, qma@math.ryukoku.ac.jp,
murata@ike.tottori-u.ac.jp

1 はじめに

昨今，インターネットを介したコミュニケーションや情報共有が盛んに行われており，QAサイトや掲示板，SNS等の利用者が増加している．その中で，身近な悩みや疑問を解決する手段としてYahoo!知恵袋やOKWave等のQAサイトが多く利用されている．

QAサイトでは多くのカテゴリが存在し，質問者は投稿する質問内容に該当するカテゴリを選択する必要がある．また，専門家を含むQAサイトの利用者が回答できる質問を探す場合，得意なカテゴリから探し出すことが基本となる．質問者が質の良い回答を得たい場合，カテゴリの選択は重要であると考えられる．多くのQAサイトでは，質問を投稿する際に，入力された質問内容から類推して適しているカテゴリの候補を表示するサービスが提供されている．しかし，カテゴリを予測する手法は明らかにされていない．また，類推の精度も必ずしも高くなかったと思われる．本研究は，機械学習を用いて質問文に適しているカテゴリを高精度に類推することを目標としている．

我々は先行研究として深層学習手法の一種であるStacked Denoising Autoencoders(SdA)とDeep Belief Network(DBN)を用いてカテゴリの類推を行った[1]．深層学習がカテゴリの類推に有効であるかを確認するため，SdA，DBNと従来の機械学習手法であるMulti Layer Perceptron(MLP)，Support Vector Machine(SVM)との類推の精度を比較した．更に，入力ベクトルの次元数の違いによる類推の精度を比較した．その結果，最も高い精度がSdAの0.735であった．

本研究では，先行研究より汎化性能を向上させ高精度に類推するため，正則化の1つであるDropoutをMLP，SdA，DBNの学習時に利用した．また，用いる学習データに対し，内容が重複するカテゴリを統合することによりデータの品質を改良した．更に，データをベクトルに変換する際に，質問の本文のみならずタイトルも使うようにした．その結果，類推精度が先行研究より大幅に向上することができ，最も高い類推精度がDropoutを加えたSdAの0.824であった．

2 データセット

本研究では，OKWaveに投稿されている質問文を用いて作成したコーパスを使用する．質問文は質問の本文とタイトルで構成されている．現在OKWaveのカテゴリは大分類，中分類，小分類の3つの分類で構成されている．コーパスを作成する際には，大分類の「デジタルライフ」カテゴリに属する中分類から，10種類のカテゴリを選んで用いる．作成されたコーパスは合計でおよそ33,000の質問文で構成される（詳細は[1]を参照されたい）．

このように作成したコーパスをこのまま1つのデータセットとして用いる．これは先行研究で使用したデータセットと同一のものであり，本稿ではデータセット1と呼ぶ．

しかし，コーパスには意味の近いカテゴリが存在する．中分類の「AV機器」に属する小分類の「カメラ全般」と，中分類の「マルチメディア」に属する小分類の「デジタルカメラ」は意味の近いカテゴリである．そのため，デジタルカメラについての質問が両方に投稿されている．そこで，コーパスの「カメラ全般」と「デジタルカメラ」を統合し，1つの小分類のカテゴリとする．加えて，統合したカテゴリを,「マルチメディア」に属するカテゴリとする．そのようにして改良したコーパスを新しいデータセットとして用い，本稿ではデータセット2と呼ぶ．

3 ベクトル変換

本章ではデータセット内の個々の質問文のベクトルへの変換について述べる．

3.1 質問の本文のみを用いたベクトルの生成

これは先行研究と同じ方法である．以下の手順(1)〜(4)でデータセット1からベクトルの要素となる単語を抽出する．

(1) 質問文を形態素解析し，名詞（固有名詞，サ変接続，一般）を抽出する．
(2) 名詞が連続しているのであれば，結合し1つの単語とみなす．
(3) 単語は全角・半角を統一し，英単語は全て大文字で統一する．

(4) 各ラベルから出現頻度がトップ N 以内の単語を抽出する．

このようにして抽出した単語をベクトルの要素とし，個々の要素はその単語が出現すれば 1，出現しなければ 0 の 2 値を取る．コーパスにはベクトルの要素を 1 つも含まない（質問文に抽出した単語が 1 度も出現しない）質問文が含まれているが，ベクトルを生成する際にその質問文は用いないこととする．

ベクトルの次元数による類推の精度を比較するため，トップ N の単語をそれぞれ 300（その結果，ベクトルが 1,276 次元），500（同 2018 次元）とした．

3.2 質問のタイトルと本文の両方を用いたベクトルの生成

コーパスを調べると，以下のような問題が存在することが判明した．例として，質問の本文を「タイトルの通りです。詳しい方宜しくお願い致します」とし，タイトルを「外付け HDD の番組を DVD にコピーする方法」とするような，タイトルに質問の内容が含まれる質問文が存在する．そのため，質問の本文とタイトルを合わせて 1 つの質問文とする．

データセット 2 に対し，上記の処理を行った後，ベクトルの要素となる単語を手順 (1)〜(4) を用いて抽出する．トップ N の単語をそれぞれ 300（その結果，ベクトルが 1,309 次元），500（同 2,083 次元）とした．

4 機械学習によるカテゴリの類推

カテゴリの類推は，入力された質問文から分類器が内容に適してるカテゴリを予測する．本研究では機械学習の MLP，SVM，SdA，DBN を分類器として用いる．ここで MLP，SVM の記述は省略し，深層学習の SdA/DBN と Dropout について簡単に述べる．

4.1 深層学習

深層学習とは，従来のニューラルネットワークを多層構造にした機械学習手法の総称である．SdA や DBN は層ごとに教師なしの事前学習（pre-training）を行い，その結果を初期値とし教師ありの事後学習（fine-tuning）を行う．SdA，DBN はそれぞれ Denoising Autoencoder（dA）[2]，Restricted Boltzmann Machine（RBM）[3] を積み重ねて多層のネットワークを構築している．dA は決定的モデルであるのに対して，RBM は確率モデルである．

4.2 Dropout

Dropout は，ニューラルネットワークのユニットを確率的に選別して学習する手法である [4]．層ごとに指定した割合だけ，毎回ランダムに選択したユニットの出力値を 0 にする．学習時のユニットの選出確率を α とした場合、推論時には各ユニットからの出力を α 倍にする．異なる一部のネットワークの結合が取り除かれた場合においても識別できるように学習するため，汎化性能を向上させることができる．

5 実験

5.1 実験条件

本研究で用いる SdA，DBN 及び MLP は Deep Learning Tutorials[5] に記載されているスクリプトを利用し，SVM は機械学習ライブラリ scikit-learn の SVM を利用する．Dropout は SdA，DBN の Pre-training には用いず，Fine-tuning にのみに用いる．

データセット 1 は 3.1 節に従い次元数の異なる 2 つのベクトルを構成する．また，データセット 2 は 3.2 節に従い次元数の異なる 2 つのベクトルを構成する．データセット 1，2 を学習用，検証用，テスト用の 3 つに分類する．1 カテゴリあたりに学習データを 2,000 個，検証データを 400 個，テストデータを 400 個用意する．すなわち，10 カテゴリの全データにおいては，学習データ，検証データ，テストデータをそれぞれ合計 20,000，4,000，4,000 個用意する．

各機械学習手法の最適なハイパーパラメータは，グリッドサーチを行うことで決定する．各機械学習手法のハイパーパラメータの組み合わせの数がほぼ同等になるように設定されており，その数は MLP で 228，SdA 及び DBN で 216，SVM で 225 通りとする．例として，SdA，DBN のハイパーパラメータは，隠れ層の構造，pre-training での学習係数及び学習回数，fine-tuning の学習係数，活性化関数である．隠れ層の構造については，1 層から 3 層とし，各層のユニット数は層ごとに減っていくものと，各層が同じものとした．活性化関数については，Sigmoid 関数と ReLU を用いる．

グリッドサーチで最適化するハイパーパラメータ以外の数値設定については，ミニバッチサイズ 100，L2 正則化項 0.0001，モメンタム項 0.9，ドロップアウトでのユニットの選出確率は入力層で 0.8，隠れ層で 0.5 とする．学習アルゴリズムはミニバッチ勾配降下法，誤差関数は交差エントロピーを用いる．

5.2 実験結果

5.2.1 データセット 1 を使用した場合

ハイパーパラメータの組み合わせを検証誤差の小さい順に並べ，上位 N 個（ただし N = 1,5,10,...,30）を用いた場合の各機械学習手法の平均精度を図 1，2 に示す．ただし，図 1 は 1,276 次元の特徴ベクトルを用いた場合のテストデータに対する平均精度であり、図 2 は 2,018 次元の特徴ベクトルを用いた場合の平均精度である．なお，本論文に用いられている平均精度はすべてマクロ平均で算出したものである．

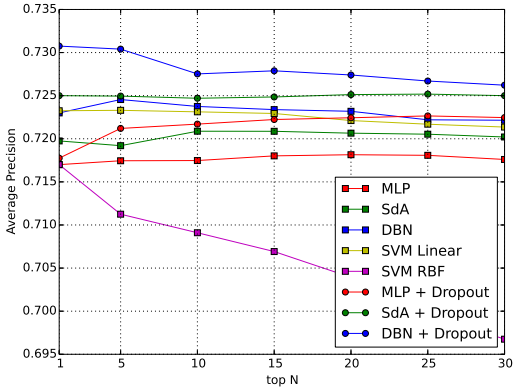

図 1: 1,276 次元での各機械学習手法の平均精度

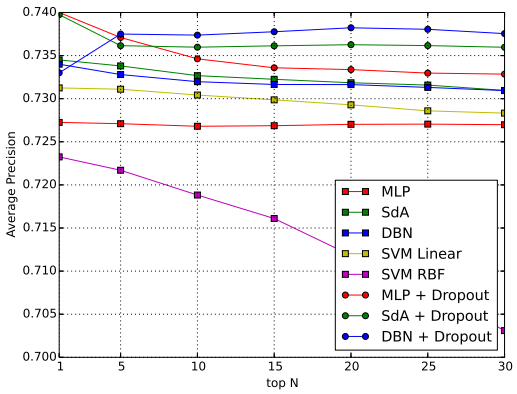

図 2: 2,018 次元での各機械学習手法の平均精度

表 1: データセット 1 を用いた類推精度

機械学習手法	1,276 次元	2,018 次元
MLP	0.717	0.727
SdA	0.720	0.735
DBN	0.723	0.734
SVM Linear	0.723	0.731
SVM RBF	0.717	0.723
MLP+Dropout	0.718	**0.740**
SdA+Dropout	0.725	**0.740**
DBN+Dropout	**0.731**	0.733

図 1 より，DBN+Dropout が最も平均精度が高いことが分かる．次いで SdA+Dropout が高い平均精度を出している．一方で，MLP+Dropout は平均精度が全般的に高くないことが見てとれる．図 2 より，全般的に見れば DBN+Dropout の平均精度が最も高いことが分かる．検証誤差が一番小さいときの平均精度を見ると MLP+Dropout，SdA+Dropout が最も高いことが分かる．図 1，2 より，MLP，SdA，DBN の全てにおいて Dropout を加えることで，平均精度を向上することができる．

表 1 に各機械学習手法の類推精度を示す．ここでの類推精度とは，最も検証誤差の小さい（N = 1）ハイパーパラメータを用いたときの精度である（以降，これを類推精度とする）．1,276 次元において DBN+Dropout が 0.731 と最も類推精度が高く，2,018 次元においては MLP+Dropout と SdA+Dropout が 0.740 と最も類推精度が高い．また，特徴ベクトルの次元数が増えることにより，全ての機械学習手法の類推精度が高くなることが分かる．

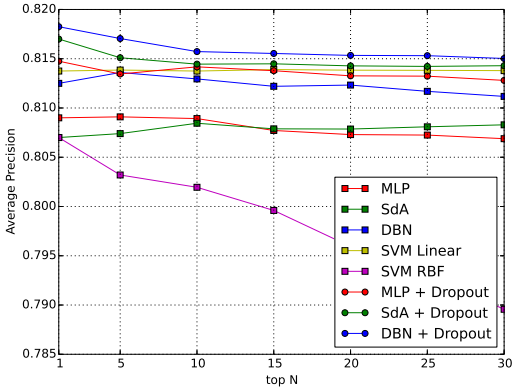

図 3: 1,309 次元での各機械学習手法の平均精度

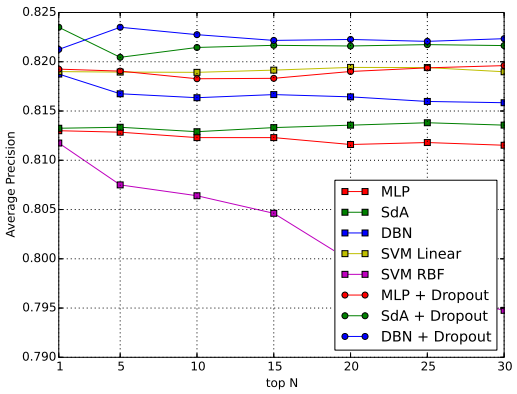

図 4: 2,083 次元での各機械学習手法の平均精度

5.2.2 データセット 2 を使用した場合

前節と同様に各機械学習手法の平均精度を図 3，4 に示す．ただし，図 3 は 1,309 次元の特徴ベクトルを用いた場合のテストデータに対する平均精度であり，図 4 は 2,083 次元の特徴ベクトルを用いた場合の平均精度である．

図 3，4 より，Dropout を加えた手法が全般的に高い平均精度である．SVM Linear は MLP+Dropout と同等の平均精度である．また，Dropout を加えていない深層学習手法と比べて SVM Linear の平均精度が高いことが見てとれる．

表 2 に各機械学習手法の類推精度を示す．1,309 次元において DBN+Dropout が 0.818 と類推精度が最も高く，2,083 次元においては SdA+Dropout が 0.824 と最も類推精度が高い．SVM Linear は 1,309 次元の

表 2: データセット 2 を用いた類推精度

機械学習手法	1,309 次元	2,083 次元
MLP	0.809	0.813
SdA	0.807	0.813
DBN	0.813	0.819
SVM Linear	0.814	0.819
SVM RBF	0.807	0.812
MLP+Dropout	0.815	0.819
SdA+Dropout	0.817	**0.824**
DBN+Dropout	**0.818**	0.821

表 3: カテゴリごとの類推精度

カテゴリ	データセット 1	データセット 2
AV 機器	0.703	0.840
携帯・スマートフォン・PHS	0.748	0.820
SNS	0.853	0.892
ネットショッピング	0.915	0.930
ウィルス対策	0.818	0.890
Windows	0.637	0.698
Macintosh	0.608	0.873
PC パーツ・周辺機器	0.762	0.848
ソフトウェア	0.667	0.708
マルチメディア	0.688	0.738

とき SdA，DBN より高い精度であり，2,083 次元のとき SdA よりも精度が高く，DBN と同等の精度であることから，Dropout を加えていない深層学習手法よりも性能が良いことが分かる．

5.3 考察

データセット 1，2 の両方において，入力の次元数を増やすことで精度が向上しており，各機械学習手法において入力次元数を増やすことが重要であることが分かる．また，Dropout を用いて学習した MLP，SdA，DBN が全て高い精度であるため，カテゴリの類推において Dropout を用いることで性能を向上させることができることが分かる．

表 3 にデータセット 1，2 を使用したときの最も類推精度が高い機械学習手法に対する，カテゴリごとの類推精度を示す．使用した手法は SdA+Dropout である（データセット 1，2 はそれぞれ，2,018 次元，2,083 次元を用いる）．

表 3 より，データセット 2 を使用した場合，全てのカテゴリでデータセット 1 を使用した場合より類推精度が高いことが分かる．これにより，コーパスの品質改良と質問文におけるタイトルの利用が精度向上に有効であることが確認できた．ただし，データセット 1，2 はデータの内容が全く同じではないため，厳密な比較ではない．

6 おわりに

本稿では，機械学習手法を用いて QA サイト質問文から適切なカテゴリを類推する手法を提案した．深層学習手法に Dropout を加えて学習させることで，他の機械学習手法より高精度にカテゴリを類推することができた．加えて，カテゴリの統合によるデータセットの品質改良や，入力データに質問文のタイトルを加えることにより，類推精度を向上することができた．先行研究で最も高い類推精度は SdA の 0.735 であり，本研究の最も高い類推精度は SdA+Dropout の 0.824 であることから，先行研究より高精度に類推することができた．

今後の課題として，深層学習の Pre-training に Dropout を用いる実験や，グリッドサーチに用いるハイパーパラメータ設定の見直しが挙げられる．また，QA サイトの全てのカテゴリの類推を実現するために，学習時に用いるカテゴリを増やすことや，小分類のカテゴリを使用することなど，データの規模を拡大した場合において高精度で類推することが挙げられる．

謝辞

本研究は科研費（25330368）の助成を受けたものである．

参考文献

[1] 加藤玲大，馬青，村田真樹． 深層学習を用いた QA サイト質問文のカテゴリ分類． 情報処理学会研究報告，Vol. 2016-NL-228, No. 10, pp. 1-6, 2016.

[2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol. Stacked Denoising Autoencoders : Learning Usefull Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, Vol. 11, pp. 3371-3408, 2010.

[3] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory, In *Parallel distributed processing: explorations in the microstructure of cognition*, Vol. 1, pp. 194-281, 1986.

[4] N. Srivastava, G. E. Hinton, A. Krizhevsky, I.Sutskever, and R. Salakhutdinov. Dropout: A Simple way to prevent neural networks form overfiting. *Journal of Machine Learning Research*, 15:1929-1958, 2014.

[5] http://www.deeplearning.net/tutorial