博 士 学 位 論 文

深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援

2022 年 3 月

滋賀県立大学大学院　工学研究科　先端工学専攻

安藤　雅行

摘要

近年，深層学習を用いたAIシステムが急速に活躍の場を増やしている．画像認識を始めとし，自動車の自動運転やドローンを用いた荷物の自動配達，医師の診断アシストなどさまざまな分野において，活用が進められている．日本の大手製薬会社でも，英のAI研究機関と共同で，AIによる新しい化合物の探索を行うことで，薬の開発を大幅に効率化する試みが行われている．

一方で，深層学習には学習による予測・分類の基準が不明で人間に理解できないという大きな問題が存在している．この問題は医療分野や自動運転など，学習結果の信頼性･安全性が重要視されている分野では特に重大な課題とされている．また，自然言語処理分野においても深層学習の判断基準を人間が深く理解することが可能であれば，例えば新人とベテランの書いた電子カルテの違いから，良い電子カルテを書く方法を理解し，アンケートやレビューを分析して商品開発のヒントになる情報を得るなど，深層学習の新しい活用が期待される．

これを受け本研究では，文章の分類問題を題材として，学習済みの深層学習の重み付きネットワークから，重みによって学習に使用した文章の特徴の重要度を算出し，深層学習の分類根拠である分類パターンの解釈を支援するシステムを提案する．特に，データ分析の初心者でも分類パターンの解釈を行えるシステムを構築することで，クラウドベースの機械学習API（Application Programming Interface）を利用するユーザや個人で簡易なテキストマイニングを行いたいユーザが，その学習結果を容易に解釈できる環境を構成できると考えている．

最初に，最も基本的な深層学習モデルである，DNN（Deep Neural Network）モデルから分類パターンを抽出し，解釈支援を行うシステムの構築を行った．ここで，学習済みのDNNの重みの値によって，そこを通過する情報の重要度が決まると考え，重みの積から出力に寄与する特徴（本研究では文章中の単語が該当する），または，各中間層ノードに寄与する特徴を取得する．取得した情報は可視化インタフェースによって解釈支援システムの利用者に提供される．可視化インタフェースでは，特定の分類先の出力ノードと強く結びついた中間層ノードが表示され，その中間層ノードにはそのノードと強く結びつく特徴が表示される．解釈支援システムの評価実験では，動物の生態に関する文章，映画のレビューに関する文章，受験に関するツイート集合を題材とし，各文章集合について14人の被験者に提案システムを用いて解釈を行ってもらった．その解釈を解釈内容が文章集合に対して

妥当かどうか調査したところ，90%近くの解釈が妥当であることがわかり，提案システムの解釈に対する有効性が確認された．

続いて，実際に自然言語処理分野で広く扱われているRNN（Recurrent Neural Network）について，分類パターンを抽出し，解釈支援を行うシステムの構築を行った．ここでは，RNNの重み付きネットワークを一つのHMM（Hidden Markov Model）と捉え，適当な単語のパターンに対する尤度を算出することで，時系列情報を含んだ分類パターンを取得を行った．取得した分類パターンは，単語をノード，単語の時系列を矢印付きエッジとする解釈支援ネットワークとして表示し，解釈支援システムの利用者に提供される．解釈支援システムの評価実験では，アニメや漫画のキャラクターのセリフ集合，Amazonの家電商品とゲームソフトのレビュー集合を題材とし，8人の被験者には提案システムを用いて各文章集合に対する解釈を行ってもらった．また，実験では，比較システムとして，TFIDF（Term Frequency Inverse Document Frequency）を用いたシステムを用意し，比較システムを用いて別の8人の被験者に同じ実験を行ってもらった．提案システムと比較システムにおいて，解釈内容が文章集合に対して妥当かどうか調査したところ，比較システムでは85%近くの解釈が妥当と判断された一方，提案システムでは95%以上の解釈が妥当であることが確認された．また，比較システムで10%程度見られた妥当でない解釈は，提案システムでは見られなかった．したがって，提案システムは解釈に有効であり，さらにその有効性は単語の時系列情報によるものだと確認できた．

以上より，本研究で提案する深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムは，学習ネットワークの重み情報を活用することで深層学習が学習した文章集合の情報を分類パターンとして抽出することができ，さらに解釈支援インタフェースによってシステム利用者が分類パターンの解釈するのに有効的であると結論づけた．

目 次

1. 緒言 **6**

1.1 深層学習の歴史 6

1.2 深層学習の活用と課題 7

1.3 本研究の目的 8

2. 関連研究 **10**

2.1 深層学習の出力に寄与する入力に注目した研究 10

2.2 深層学習の学習ネットワークの解釈に注目した研究 10

2.3 深層学習の説明可能性に注目した研究 11

3. DNNの重みネットワークを用いたテキスト分類パターンの解釈支援 **13**

3.1 DNNモデルの構造 13

3.1.1 ノードの学習 15

3.1.2 重みの更新 16

3.2 DNNを用いたテキストベースの分類パターン解釈支援システムの構成 18

3.3 学習ネットワークからの分類パターンの抽出処理 18

3.3.1 単語集合の抽出 18

3.3.2 学習によるネットワークの重み付け 19

3.3.3 出力ラベルに対する重要パスの抽出 19

3.3.4 重要パス上のノードへのラベルづけ 21

3.4 学習ネットワークからの分類パターンの可視化処理 22

3.4.1 解釈支援ネットワークの形成 22

3.4.2 解釈対象出力の選択 24

3.4.3 解釈対象中間層の選択 26

3.4.4 ノード数・ノードラベル数の変更 28

3.4.5 原文表示機能 31

3.5 システムの使用例 32

3.6 DNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムの有効性の検証実験 39

3.6.1 使用テキストデータと学習モデル 39

3.6.2 実験手順 40
3.6.3 結果と考察 41

4. RNNの重みネットワークを用いたテキスト分類パターンの解釈支援 46
4.1 RNNモデルの構造 46
4.2 RNNを用いたテキストベースの分類パターン解釈支援システムの構成 48
4.3 深層学習による学習ネットワークの形成 49
4.3.1 文中の単語のベクトル化 49
4.3.2 学習によるネットワークの重み付け 49
4.4 学習ネットワークからの分類パターンの抽出処理 50
4.4.1 単語出現パターンの作成 50
4.4.2 重みネットワークのHMMへの変換手法 51
4.4.3 単語出現パターンの尤度算出 53
4.5 学習ネットワークからの分類パターンの可視化処理 54
4.5.1 解釈支援ネットワークの形成 54
4.5.2 原文表示機能 55
4.6 システムの使用例 57
4.7 RNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムの有効性の検証実験 59
4.7.1 使用テキストデータと学習モデル 59
4.7.2 実験手順 60
4.7.3 結果と考察 63

5. 結言 69

謝辞 71

参考文献 72

研究業績 77

1. 緒言

1.1 深層学習の歴史

深層学習とは，ニューラルネットワークを用いた機械学習の一種である．基本的に多層構造をとり，それまでの機械学習と比較して優れた分類精度，予測精度が特徴である．その起源は19世紀ごろにアメリカの心理学者ローゼンブラット[1]が「脳の構造を真似た情報伝達回路を作成すれば，優れた人工知能が作れるのでは」というコンセプトをもとに提唱されたパーセプトロンから成る．このパーセプトロンは，人間の脳神経細胞（ニューロン）をモデル化したもので，主に画像分野を中心に，急激に世界中で研究が進められるようになった．

しかし，1970年代にアメリカのコンピュータ科学者ミンスキー[2]らが「パーセプトロンはノイズに弱く収束が遅い．何より線形分離不可能（データ集合が一本の直線で分類することができない）問題に対して学習ができない」と批判し，一時，現在の深層学習につながる研究は冬の時代になった．その後，中間層の追加，つまりパーセプトロンの多層化や学習方法として勾配法と誤差逆伝播法[3]の発明により，深層学習の基本モデルとなる深層ニューラルネットワーク（DNN : Deep Neural Network）が作られた．それから深層学習のブームは続いたが，1990年代になると，過学習（モデルに対して学習データが少なく，学習データの特徴を不必要に学習してしまい，未知のデータへの対応力が下がる状態）や勾配消失（多層モデルにおいて，誤差逆伝播法で誤差の情報が途中の層で消える状態）の問題などが壁となり，再び現在の深層学習につながる研究は下火になった．

そこで，イギリスのコンピュータ科学者ヒントン[4]らが教師なし学習やドロップアウトなどの新しい学習方法を提唱，2006年ごろからは深層学習の学習周りの諸問題が解決し始めたのを皮切りに，モデルの大規模化や画像認識コンペで深層学習モデルが圧勝する[5]など，深層学習ブームが加熱した．また，この時使われた深層学習モデルは，畳み込みニューラルネットワーク（CNN : Convolutional Neural Network）と呼ばれるモデルで，本論文では詳しくは述べないが，DNNに畳み込み層と呼ばれる，画像の特定の部分に反応するようなフィルタによって，画像の特徴を特徴を取り出して学習する層を追加することで，画像の分類に特化したモデルとなっている[10]．

2010年代に入ると，深層学習は自然言語処理分野にも応用され始め，時系列データを学習できるモデルとして，回帰的ニューラルネットワーク（RNN : Recurrent

Neural Network）とその発展型であるLSTM（Long Short-Term Memory）が機械翻訳などの現場で，大きな成果をあげるようになった[6]．RNN自体は1986年にアメリカの認知学研究者ラメルハートらの研究に基づき考案されたもので，詳細は4章で述べるが，DNNの中間層を時間方向に展開することで多層化したモデルで，主に音声認識[7]の分野で高い精度を出したことで注目され[8]，手書き文字認識[9]など，自然言語処理分野以外でも広く活用されている．

1.2　深層学習の活用と課題

深層学習の歴史について簡単に解説したが，ここからは実際に深層学習がどのように身近で役立てられているか，そしてどのような課題があるのかを述べる．

近年，深層学習を用いたAIシステムが急速に活躍の場を増やしている．画像処理分野では画像認識を始めとし，自動車の自動運転[11, 12]やドローンを用いた荷物の自動配達[13]，医師の診断アシスト[14]などさまざまな場面において，活用が進められている．また，自然言語処理分野では，多言語に対応した文章の自動翻訳[15]に始まり，商品紹介記事の自動生成[16]，新しいものでは企業の採用面接をAIに行わせる技術も出てきている[17]．その他の分野においても，例えば，台湾では，レジに通さずとも商品を手に持って店を出るだけで，正確に商品を判別して決済が行われる無人のコンビニが営業されていたり[18]，日本の大手製薬会社でも，英のAI研究機関と共同で，AIによる新しい化合物の探索を行うことで，薬の開発を大幅に効率化する試みが行われている[19]．

一方で，深層学習には学習による予測・分類の結果の精度自体は非常に優れているが，その判断根拠が不明（正確には，人間に理解することが非常に難しい）という大きな問題が存在している．これは，深層学習が数万，大規模モデルでは数億からなるニューロンで構築されたモデルを持ち，非常に膨大なデータから得られた複雑なロジックで動いているからである．とは言え，AIの挙動を完全にコントロールするためには，その技術を構築しているロジックを把握しなければならず，また，人間に理解できないと言うことは，理解できれば人間にとって新しい知見が得られると言うことでもある．このような背景をもとに，2018年ごろから，深層学習のブラックボックス問題に関する研究は話題になり，政府からも大きく注目されるようになった[20]．

特に，現実的な視点では，医療分野や自動運転など，学習結果の信頼性･安全性が重要視されている分野において，使用されている技術に不明な点があるという

ことは非常に重大な問題であり，早急な解決が強く望まれている．また，自然言語処理分野においても，深層学習の判断基準を人間が深く理解することが可能であれば，例えば新人とベテランの書いた電子カルテを学習したネットワークから，その両者の分類基準を理解し，ベテランの書くカルテの特徴（良い点）と新人の書くカルテの特徴（悪い点）を知識として得られ，結果として良い電子カルテを書く方法を習得することができる．その他では，アンケートやレビューの良い評価と悪い評価を学習したネットワークから，商品の良い点と悪い点の特徴を理解し，売れる商品開発のヒントになる情報を得るなど，深層学習の新しい活用が期待される．

1.3 本研究の目的

前節により，深層学習のブラックボックス問題の解決は早急な問題であると述べた．そこで本研究では，深層学習のブラックボックス問題の解決の一歩として，深層学習モデルが学習した情報を人間が理解できる形に抽出し，その抽出した情報の意味を人間が解釈する支援を行うシステムの構築を目指す．具体的には，分野は自然言語処理分野で文章の分類問題を題材として，学習済みの深層学習の重み付きネットワークから，重みによって学習に使用した文章の特徴の重要度を算出し，深層学習の分類基準である分類パターンを抽出する手法と，抽出した分類パターンの解釈を支援するためのインタフェースや機能を持つシステムを提案する．特に，データ分析の初心者でも分類パターンの解釈を行えるシステムを構築することで，クラウドベースの機械学習APIを利用するユーザや個人で簡易なテキストマイニングを行いたいユーザが，その学習結果を容易に解釈できる環境を構成できると考えている．目的を自然言語処理分野に絞ったのは，画像処理分野では比較的容易に中間層の情報を可視化しやすく，そのためネットワーク解釈の研究がすでに多くされていたためである（詳しくは2章で述べる）．

本研究は，深層学習における，分類基準の判断根拠の理解へのひとつのアプローチとして，分類パターンの解釈支援を行う．そのため，分類結果の精度の向上より，人間がいかに分類基準の根拠を理解できるかという点を重視する．また，本研究で使用する深層学習モデルについては，最も単純な形である多層型のニューラルネットワーク，DNNと，実際に自然言語処理分野で広く使用されているモデルである，単語の時系列情報も学習可能なRNNの２つのモデルを利用する．

以下本論文では，2章では関連研究として，深層学習の出力に寄与する入力に注

目した研究と，深層学習の学習ネットワークの解釈に注目した研究，そして深層学習の説明可能性に注目した研究について述べる．3章ではDNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムとして，提案システムの構成，学習ネットワークの重みから分類パターンを抽出する手法などについて述べ，提案システムの評価実験と結果の考察を述べる．4章ではRNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムとして，提案システムの構成，学習ネットワークの重みをHMMに変換して分類パターンを抽出する手法などについて述べ，提案システムの評価実験と結果の考察を述べる．5章では結論として，本研究で提案する深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムは，システム利用者が分類パターンの解釈するのに有効的であると結論付け，本論文を締めくくる．

2. 関連研究

本章では，本研究に関連する先行研究として,「深層学習の出力に寄与する入力に注目した研究」,「深層学習の学習ネットワークの解釈に注目した研究」,「深層学習の説明可能性に注目した研究」の３つについて解説し，それぞれで本研究との関係について述べる.

2.1 深層学習の出力に寄与する入力に注目した研究

深層学習の分類基準を示す研究としては，アテンションと呼ばれる手法を用いた研究[21]が一般的である．アテンションとは，深層学習において分類・予測を行う際，その出力結果を算出するにあたって入力データのどこに注目したかを学習時に計算し，その結果を利用して分類・予想を行うことで，分類時に注目した特徴の可視化，またはより学習精度を向上させる手法である．主に，画像分野では元々深層学習の注目箇所（分類･予測に寄与する特徴的な人や物）を明示する目的で使われており[22]，後に自然言語処理分野での機械翻訳において，それまで入力データを固定のベクトルに圧縮していたのを，アテンションにより可変長のデータをそのまま保持した状態で学習を進めることで，学習精度が大幅に上がる[23]と注目されるようになった．最新の研究では，アテンション計算を層ごとに行い，より分類･予測精度を高めた研究[24]や，アテンションのみで構築された深層学習[15]なども登場している．しかし，アテンションはあくまで入力と出力の関係（簡単に言えば，入力に対する出力の反応）のみに注目し，内部でどのような学習が行われているか，学習されたネットワークに一体どのような情報が蓄積されたのかは，考慮されているとは言える状況ではない.

そこで本研究では，深層学習が学習時，内部で入力データのどの特徴がどのように出力へ伝達されるのかを算出し，学習ネットワークの解釈を行うための，出力に貢献する分類パターンの抽出を目指す.

2.2 深層学習の学習ネットワークの解釈に注目した研究

深層学習の学習ネットワークの解釈に焦点を当てた研究としては，1.3節でも述べたように画像処理分野でだが，いくつかの研究がなされている[25, 26]．これは，簡単に説明すると，画像を深層学習にて学習させた後，出力ごとに強く反応している学習ネットワークの中間層のノードを取り出し，画像形式に戻すことで，深

層学習が入力画像のどの部分に注目して判断しているか解析するという内容である．しかし，このような中間層の学習過程の推測ができたのは，CNNの構造上，ノードを画素として画像に再変換が容易であったためで，あくまで画像分野に限っての話である．よって，自然言語処理分野になると，中間層の学習した情報を可視化することが難しく，別のアプローチが必要である．

一方，ニューラルネットワークを用いた研究では，中間層の役割と重みの持つ意味について考察したもの[9]や中間層が学習した独自の特徴量を用いて，入出力の様々な関係を得ることを目指したもの[27]があり，中間層に注目すれば，出力に関して意味のある情報が得られることを示している．また，ニューラルネットワークの中で，特に重みに注目した研究として，重みを用いて入力データ間の相関関係を抽出する研究[28]があり，学習によってネットワークに付与された重みが，中間層が学習した情報と取り出す手がかりになることは明確である．

そこで本研究では，画像ではなく自然処理言語の分野において，重みを用いて中間層ごとに学習した特徴を抽出して可視化することで，DNNならば出力に近付くにつれて重要な情報が収束していく様子がわかり，RNNならば学習した特徴の時系列関係がわかるのではと考えた．

2.3　深層学習の説明可能性に注目した研究

近年，深層学習のモデルへの信頼性・公平性の説明や判断基準への理解を重視した研究分野として，XAI（Explainable AI：説明可能なAI）[29]が注目されてきている．XAIの研究としては，深層学習モデルの動作を理解･信頼するために，何を学習したか説明を行うことの必要性の提唱[30, 31]から始まり，実際に，モデル内のデータや変数間の相関関係から動作の説明を試みたり[32]，反事実的条件文を用いてモデルの動作をユーザに理解させる研究[33]等，モデルの動作自体を説明できないか試みる研究が行われている．また，モデルの動作の解釈だけではなく，モデルの動作の安定性･信頼性に注目し，悪意のあるデータへの対策[34]や，モデルの動作を別の論理回路や決定木に当てはめ，モデルの動作やその安定性を評価する研究[35, 36]も存在する．

そこで，本研究では，XAIの問題意識の元，モデル自体の評価ではなく，あくまで深層学習モデルの学習結果に対し，人間が納得できる理由を探す手伝い，つまり分類の根拠となる分類パターンの解釈の支援をするシステムの構築を目指す．モデルの評価については，すでに多くの研究で題材とされていることと，モデルが

信頼できるからと言って，やはり結果がどのように得られたのかが不明では説明可能になっているとは言えないと考え，今回は考察しないこととした．また，本研究は深層学習モデルの学習結果を全て自動で説明するわけではなく，提供された情報から，人間が解釈を行って最終的に判断するシステムとなっている．これは，深層学習（AIシステム）はあくまで人間の判断を補助するための道具であり，人間の判断を肩代わりするようなものではないと考えているためである．

3. DNNの重みネットワークを用いたテキスト分類パターンの解釈支援

本章では，本研究で構築したDNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムについて，システムの構成，分類パターンの抽出手法，解釈支援機能，システムの評価実験について述べる．

3.1 DNNモデルの構造

本研究で構築する深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムについて，本章では，まず，基本的な深層学習モデルであるDNNを使ったシステムの構築を行う．これは，比較的構造が単純な深層学習モデルを使うことで，本研究で提案する，学習済みネットワークに付与された重みから深層学習が学習した情報を抽出する手法の基礎を構築するためである．

図1に，DNNが文章集合を学習し，分類を行うイメージ図を示す．DNNとは，最も単純な深層学習モデルであり，基本的にデータを入力する入力層，学習を行う中間層，学習結果を出力する出力層からなる．中間層は2層以上が一般的であり，基本的に中間層数が多くなるほど高い精度で学習が行える．その反面，中間層が多くなると1章で述べたような勾配消失問題などの影響が出てくるため注意する．層はノードで構成されており，ノードは一つ一つがパーセプトロンとなっている．そしてノード同士はエッジで繋がっており，エッジには学習によって重みが付与される．

図1では，まず正解ラベルが与えられた文章集合について，各文章中の単語をベクトルに変換している．この時，単語ベクトルは単語の出現頻度を0,1で表現するBoWや，単語を200から300次元の固定長のベクトルで表現する分散表現（主にWord2vec）[38]があるが，図ではBoWを使っている．ベクトルに変換された文章集合は，入力層のノードに数値として入力され，その情報は中間層に送られる．中間層では送られた情報を学習していき（詳細は次節で述べる），最終的に出力層に送られ，出力層ノードから分類結果として出力される．この時，中間層はエッジの重みを，その分類精度が高くなるように反復して学習が行われていく．この学習により，中間層は入力された特徴の組合せの重要性を判断していき，出力層に近づくほど，組合せ同士を対象とした大きな塊の組合せなど，より複雑な組合せを対象とした判断が行われる．よって，多層構造のDNNにおいて，中間層ごとの学習された情報

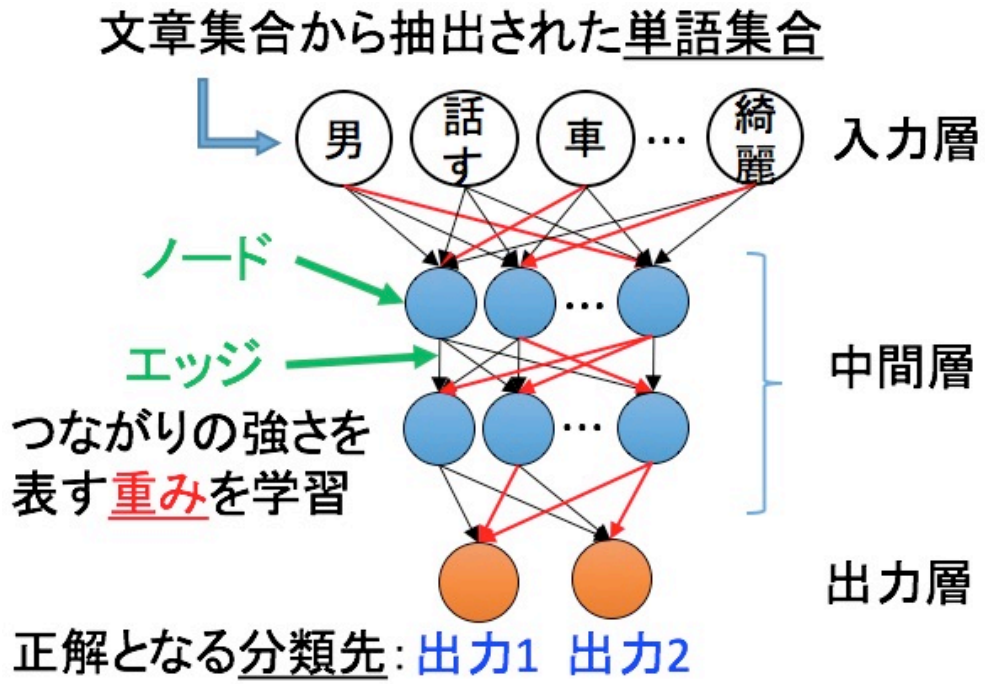

図 1　テキスト分類ネットワーク

を取得できれば，学習の様子を細かく観察することが可能であると言える．本研究においては，学習されたネットワークをもとに，中間層ごとのテキストの分類パターン（分類の根拠を示す情報）を人間が解釈できるように支援を行う．

ここで，学習された分類パターンを解釈するためには，適切な学習が行われていることが前提と考えられる．そのため，学習時の交差検定におけるテストデータ，または学習用のデータセットとは異なるテストデータの分類精度が，90%以上となるよう深層学習で学習させたネットワークを想定し，ネットワークには大きな誤りを含まないと仮定する．

テキストベースの深層学習においては，分類精度を高めるために入力に単語の分散表現を用いることが一般的であるが，本研究は分類精度を犠牲にして解釈性を高めるための手法として位置付けられる．一方で，90%以上の分類精度が得られるネットワークを解釈の対象としているため，必要以上に高い精度を求めても，解釈の結果は大きく変わらないと考えている．

加えて，word2vecなどの分散表現を用いるためには，あらかじめ分類対象テキストに応じた分散表現を事前に学習させるコーパスが必要となる．本研究による分散表現を用いない手法の利点として，分散表現学習のためのコーパスが得られない場合にも適用可能となることや，分散表現学習のための手間を必要としない

点が挙げられる．

3.1.1 ノードの学習

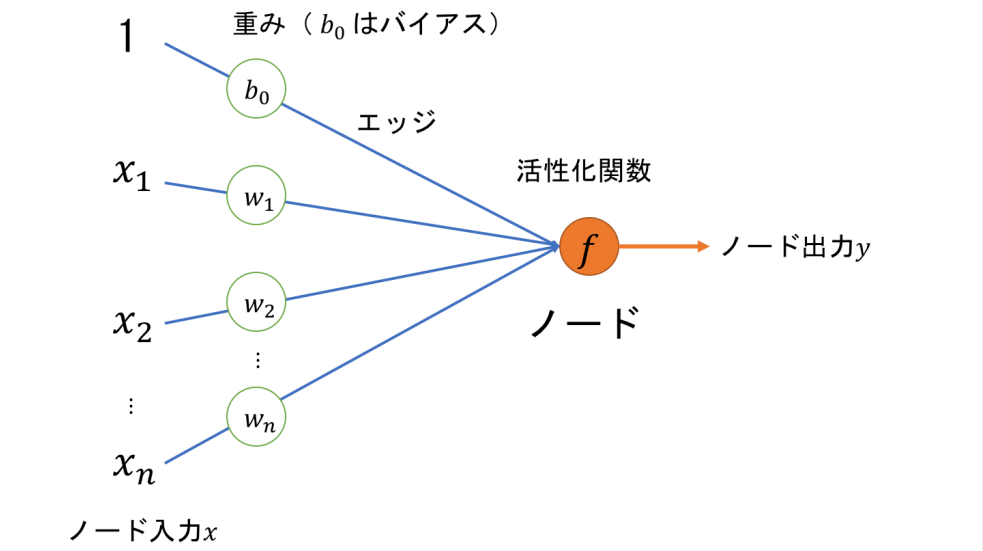

図 2 ノードの構造

中間層ノード（正確には出力層ノードも）はひとつひとつが活性化関数による識別器として機能する．ノードの構造を図2に示す．ノードに繋がる全てのエッジで，通過する数値データに重み（図2ではw）が積算され，それらの和にバイアス（図2ではb）を加えた値が，ノードへの入力となる．一連の流れを式にすると式(1)となる．ノードでは入力を活性化関数（図2ではf）によって識別し，-1から1（活性化関数によっては0以上）の範囲で識別結果を出力（図2ではy）する．活性化関数の主な種類はシグモイド関数，ハイパボリックタンジェント関数，ランプ関数（もしくはReLu関数とも呼ぶ）がある．それぞれの式と関数グラフを図3に示す．このノード出力は，エッジで繋がる次の層のノードへの入力となり，この結合を繰り返す事で，複雑なネットワークを形成する．

シグモイド関数，ハイパボリックタンジェント関数は，比較的単純な非線形関数であり，微分の計算が容易であったため，深層学習をはじめとする機械学習の識別器に使用されてきた．一方で，ランプ関数は2011年に前述の2つの活性化関数より学習の精度がよくなると発表された[39]関数であり，現在では，深層学習の中間層は主にランプ関数が使用されている．

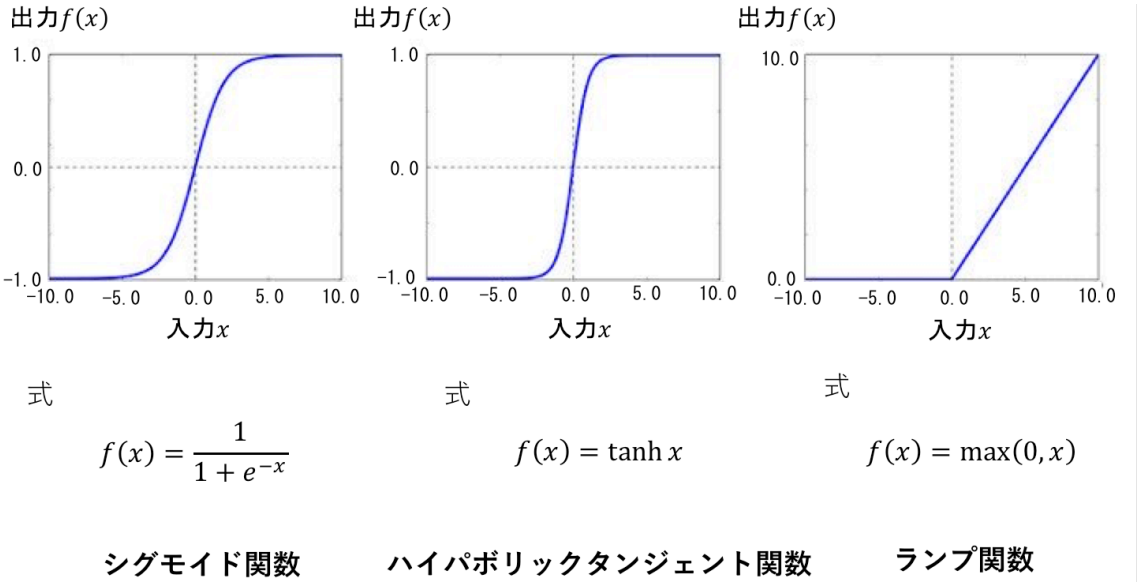

図 3　活性化関数

$$y = f\left(\sum_{i=1}^{n} x_i + b_0\right) \tag{1}$$

3.1.2　重みの更新

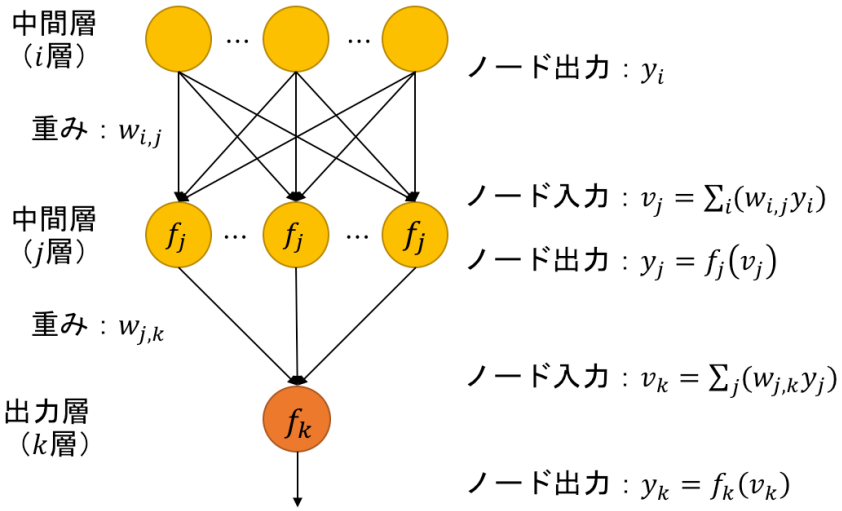

図 4　重みの更新

エッジに与えられる重み（とバイアス）は，深層学習の学習時に誤差逆伝播法[40][41]によって算出される．ここでは図4に示す，ディープニューラルネットワークの学習を例にして説明を行う（説明が複雑になるため，バイアスについては説

明から省く）．深層学習では，図4のように，まず入力データを入力層から出力層へ伝播させていき，出力y_kを算出する．この時，重みの初期値はランダムに与えられる．その後，出力y_kと理想の出力（分類なら入力データに対応するラベルを示す値）t_kとの差を算出する．この差を損失関数Eとし，$E = |y_k - t_k|$とする．続いて，損失関数に基づいて出力層側の重みの更新を行っていく．まず，$w_{j,k}$の更新式は式(2)となり，$\frac{\partial E}{\partial w_{j,k}}$は展開すると式(3)と表せる．$\delta_k$は勾配と呼ぶ．同様に$w_{i,j}$の更新式は式(4)となり，$\frac{\partial E}{\partial w_{i,j}}$は展開すると$\delta_k$を用いて式(5)と表せる．重みの更新は，損失関数$E$が0に近づく$w_{j,k}$，$w_{i,j}$を見つければ良いが，探しやすくするために勾配を利用する．

$$w_{j,k} = w_{j,k} - \frac{\partial E}{\partial w_{j,k}} \tag{2}$$

$$\frac{\partial E}{\partial w_{j,k}} = \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial v_k}\frac{\partial v_k}{\partial w_{j,k}} = f_k'(v_k)\frac{\partial E}{\partial y_k}\frac{\partial v_k}{\partial w_{j,k}} = \delta_k\frac{\partial v_k}{\partial w_{j,k}} \tag{3}$$

$$w_{i,j} = w_{i,j} - \frac{\partial E}{\partial w_{i,j}} \tag{4}$$

$$\frac{\partial E}{\partial w_{i,j}} = \sum_k \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial v_k}\frac{\partial v_k}{\partial y_j}\frac{\partial y_j}{\partial v_j}\frac{\partial v_j}{\partial w_{i,j}} = f_j'(v_j)\sum_k \delta_k w_{j,k}\frac{\partial v_j}{\partial w_{i,j}} = \delta_j\frac{\partial v_j}{\partial w_{i,j}} \tag{5}$$

式(3)と式(5)を整理すると，重み$w_{j,k}$，$w_{i,j}$に関する勾配はそれぞれ式(6)，式(7)となり，損失関数Eが0に近づく$w_{j,k}$，$w_{i,j}$を探すには，勾配が負なら重みを大きくし，勾配が正なら重みを小さくしていき，勾配が0に近づく値を探せば良い．ただし，式(5)のように，ある層の勾配を算出するには，それより下の層の出力が必要なため，重みの更新は出力層から始める必要がある．これを誤差逆伝播法と呼ぶ．誤差逆伝播法によって与えられた重みは，正しい出力を導くための，出力を特徴付ける入力が強調されるように値が大きくなるため，重みを辿ることは，分類基準に寄与する入力データを探すことになる．

$$\frac{\partial E}{\partial w_{j,k}} = \delta_k y_j \tag{6}$$

$$\frac{\partial E}{\partial w_{i,j}} = \delta_j y_i \tag{7}$$

3.2 DNNを用いたテキストベースの分類パターン解釈支援システムの構成

提案するテキスト分類パターン解釈支援システムの構成を図5に示す．まず，深層学習により学習されたテキストの分類パターンを表すネットワークを入力とする．次に，入力されたネットワークの中から，特定の出力に結びつく部分ネットワークをパスとして抽出し，パス上のノードにテキスト中の単語を用いたラベルづけを行う．その後，これらの抽出されたパスとラベルを可視化した分類パターンの解釈支援機能を，ユーザに提供する．ユーザは，可視化される重要パスとパス上のラベル，ラベル集合を要約したネットワーク，ラベルとなった単語が用いられている原文，などを参照しながら，分類パターンに解釈を与える．

深層学習による分類精度が高いことは広く知られているが，これは人間では網羅しきれない細かいパターンも含めて学習が行われていることが理由の 1 つと考えられる．すなわち，ある出力を導く分類パターンは細かいものを含めると無数にあり，そのすべての解釈を行うことはできない．そこで本論文では，分類パターンの中でも典型的な（適用されるデータ数が多い）パターンの解釈を促す．

また本研究では,「ある単語を含むと，ある出力に分類される」という分類パターンを解釈に利用することを仮定する．これは，テキスト中で用いられている単語情報と，分類先となる出力との関係を表す最もシンプルな表現として設定する．

3.3 学習ネットワークからの分類パターンの抽出処理

3.3.1 単語集合の抽出

深層学習で学習を行う前に，文章データ群は，まず，Bag of Words(BoW)[37]で各文章ごとに単語ベクトルに直される．BoWとは，全文書中に登場する単語を抽出して並べ，各単語の出現頻度をベクトルで表す表現である．BoWを用いた理由は，この手法では単語の並び順や出現場所が考慮されない代わりに，非常に簡単に文章の情報を表現でき，文章の情報を抽出する際によく用いられるためである．

なお，本研究では，単語抽出にはテキストマイニングのフリーソフトウェアであるKH coder[1]の単語抽出機能を用いてBoWの作成を行っている．また，KH coderは，形態素解析ツールの茶筌（IPADIC）の形態素解析の結果をほぼそのまま利用している．

[1]KH coder：http://khc.sourceforge.net

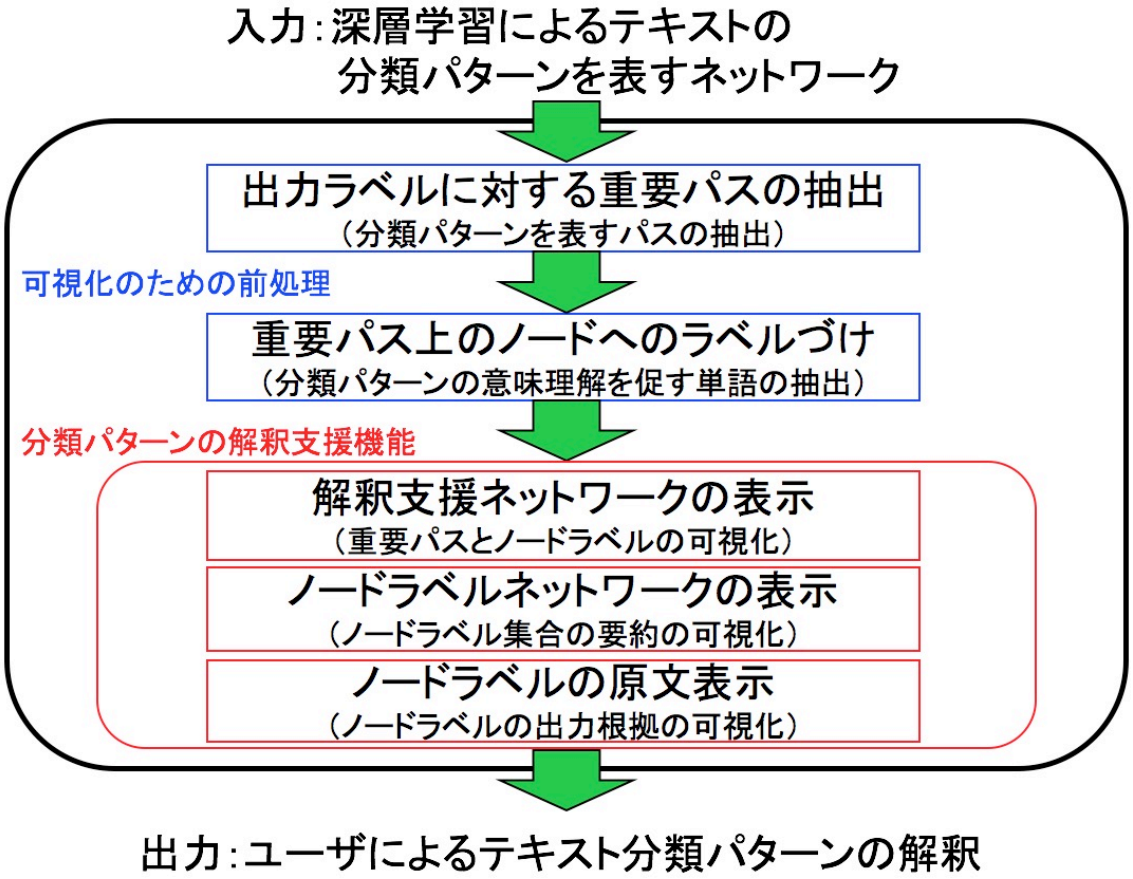

図 5　テキスト分類パターン解釈支援システムの構成

3.3.2　学習によるネットワークの重み付け

BoWによってベクトル化され，さらに文章ごとにラベル付けされた文章データは，DNNにて，それぞれの出力（ラベル）を導くネットワーク構造を構成するよう，重み付けがされていく．ノードの学習とエッジの重み付けについては，すでに3.1節で説明したため，ここでは省く．なお，今回の研究では，深層学習の実装には，深層学習ライブラリであるDeep Learning 4j[2]（DL4j）を使用している．また，中間層で使用した活性関数はtanh関数を用いているため，重みの値は正の値だけでなく負の値もとる．

3.3.3　出力ラベルに対する重要パスの抽出

深層学習により学習されたネットワークにおいて，ある出力ラベルXが指定された時に，出力層ノードXに到達する部分ネットワークの中から，分類に大きく

[2]Deep Learning 4j：https://deeplearning4j.org/index.html

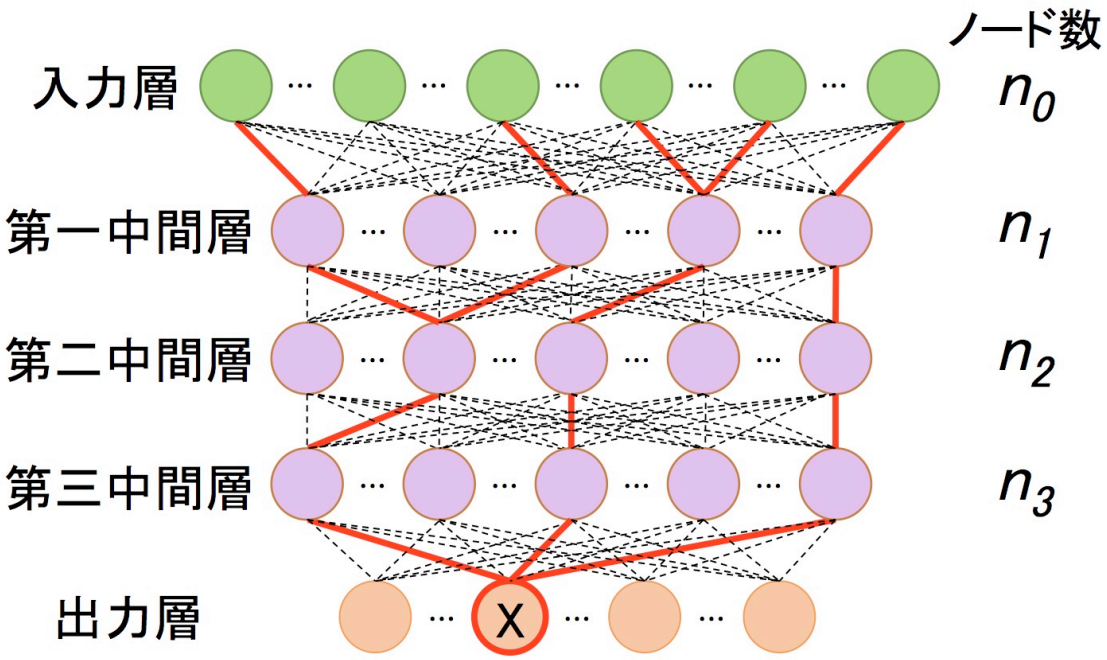

図 6 入力層ノードから出力層ノードXに到達する重要度の上位5本のパス抽出のイメージ（各エッジには学習した重みが与えられており，各パスの重要度はそれらの重みの積で与えられる）

寄与する重要パスを抽出する．ここで1つのパスを，「ある入力層ノードから出力層ノードXまで各中間層のノードを1つずつ辿った一本道のネットワーク」と定義し，これらのパス集合の中から重要度の高いパスを，別途指定する数（例えば5本）だけ抽出する（図6）．

すなわち，入力層のノード数をn_0，第一中間層，第二中間層，第三中間層のノード数をそれぞれn_1, n_2, n_3としたとき，入力層のノードから出力層ノードXに至る全部で$n_0 \times n_1 \times n_2 \times n_3$本のパスの中から，重要度の高いパスを取り出す操作を行う．

パス$path$の重要度$\mathrm{Imp}(path)$は，パスが含むエッジの重みの積として，式(8)で定義する．ただし$e(i,j)$は，ノードiからノードjへのエッジを，$W_{i,j}$はエッジ$e(i,j)$の重みを表す．

$$\mathrm{Imp}(path) = \prod_{e(i,j) \in path} W_{i,j} \tag{8}$$

エッジの重みには負の値を持つものもあるが，重要度は重みの積として計算されるため，最終的に重要度の正の値の大きい順に，重要なパスとして抽出する．す

表 1　重要パスの抽出に用いるエッジ数の上限

エッジ位置	エッジ数
入力層と第 1 中間層	2,000
第 1 中間層と第 2 中間層	300
第 2 中間層と第 3 中間層	300
第 3 中間層と出力層	100

なわち，ある出力ラベルXに至るパス上のエッジ$e(i,j)$の重み$W_{i,j}$は，あるノードiが表す単語を含むと出力Xに分類される場合に正の値に，あるノードiが表す単語を含むと出力Xに分類されない場合に負の値となり，ノードiが表す単語を利用することが出力Xに繋がるか否かが正負で表されている．そのため，重みが負の値のエッジが偶数個含まれているパスを重要パスの候補とする[3]．

この重要度を計算するためのPCの処理能力が高ければ，すべてのパスについての重要度を計算することができる．しかし，中間層の数やノード数が多くなるにつれて，実時間での計算が困難になるため，各層間ごとに重要パスの抽出に用いるエッジの数に制限を設ける．すなわち，各層間のエッジについて，その重みの絶対値が大きい方から順にこの上限値まで抽出して，パスを構成するエッジ候補とする．3.6節で後述する実験の際に設定したエッジ数の上限値を表1に示す．

このエッジ数の上限値は，利用する計算機の処理能力と合わせて，計算に費やせる時間を考慮して設定する．各層における上位のエッジのみを利用する理由は，パスの重要度が重みの積により計算されるため，途中のエッジの重みに0や0に近い値が入ると，結果的にパス全体の重要度が低く算出されると考えたことによる．

しかし，上限値が小さい場合など，上限値付近であっても重みが一定の値を持つ場合には，結果として重要パスが抽出されなくなるリスクも存在する．そのため，より有効なエッジの絞り込み方法については，今後さらなる検討が必要と考えている．

3.3.4　重要パス上のノードへのラベルづけ

本節では，前節の方法により抽出された重要パスの解釈を支援するために，重要パス上の中間層ノードの意味を表すラベルを付与する方法について述べる．

[3]否定の否定（二重否定）は肯定として捉えることができ，否定表現を重ねた日本語（たとえば,「今日はご飯を食べ[ない]わけでは[ない]ことも[ない]」のように，否定が3回出現すると最終的な意味も否定になる）と同様に，否定の数の偶数と奇数により最終的な意味が変わると考えられる．

重要パスが含む中間層のノードについて，各ノードと強いつながりのある入力層のノードが表す単語を，つながりの強い順に抽出し，抽出された単語集合を，各中間層ノードのラベルとする．各ノードと強いつながりのある入力層のノードの抽出には，前節で述べた重要パス抽出の方法を転用する．すなわち，中間層ノードmにおける入力層ノードwの重要度Label(w, m)を，wからmに至るすべての部分パス$kpath$の重要度の和として，式(9)で定義する．ただし式中のAllkpath (w, m)は，入力層ノードwから中間層ノードmに至るすべての部分パスの集合を表す．

$$\mathrm{Label}(w, m) = \sum_{kpath \in \mathrm{Allkpath}(w,m)} \mathrm{Imp}(kpath) \tag{9}$$

この重要度の値の高い順に，指定された数だけ入力層ノードの単語を取り出して，中間層ノードのラベルとする．なお本論文で取り扱った範囲においては，すべての部分パスを対象として計算することができたが，より大きなネットワークを対象とする場合には，前節と同様にあらかじめ重要エッジを選定した上で計算するなどの処理が必要になると考えている．

3.4 学習ネットワークからの分類パターンの可視化処理

3.4.1 解釈支援ネットワークの形成

本研究で開発した分類パターンの意味付け支援システムでは，意味付け支援ネットワークとして，各出力を導き出すネットワークと，ネットワーク上のノードの情報が表示される．このネットワークは，出力ノードから互いに関係が強い中間層ノードを上層（入力層側）へ辿っていき，注目している中間層上で，出力と結び付きの強いノード（と繋がるエッジ）を順番に表示したものである．この順番は重要度の値に基づいている．また，ノードの情報は，ノード上に単語集合として表示される．なお，注目している中間層，ノードの表示する個数については，別の章で説明する．

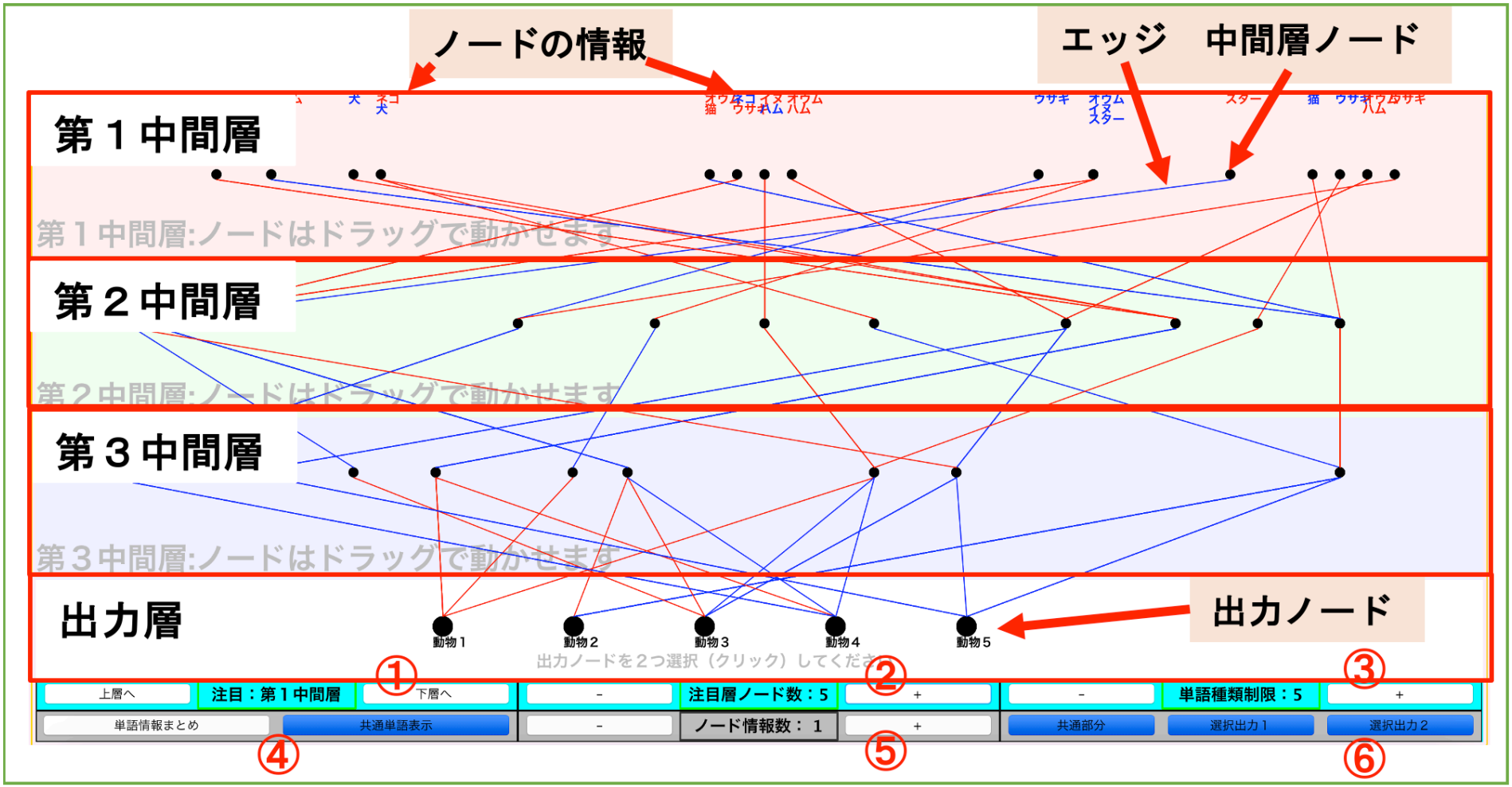

図 7　システムの画面（出力ノードを2つ選択する前）

例として中間層数が3つで，ラベル付けされた5種類の動物の生態に関する文章の分類を行った場合のシステムのメイン画面を，図7に示す．この時，注目している中間層は一番上層の第1中間層であり，注目中間層でのノードの表示数は各出力ごとに5つである．なお，本研究では，システムの実装ツールとして，TETDM[4]と呼ばれる総合開発環境ツールを利用している．よってシステムの表示画面は，全てTETDM上のものである．

図7では，入力層を除いて全ての中間層と出力層が表示され，出力層には分類されたそれぞれの出力を示す出力ノードが表示されている．また，中間層にはその出力と強く結びつくノードが表示されている．この時，出力ノードにマウスカーソルを重ねると，その出力を導くネットワークのみが表示される．また，出力ノードを2つ選択すると，選択したネットワーク同士を比較できるモードに移行するが，詳しくは後で述べる．その他，画面の下部には，ネットワークの表示を操作する操作ボタンが並んでおり，機能ごとに① から⑥に分かれている．次の項から，これらの機能について詳しく見ていく．

3.4.2 解釈対象出力の選択

本研究で開発したシステムでは，2つの出力を取り出してネットワークを比較し，各中間層のノード情報，共通のノードや単語，共通以外の単語などが画面上でわかるようになっている．この機能により，意味付けをしたい出力が，他の出力とどのような共通点，または特有の特徴があるかを，理解する助けとなる．

なお出力の選択は，システム画面上で出力ノードをクリックすれば良い．これにより，図8に示したように，ノードが黄緑色になり，その出力が選択状態となる．

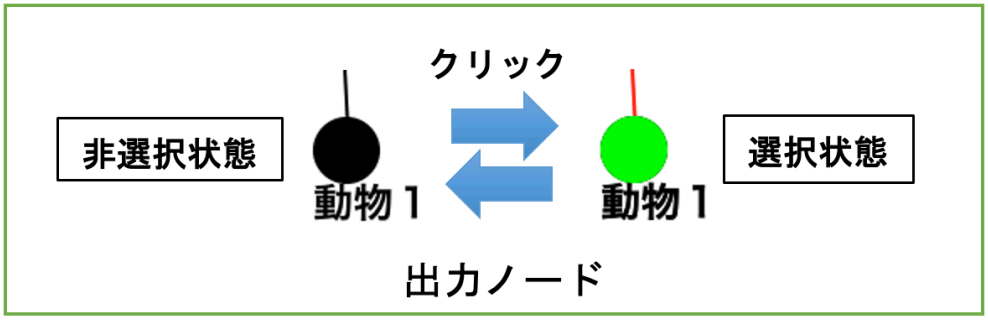

図 8 出力ノードの選択

[4]TETDM：http://tetdm.jp

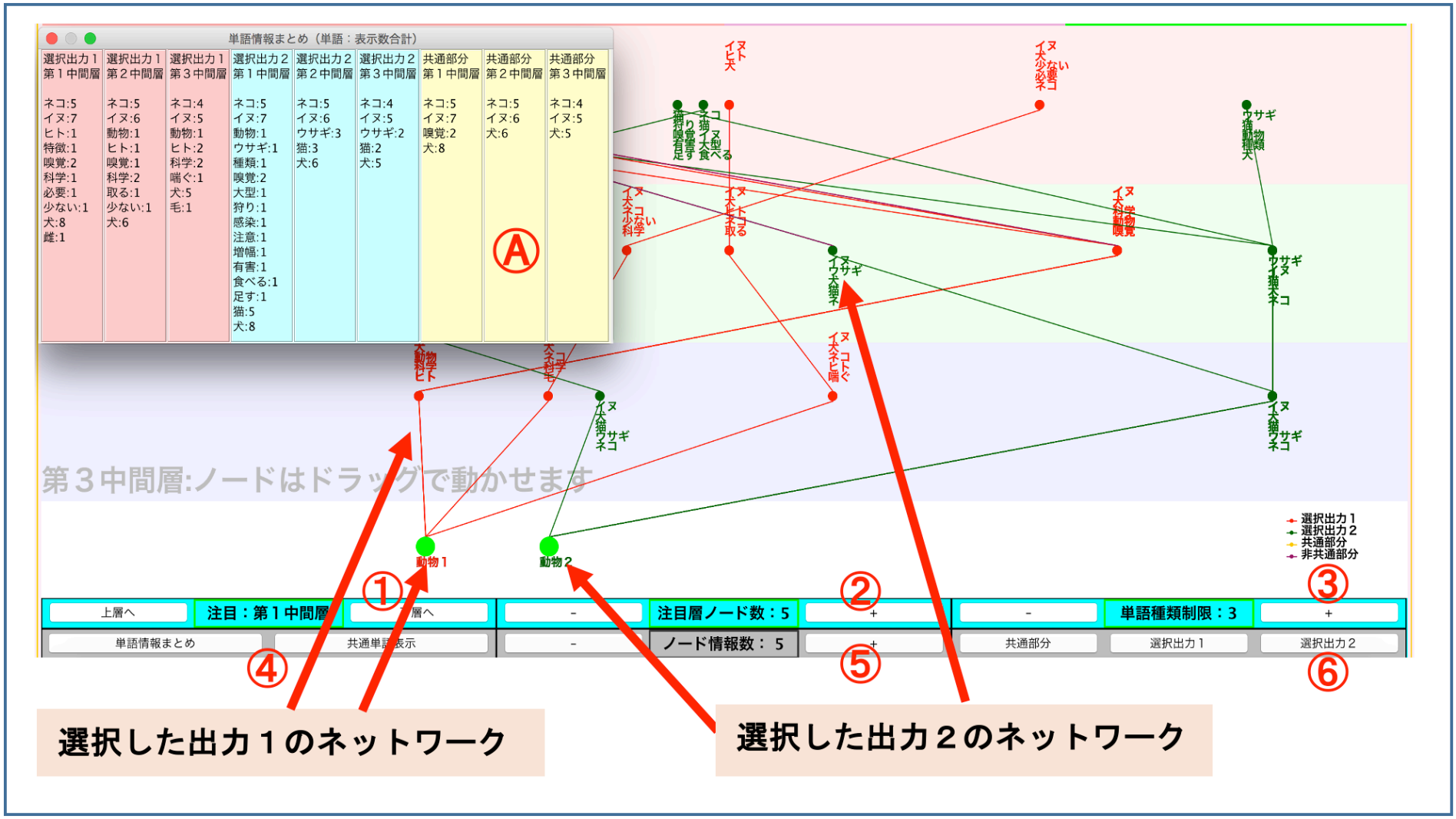

図 9　システムの画面（出力ノードを2つ選択した後）

出力ノードを2つ選択すると，ネットワークの表示画面では，各中間層のノード情報を含めて，選択した出力それぞれを導くネットワークと，共通のネットワーク，そして表示されている単語を集計した表が表示されるようになっている．その様子を図9に示す．図9では，選択した2つの出力の内，左側の方を選択出力1，右側の方を選択出力2としている．その上で，選択出力1のネットワークは赤色，選択出力2のネットワークは緑色で表示される．また，2つの出力に共通のネットワークはオレンジ色で，ネットワークが共通していても，ノード情報は共通していないものは紫色で表示される．単語を集計した表は左上のⒶの表で，後で説明する単語情報まとめ表示ボタンがオンの時に表示される．

3.4.3 解釈対象中間層の選択

ある層にて最も出力と結び付きの強い順にノードを表示したからといって，その層より下の層で表示されているノードが，その中間層内で重要度が大きい順に表示されているとは限らない．そこで注目する中間層を決めることで，注目中間層内でのノードが重要度順に表示される機能を設けた．この機能により，重要度の大きい順にノードを表示させたい中間層がある時，注目中間層をその中間層に選択すれば，出力との結び付きが強い順にノードを表示させることができる．また，注目している中間層より上の層は表示されなくなる．

この機能は図7の① の部分に相当し，「上層へ」と「下層へ」のボタンを押すことで，注目する中間層を変更することができる．図7の例では，中間層は入力層側から第1中間層，第2中間層，第3中間層となっており，注目する中間層は第1中間層となっている．この状態で注目中間層を下げると，それより上の層は表示されなくなり，注目中間層内で出力と結び付きが大きい順にノードが表示されるようになる．例として第3中間層に注目すると，図10に示すように第1中間層，第2中間層は表示されなくなり，さらに各ノードの真上には，新しくノードの情報を表す単語が表示される．

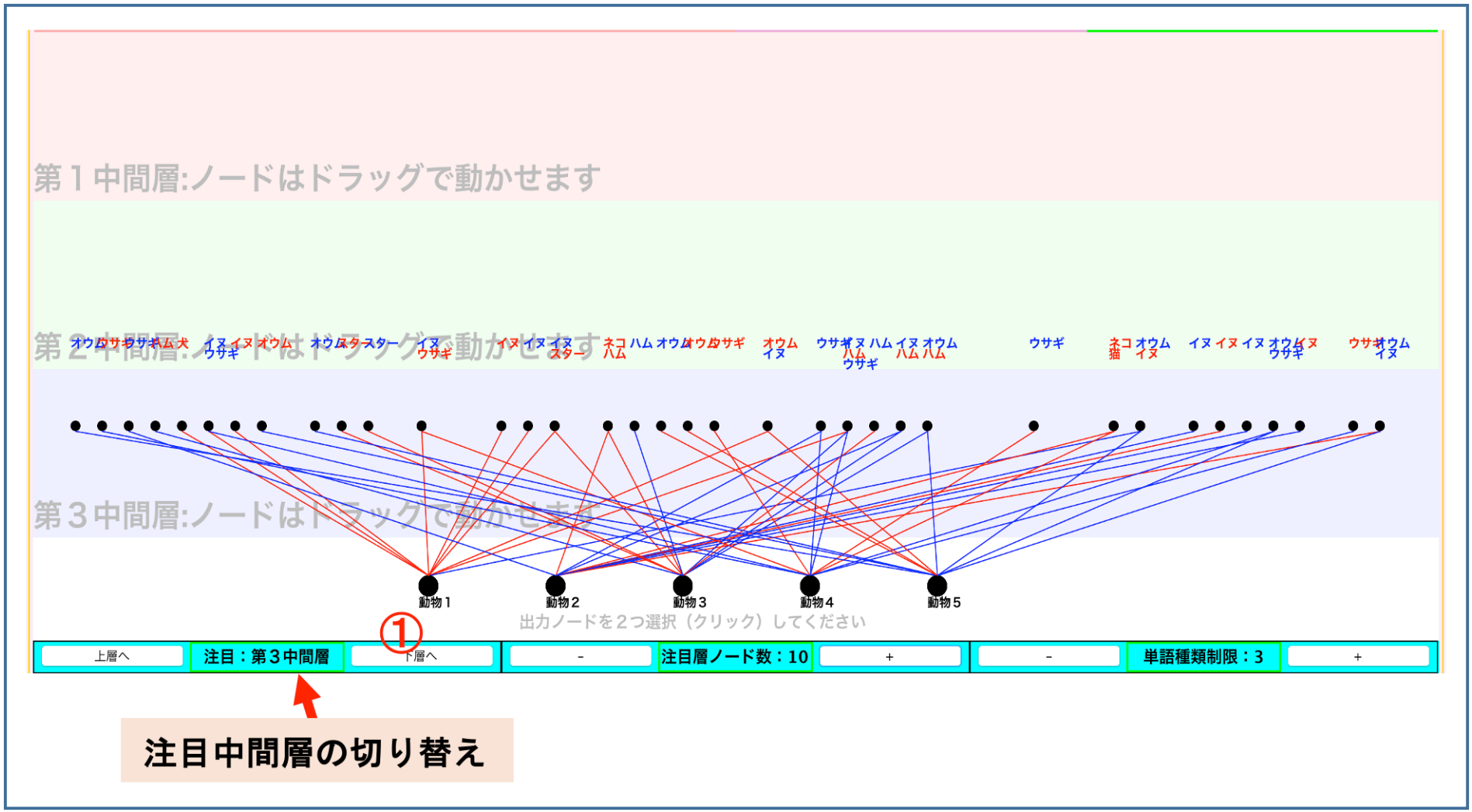

図 10　注目中間層の変更（第3中間層に注目した場合）

3.4.4　ノード数・ノードラベル数の変更

注目する中間層内で，出力との繋がりが強い順に表示されているノードの個数を調整できるように，注目層ノードの表示数選択機能を設けた．この機能により，表示数を1から順番に増やしていけば，注目層ごとに出力との繋がりが強いノードを順番に取得できる．また，1つのノードの情報から分類のルールを得るのはとても難しいため，表示ノード数を増やせば，より分類パターンの意味付けがしやすくなる．

この機能は，図7の②の部分に相当し，「＋」「—」ボタンを操作することで，現在注目している中間層の表示ノード数を変更することができる．この時，注目層ノード数の個数は，各出力ごとの値である．例として，注目層を第1中間層，注目層ノード数を5にして，ある1つの出力のネットワークを表示させたのが図11である．図11より，第1中間層のノード数を見れば，きちんと注目層ノード数と等しくなっていることがわかる．

ノードの情報は，そのままでは全ての出力に関わる情報が含まれる，したがって，特定の出力に結びつく情報がわかりづらい．そこで本研究では，各出力ごとに元の文章内で一定個数以上出現している単語のみ表示させている．各文章の単語の出現数は，深層学習への入力データ（単語の集合ベクトル）を利用して取得している．この機能を用いれば，頻出度が高い単語のみを表示でき，より文章の特徴を表している単語だけを残すことができる．しかし，結果的に表示される単語の種類が減るので，分類パターンの意味付けがしにくくなる可能性もある．

この機能は，図7の③の部分に相当し，「＋」「—」ボタンにより単語種類制限の数値を変更すると，各文章中にその個数以上出現している単語だけが表示されるようになる．

出力ノードを2つ選択した後は，ネットワーク上で各ノード上にノードの情報が単語として表示されている．ここで，ノード上の情報量，つまり単語の表示数を増やして，各ノードが持っている情報の意味を考察しやすくする機能を設けた．この機能によりノードの情報数を増やせば，ノードの持っている情報の意味を，単語単体ではなく単語を組み合わせた文章として考察でき，より分類パターンの意味付けをしやすくなるとした．

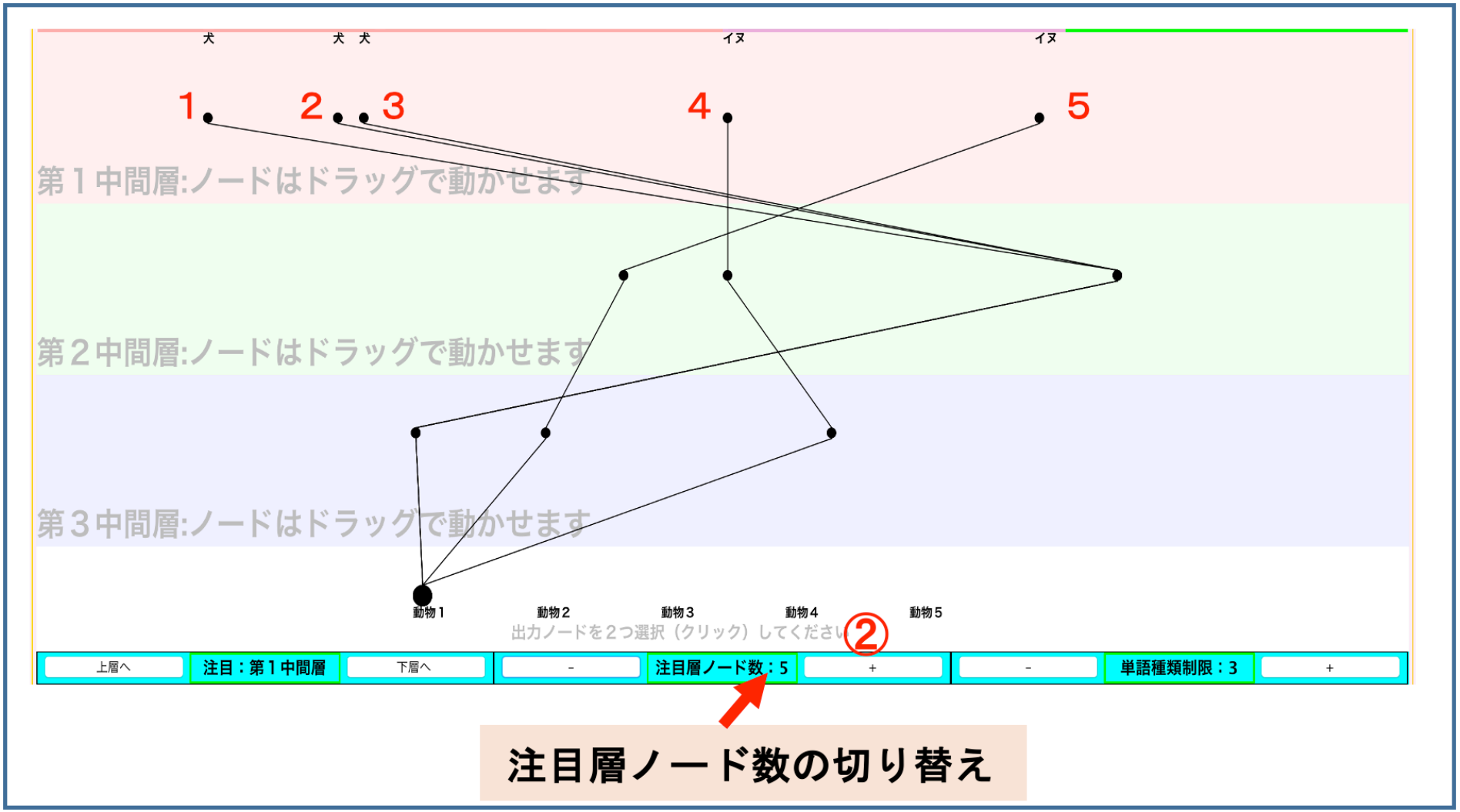

図 11　注目層ノード数を5に変更した場合（出力ノード「動物1」をマウスで触っている状態）

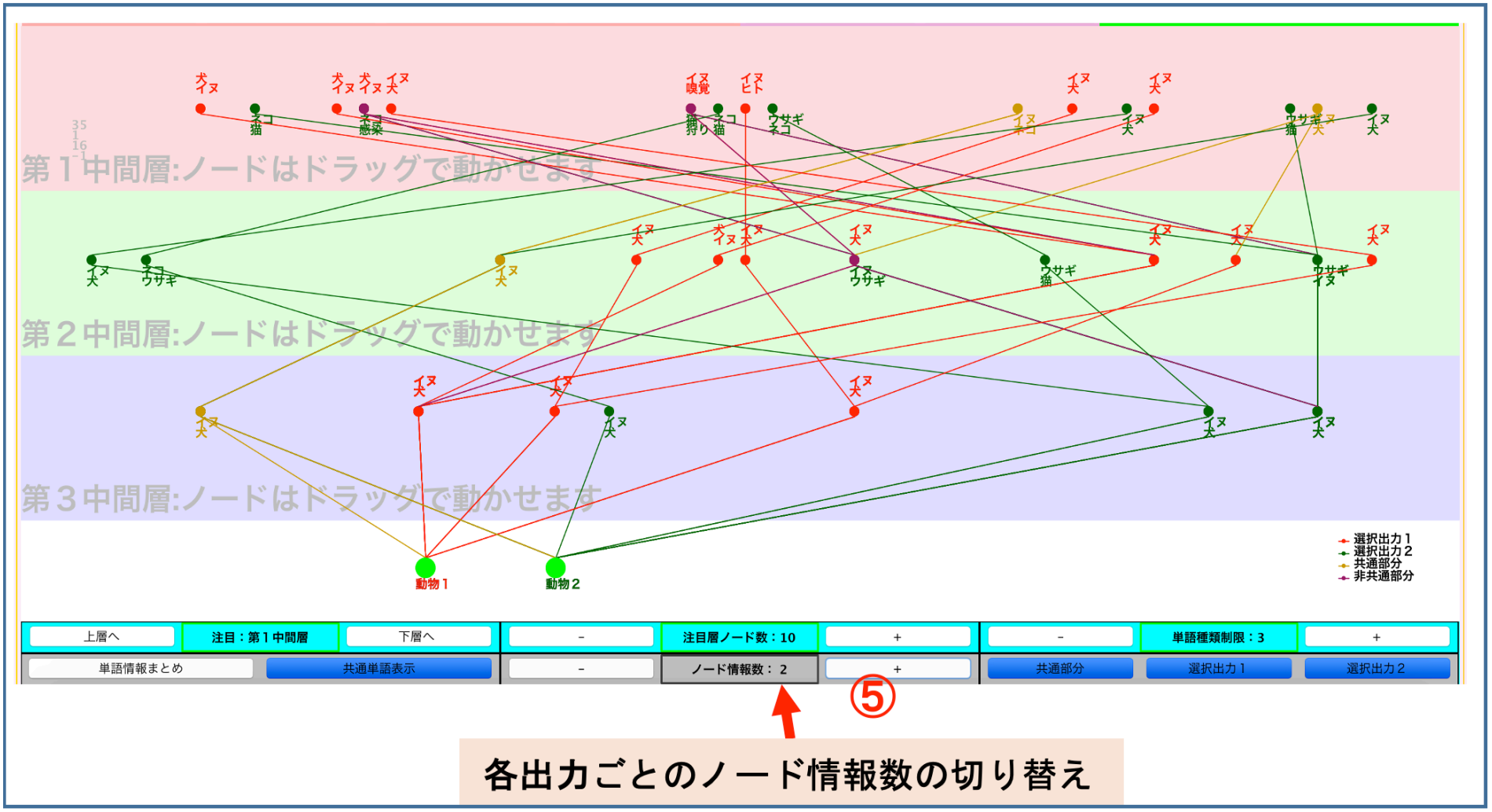

図 12　ノードの情報数の変更

この機能は図7の⑤の部分に相当し，出力ノードを2つ選択した状態で「＋」「—」ボタンを押すことで，各出力ごとのノードの情報数を変更できる．なお，1つのノード上に表示される単語は，重要度が大きい順に上から表示されていく．図12は，ノード情報数を2に変更した場合のネットワークを示している．図12より，選択出力1，選択出力2共に，中間層ノード上単語の表示数が2つずつになっていることがわかる．

3.4.5　原文表示機能

出力ノードを2つ選択した後は，全てのノード情報が表示され，さらに各ノードの情報数が増えてくると，ネットワーク中のノード情報を見ていくのが大変になる．そこで表示中の単語情報をまとめて表示する，単語情報まとめ表の表示機能を設けた．この単語情報のまとめ表は，表示されているネットワーク上の単語を各中間層ごとにまとめ，さらに選択出力ごとと，共通部分の3つの項目に分けて表示する．また，単語ごとの表示数の合計も合わせて表示される．なお，共通部分とは，共通ネットワーク上の単語に合わせて，選択出力それぞれのネットワーク上に共通で出現している単語も含む．その他，各選択出力の項目内から共通の単語の表示を消すこともできる．この機能により，表示されている単語が増加して見づらくなってきても，現在出現している単語を表から一目で確認できる．また，共通部分の項は比較したい出力との共通点を見つけるのに利用でき，共通単語の表示を消せば，それぞれの出力に特有の特徴を掴むのにも役立つ．

図13の左図に，単語情報のまとめ表の例を拡大したものを示す．

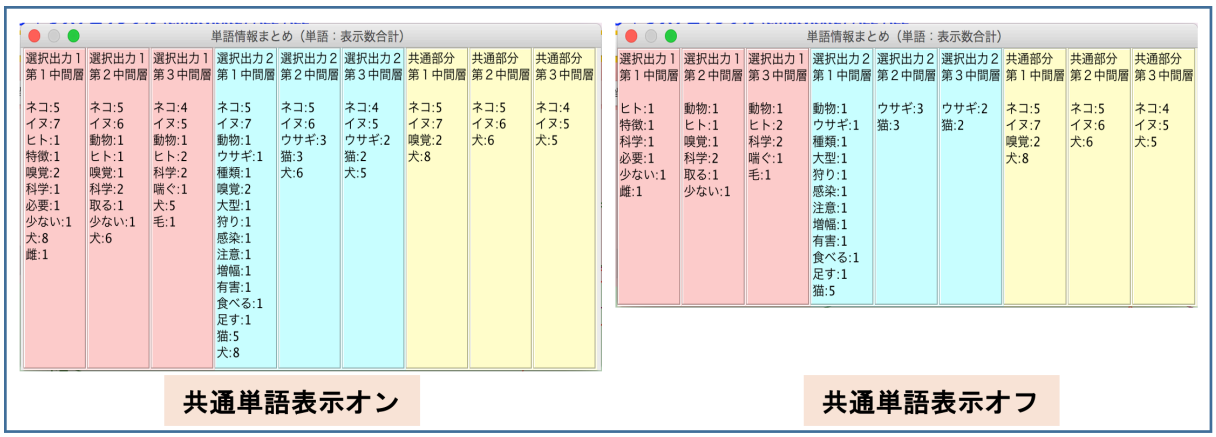

図 13　単語情報まとめ表

この機能は図7の④の部分に相当し，「単語情報まとめ表示」ボタンをオンにす

ると，単語情報まとめ表が表示される．また，各選択出力の項目から共通単語を消すには,「単語情報まとめ表示」ボタンの隣にある「共通単語表示」ボタンをオフにすれば良い.「共通単語表示」ボタンをオフにすると，表は図13の右図に示す状態になる．この状態では，表の選択出力1，選択出力2の項から，共通単語が除かれているのがわかる．

3.5　システムの使用例

システムの使用例として，5種類の動物「犬」,「猫」,「ハムスター」,「うさぎ」,「オウム」の生態や特性の説明文として書かれた文章データを用いて，2種類の動物の特徴，共通点などから分類パターンの意味付けを考察する手順を示す．深層学習に用いた学習ネットワークは，入力層のノード数が2,903個，中間層が3層でノード数は各50個，出力層のノード数が5個となっている．

利用者はまず，自分が分類パターンの意味付けを行いたい出力の出力ノードと，比較を行いたい出力の出力ノードの2つを選択する必要がある．ただし，① から⑥の機能は，出力ノードを2つ選択しているかいないかで，使用できるか出来ないかが決まっている．まず，出力ノードを2つ選択していない状態では，① から③の機能が使える代わりに，④ から⑥の機能が使えない．また，出力ノードを2つ選択している状態では，① から③の機能が使えない代わりに，④ から⑥の機能が使える．よって，最初は何も選択しないで，① から③の機能である，注目層中間層の選択，注目層ノードの表示数の選択，単語種類の選択を行った方が良い．

最初にシステムを立ち上げると，図14に示す画面が表示される．今回は5つの出力のうち,「ハムスター」と「オウム」の比較をし，この2つの出力に関して，大まかな特徴や共通の特徴，共通情報を除いた特有の特徴を考察していき，分類パターンの意味付けを行う．

まず，出力ノードを選択する前に，注目層の変更，注目層ノード数の変更，単語種類制限の変更を行う．今回は注目層，単語種類制限は初期状態のままにし，情報量を増やすために注目層ノード数を15まで増やした．この数値は20まで増やせるが，多すぎると画面が見にくくなるので注意する．また，出力ノード「ハムスター」と「オウム」をクリックして選択状態にする．図15に，2つの出力ノードを選択した後のシステムの画面を示す．これにより，図7の④ から⑥の機能が使えるようになった．

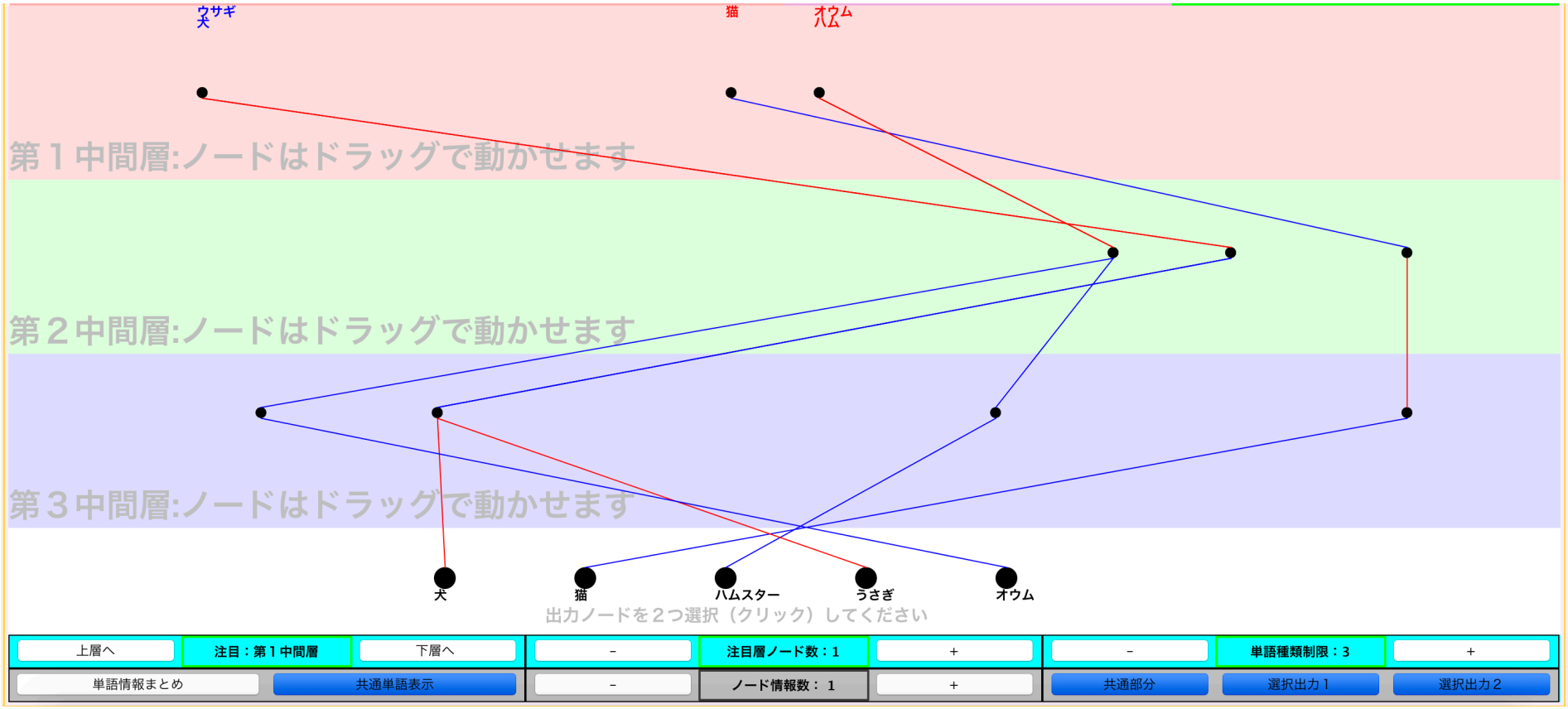
ウサギ
犬
猫
オウム
ハム
第１中間層:ノードはドラッグで動かせます
第２中間層:ノードはドラッグで動かせます
第３中間層:ノードはドラッグで動かせます
犬
猫
ハムスター
うさぎ
オウム
出力ノードを２つ選択（クリック）してください
上層へ
注目：第１中間層
下層へ
－
注目層ノード数：1
＋
－
単語種類制限：3
＋
単語情報まとめ
共通単語表示
－
ノード情報数： 1
＋
共通部分
選択出力１
選択出力２

図 14　操作開始画面

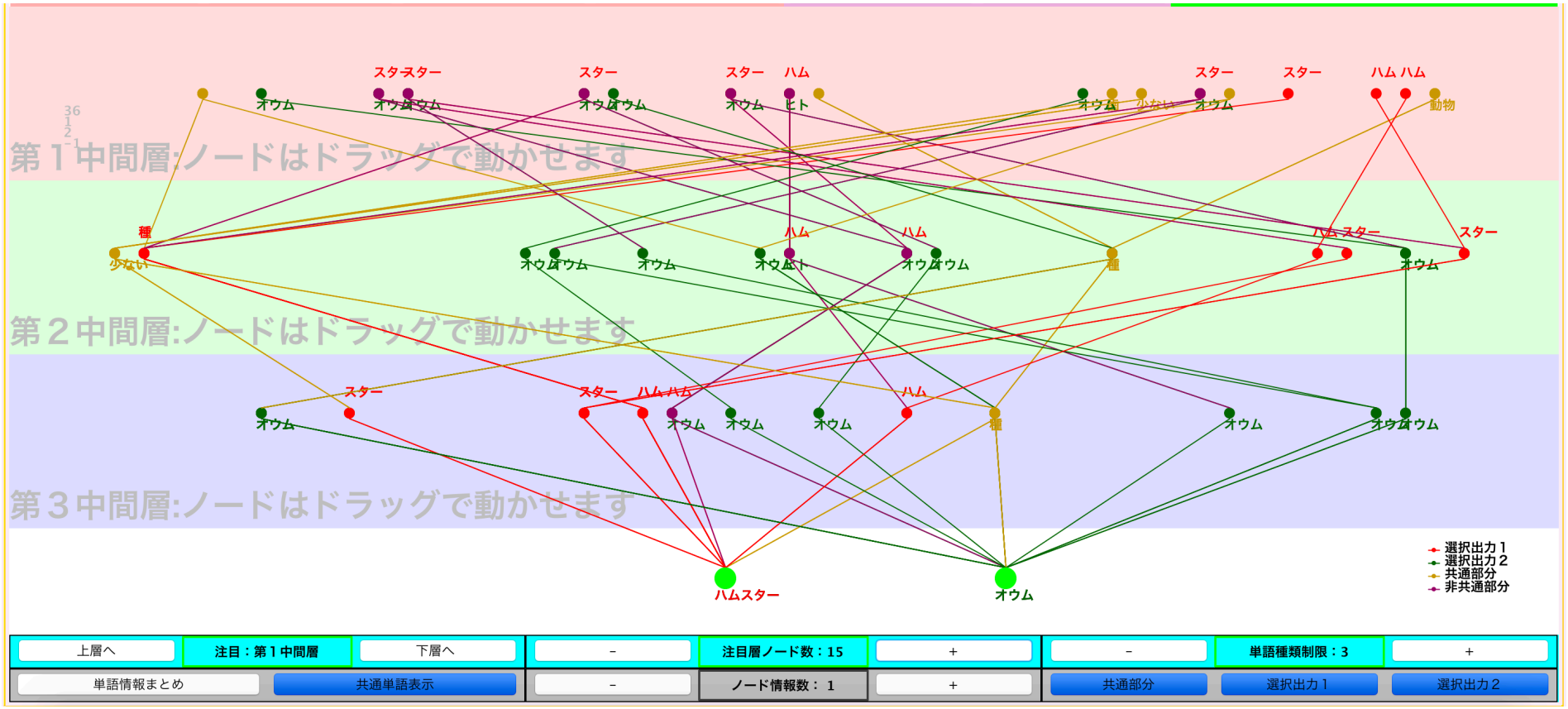

図 15　「ハムスター」,「オウム」選択後のシステム画面

なお，ノードの情報量を増やすために，ノード情報数を変更して10にした．また，最初は「ハムスター」についての分類される条件を調べてたいので，ネットワークの表示非表示を変更し，共通部分と選択出力1（出力「ハムスター」）のネットワークのみが表示されるようにした．ただし，そのままでは第1中間層のノード情報が隠れて見にくいので，ノードをドラッグして少し位置を下にずらした．その状態を図16に示す．

まず，図16の第1中間層のノード情報を見ると,「ハム」と「スター」が1つのノードに並んで表示されているのがわかるが，この文章はハムスターについて書かれているので，これ以外の何らかの文を考察できそうな単語を探す．そのまま第2中間層，第3中間層のノード情報を見ると,「穴」「動物」と言った組み合わせや,「歯」「食べる」などの組み合わせから，穴に住んでいる動物，特別な歯の使い方で物を食べる，などの特徴を表す文が考察できた．また,「個体」,「多い」から，個体数が多い，という特徴も推測できる．

続いて,「オウム」についての分類条件を調べるため，選択出力2（出力「オウム」）と共通部分だけを表示させる．その状態を図17に示す．図17の第1中間層の情報を見ると，オウム，インコと言った単語が並んでいる他，第2，第3中間層のノード情報を見ると,「絆」「飼い主」「ヒト」などの単語から，人によく懐く動物,「感性」という単語から，注意しなければならない感染症がある,「穴」「動物」「生息」などから，穴に住む動物である，といった内容の文が考察できた．

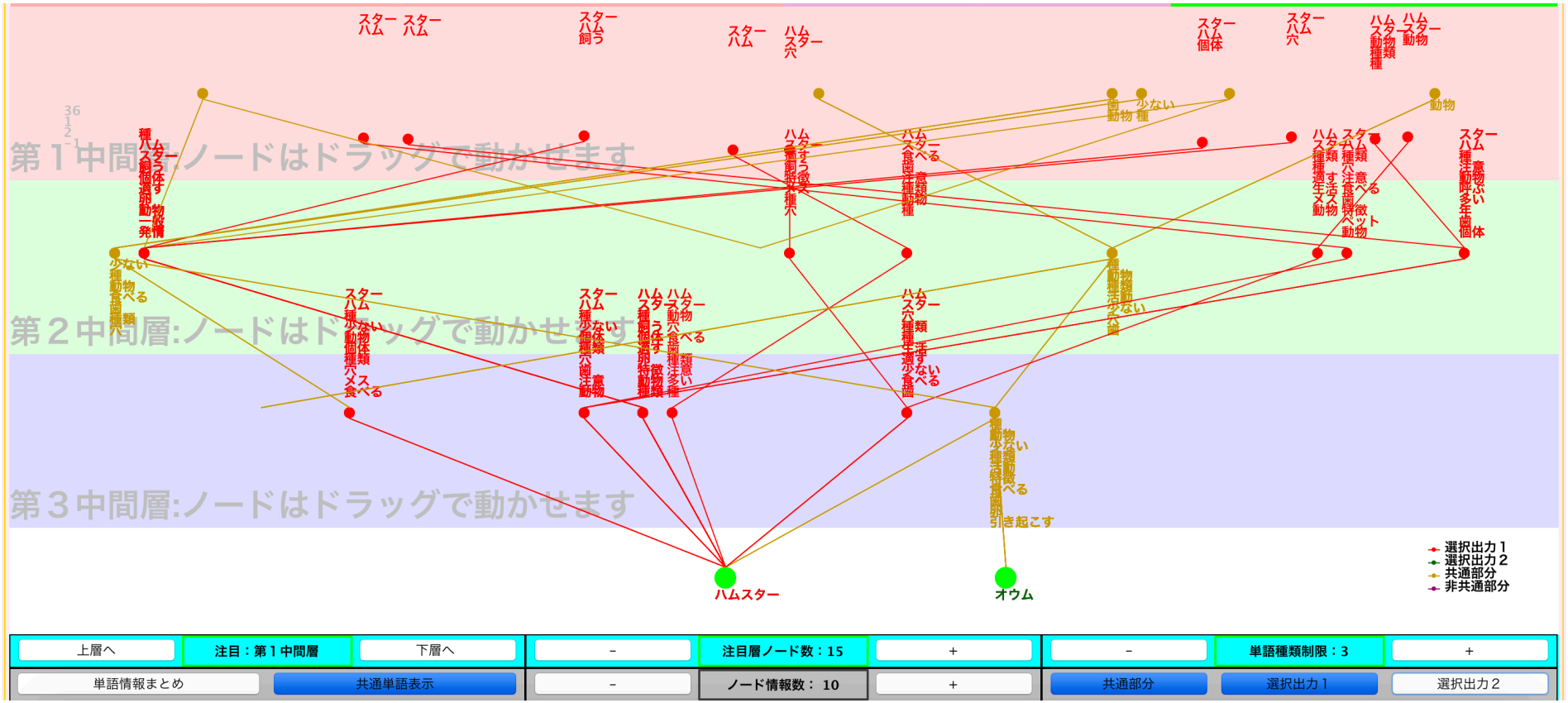

図 16　「ハムスター」が強調されたネットワークの表示

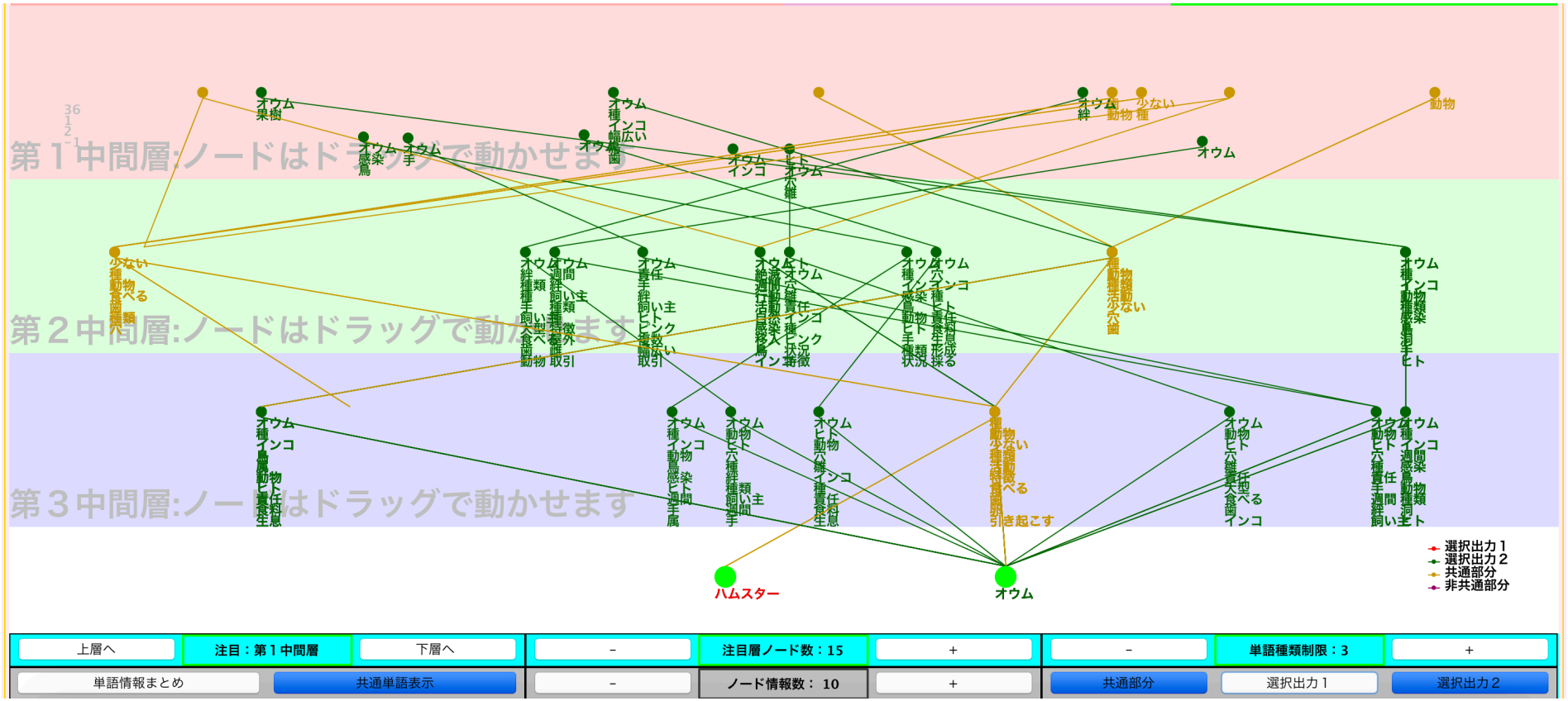

図 17　「オウム」が強調されたネットワークの表示

最後に，共通の情報とそれぞれの出力特有の情報を知るために，ネットワークの表示を全てオンにした状態で共通単語表示をオフにして，単語情報まとめ表を表示させる．その表を拡大したものを図18に示す．

図 18　共通単語表示をオフにした「ハムスター」と「オウム」の単語まとめ表

図18では，単語の表示数の多いものに注目していく．まず，共通の特徴を知るために黄色の共通部分の項目を見ると,「種類」「少ない」「穴」「歯」などの単語から，どちらの動物も穴に住んでいて，歯に特徴があり，種類が少ないといった文が考察できる．次に，それぞれの特有の特徴を見ると，選択出力1の項目では「個体」「多い」「ペット」「飼う」「適す」などから,「ハムスター」特有の特徴は，個体数が多く，ペットとして適しているという点が挙げられる．また，選択出力2の項目では,「感染」「飼い主」「絆」という単語から,「オウム」特有の特徴は，気をつけるべき感染症があり，飼い主にはよく懐くという点が挙げられる．

以上より，穴に住んでいて，歯に特徴があり，種類の少ない動物についての文章は「ハムスター」「オウム」に分類される可能性があり，その中でもペットに適していて，個体数が多い動物について書かれていたら「ハムスター」に分類され

る可能性が高く，もしくは危険な感性症があったり，飼い主に良く懐いたりする動物について書かれていたら「オウム」に分類される可能性が高くなる，という分類パターンの意味付けが行えた．これにより，例えば中身がハムスターについての文章を探したい時に，同システムで様々な文章の分類をさせた後で，穴や歯という単語が強調されている出力を絞り，さらにそこから，個体数の多さについてや，ペットとしての適正の良さを示すような単語が含まれている出力を探せば，中身がハムスターについて書かれている文章を探し出すことができる．

3.6 DNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムの有効性の検証実験

本章では，提案システムが，分類パターンの解釈に有効に用いられるかを確認するために行った実験について述べる．本実験では，理系の大学生，大学院生14名に，システムの利用方法について理解してもらった後，3種類のデータ分析課題を与え，深層学習により学習された分類パターンに意味づけを行ってもらった．被験者のほとんどは，深層学習について簡単な予備知識がある程度で，ネットワーク構造の検討や学習結果の解釈に携わったことはなく，データ分析の経験についても初心者と言える人が多かった．

3.6.1 使用テキストデータと学習モデル

表2に，本実験で用いた課題ごとの深層学習データを示す．学習データは，課題「動物」においては，50種類の動物において，動物名と「生態」というキーワードで検索エンジンによる検索結果の上位のWebページから，明らかに生態の説明としては不適切なものを除いて100件ずつ，合計5,000件のテキストを利用した．課題「映画」においては，映画レビューサイト[5]において，映画「シン・ゴジラ」の映画レビューの項目から，レビュー記事と評価値を1,000件（評価1と2を評価153件，評価3を119件，評価4を330件，評価5を398件）利用した．課題「ツイート」においては，キーワード「受験」で検索したTwitter[6]のツイートから，フォロワー数が100人以上のユーザのツイートを6,000件（「いいね」が0個,「いいね」が1から10個,「いいね」が11個以上をそれぞれ2,000件）を利用した．またデータの前処理として，各テキスト中の名詞，動詞，形容詞を抽出し，BoWでテキストを数値に変換している．

[5]映画.com (URL) http://eiga.com
[6]Twitter: (URL) http://twitter.com

表 2　課題ごとの深層学習データ

	動物	映画	ツイート
学習テキスト数	5,000	1,000	6,000
1テキストの文字数(平均)	400	280	100
入力層ノード	7,561	6,875	4,040
中間層ノード	50	40	30
出力層ノード	50	4	3
分類精度	92.7%	95.8%	90.1%
表示単語の文書頻度のしきい値	5% (5件)	5% (6件)	1% (20件)

学習は全結合型のディープニューラルネットワークによって行い，いずれの課題も中間層は3層とした．中間層のノード数は，分類精度が90%を下回らない範囲で，ノード数を減らす操作を行った．学習率は0.1，l1ノルム係数，l2ノルム係数はともに0.0001，学習回数は2,000で学習を行った．

表2の最下段に，重要パス上の入力層ノードとして選ばれるためのしきい値を示す．意味づけ支援として表示される単語は，このしきい値以上の数のテキストに出現する単語に限られる．

3.6.2　実験手順

実験の手順を以下に示す．

1. サンプルデータの分析（10分間）：システムの操作説明に目を通し，可能な操作を一通り試してもらう．
2. 課題１「動物」：動物の生態を説明するテキストの分類（30分間）：２種類の動物の生態の違いをまとめてもらう．
3. 課題２「映画」：映画「シン・ゴジラ」のレビュー記事の分類（30分間）：映画「シン・ゴジラ」の評価されている箇所と，評価されていない箇所を分析してもらう．
4. 課題３「ツイート」：受験にまつわるツイートの分類（30分間）：受験にまつわるツイートの中で，どのようなツイートが，より「いいね」をもらいやすいかを分析してもらう．

3つの課題のそれぞれについて，以下のステップで分類パターンに意味づけを行ってもらった．

1) 2種類の出力ラベルを選択する：「動物」課題では2種類の動物を被験者に選んでもらった．「映画」課題では評価が5のレビューと，評価が1または2のレビュー，「ツイート」課題では，「いいね」が11以上のツイートと，「いいね」が0のツイートとした．

2) 「単語情報まとめ」を表示させて，3つの中間層のそれぞれに表示されている単語を元に，2種類の出力のそれぞれにつながるルールを推定して，その意味と合わせて答えてもらう．

3) 「意味づけ支援ネットワーク」を表示させて，2種類の出力について，それぞれにつながる5つのパスとパス上の単語を元に，各出力につながるルールを推定して，その意味と合わせて答えてもらう．

4) ステップ2)とステップ3)で答えた内容を整理して，2種類の出力の相違点についてまとめてもらう．

3.6.3 結果と考察

被験者により記述された解釈の内訳を図19に示す．ただし，図に示す解釈の内訳は以下に定義する内容をもとに，著者が分類を行った．また，実験手順詳細のステップ2)から4)に対応する，中間層，パス，まとめごとに被験者平均を算出している．

- 表示単語：提案システムが表示する単語を使っており，その内容の正しさが原文から確認できる．
- 関連単語：表示単語は使われていないが，表示単語が使われている原文に含まれる単語を使っており，その正しさが原文から確認できる．
- 抽象解釈：表示単語，および表示単語が使われている原文に含まれる単語が使われていないが，その内容の正しさが原文から確認できる．
- 判定不能：解釈の正しさが判定できない．
- 不正解：解釈の内容が確実に誤っている．

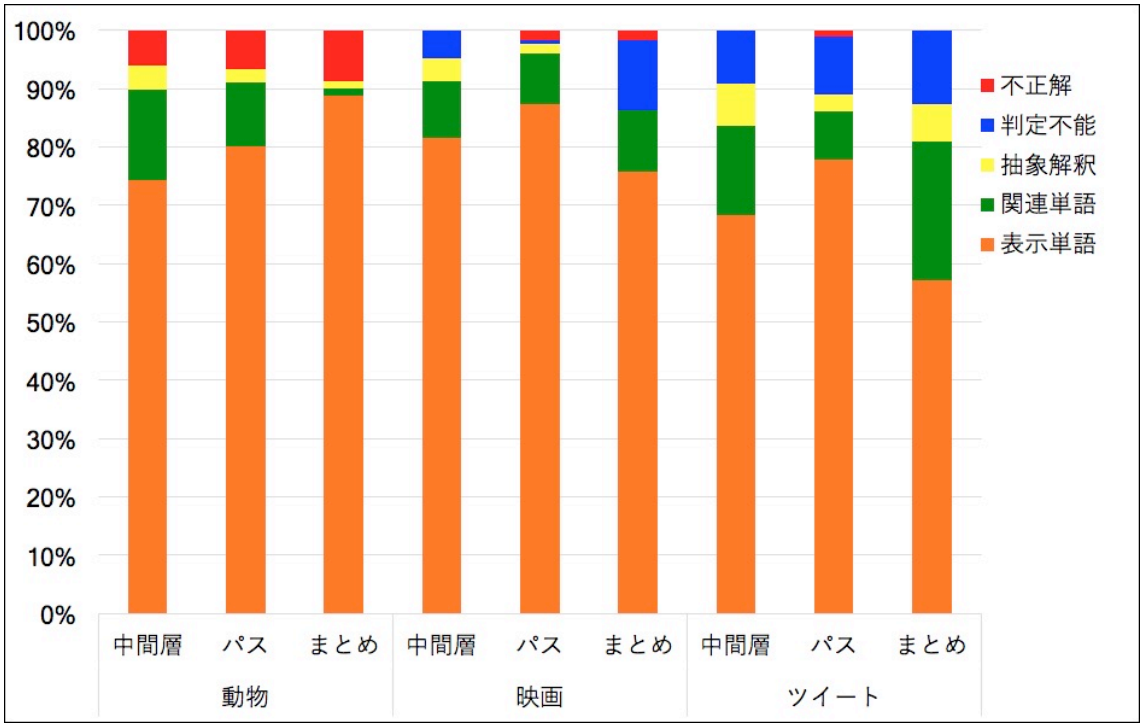

図 19　被験者に記述された解釈の内訳（被験者平均）

図19の結果から，多くの解釈は「不正解」と「判定不能」以外に分類され，その正しさが確認できており，実際に当てはまる事例を具体的な知識として捉えることができていた．このことから，中間層とパスの解釈において，本システムが出力する単語情報まとめと意味づけ支援ネットワークが，分類パターンの解釈に役立てられたと考えられる．また図19において,「表示単語」だけでなく「関連単語」や「抽象解釈」に分類される解釈も一定の割合で存在しており，システムが提示する単語を活用した解釈が行われていたことがわかる．

ほとんどの被験者が原文表示機能を用いることで，解釈を行っていた．また「不正解」に分類された解釈を書いた被験者は，多くの場合に原文表示機能を利用していなかった．このことから，テキストベースの深層学習における分類パターンの解釈においては，単語集合のみからの解釈は困難であり，原文との組み合わせで解釈していくことが有効かつ不可欠と考えられる．

課題ごとの比較として,「動物」の課題では不正解が多く,「映画」と「ツイート」の課題では判定不能が多くなった．これは動物の生態はその正しさが確認しやすかったのに対して，レビューやいいねの数などの人間の感覚に依存する評価においては，正しさの絶対的な基準がなかったことによると思われる．そのため，即座にその正しさを確認できない解釈について，その判断を支援する情報を提供で

表 3　「映画」のまとめ解釈の例と件数（○は高評価，×は低評価につながる解釈）

解釈の主旨	解釈例と件数
○音楽や映像などの演出	「音楽やCG表現により一定の評価を得ている」など8件
○映画が扱っているテーマ	「災害映画として学べるところがあった」など3件
×ストーリーやセリフ	「セリフが長い」など7件
×これまでのシリーズとの違い	「今までのゴジラを期待していた」など2件

表 4　「ツイート」のまとめ解釈の例と件数（○は「いいね」がある，×は「いいね」がないにつながる解釈）

解釈の主旨	解釈例と件数
○近況報告	「個人の近況報告についてのツイート」など6件
○人との関わり	「友達を誘うツイート」など4件
○自分の心境	「受験に対する不安」など3件
×勉強方法	「受験勉強を助けるための効率の良い勉強方法」など5件
×広告	「無料で学べる方法などのツイート」など6件

きれば，より効果的なシステムに改善できると考えられる．

また課題「映画」と「ツイート」においては，中間層とパスの解釈に比べ，まとめの解釈において表示単語の割合が少なくなり，判定不能に分類される解釈が多くなった．このことから，個別の中間層やパスの解釈をまとめる際に，一定の抽象化が行われ，表示単語そのもの以外の言葉による解釈が行われていたことがわかる．そのため，複数の解釈を集めてまとめることは，より汎用的な知識を得る上で一定の意味があると考えられるが，先の考察と同様に，その正しさを検証できる仕組みが必要と考えられる．

表3と表4に，課題「映画」と「ツイート」のまとめ解釈の例と，多かった解釈の主旨を示す．本実験における，14名の被験者のまとめ解釈は，解釈対象ごとにおよそ2,3種類にまとめることができ，データ分析の初心者が，お互いに類似する解釈にたどり着くことができたことには意味があると考えられる[7]．

これらの解釈をさらに著者がまとめると，課題「映画」においては,『音楽や映像などの演出，映画が扱っているテーマ，については高い評価が得られ，ストーリーやセリフ，これまでのシリーズとの違い，については低い評価が得られた』となる．また課題「ツイート」においては,『近況報告，人との関わり，自分の心境，のツイートについては「いいね」をもらいやすく，勉強方法や，広告，のツイートは「いいね」がもらえない』となる．特に直感的には「いいね」がもらえると

[7]課題「動物」においては，被験者に自由に動物を選択してもらったため，まとめる対象の動物がさまざまに異なっており，被験者間の解釈の類似性は確認できなかった．

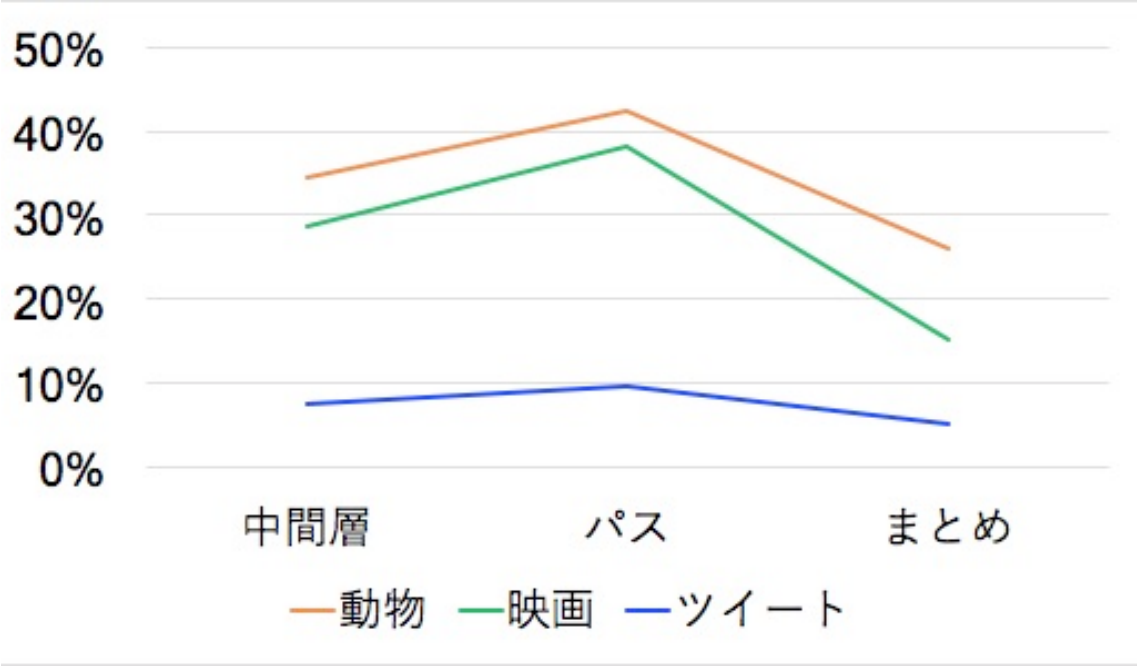

図 20　被験者の解釈が当てはまる原文の割合（被験者平均）

思われる勉強方法のツイートは，かえっていいねに結びつかないという興味深い結果も得られている．

実際の分析の場面で同様のまとめを得るためには，一人の分析者が主要な複数の解釈を列挙してまとめる道筋を用意することや，分析の初心者であっても複数人が分析を行うことで，一人のメタ分析者が結果をまとめあげる形式をとることで，有効な知見を得られる可能性もあると考えている．

被験者の解釈が当てはまる原文の割合を，図20に示す．ただし集計対象は，先の解釈の内訳で「表示単語」「関連単語」に分類されたもののみとし，それらの単語を含む原文の数を，各課題ごとに2種類の出力ラベルの学習に用いたテキスト数[8]で除した値を，解釈が当てはまる原文の割合としている[9]．ただし，図20の中間層の項目では，3つの中間層に対する解釈が当てはまる原文の合計を，パスの項目では，5つのパスに対する解釈が当てはまる原文の合計を，まとめの項目ではまとめとして書かれた1つの解釈が当てはまる原文の数を用いている．

まとめの解釈においては，課題「動物」「映画」「ツイート」の順に，26%(52件)，15%(84件)，5%(208件)の原文に当てはまっている．この数値には判定不能となった

[8]課題「動物」では200件,「映画」では551件,「ツイート」では4000件.

[9]解釈に用いられた単語が使われている原文においても，解釈が当てはまらない可能性も考慮して，一通り原文の内容を確認し，解釈の当てはまりに問題ないことを確認している．

表 5　被験者が解釈に用いた単語がTFIDF上位45単語に含まれる割合

課題と出力ラベル	単語種類数	単語総数
動物「犬」	40%(10/25)	53%(47/88)
映画「評価1・2」	56%(10/18)	63%(77/122)
映画「評価5」	26%(6/23)	32%(41/129)
ツイート「いいね0」	31%(4/13)	27%(29/106)
ツイート「いいね11+」	21%(4/19)	25%(33/134)

解釈の数値は含まれていないこと，また先述の複数人の解釈をまとめる方法をとることなどにより，この割合を増やしていける可能性があることから，本システムによって，学習データとなったテキスト集合に対して一定の解釈を与える支援を行うことができたと考えている．

課題「動物」「映画」「ツイート」の順に，まとめの解釈が原文に当てはまる割合が下がったのは，テキストが対象とする内容の範囲の広さによるものと考えられる．すなわち，課題「動物」に用いたテキストは特定の動物の生態についての話であり，内容には一定のまとまりがあると考えられるのに対して,「映画」は特定の映画に対するレビューであるものの，個人の主観に依存する部分があるため，内容に広がりが生じたと考えられる.「ツイート」は，受験をテーマにしている以外には制約がないため，個人の主観に加え，個人の置かれている状況や環境の違いがあるため，さらに内容の幅が広くなり，全体に当てはまる解釈を導くことは難しかったと考えられる．

最後に，被験者が解釈に用いた単語がTFIDF上位45単語[10] に含まれる割合を，表5に示す．TFIDFによっても，被験者が分類パターンの解釈に用いた単語を一定の割合で取り出すことができているが，その割合は必ずしも高くない．これは，DNNによる深層学習が単語の組み合わせを評価し，その組み合わせの一端を担う単語を，提案システムが表示しているためと考えられる．すなわち，TFIDFのような個々の単語を評価する手法によっては，分類パターンを解釈するための単語を十分に取り出すことはできないと考えられる．

[10]TFIDF値は，各出力ラベルが割り当てられたテキストの集合を1文書とみなし，DF値の計算に際しては，1文書中の5つ以上のテキストに単語が出現する場合に，文書頻度を1として計算した．また，各出力ラベルに対して，実験時の提案システムが出力する単語種類数の最大値が45種類（ここから中間層間で重複する単語がカットされる）となることから45位以内と比較した．

4. RNNの重みネットワークを用いたテキスト分類パターンの解釈支援

本章では，本研究で構築したRNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムについて，システムの構成，分類パターンの抽出手法，解釈支援機能，システムの評価実験について述べる．

前章では，本研究で構築する深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムについて，DNNを使用したシステムを構築し，評価実験では一定の成果が示された．続いて本章で使用する深層学習モデルは，中間層同士の繋がりが並列（簡単に言えばニューラルネットワークを横にいくつも並べ，各ニューラルネットワークの中間層同士を繋げたもの）の構造を持つ，RNN[42]である．RNNを用いるのは，RNNが時系列データの学習に適しており，文章中の単語の時系列順序を学習することで，より精度の高い分類が行え，本研究の提案手法を用いることで，単語の時系列順序の情報を持った分類パターンを抽出することが可能になると考えたためである．また，今回は学習済みのRNNの重みネットワークを一つのHMM(Hidden Markov Model)[44]として考え，時系列情報を含んだ分類パターンを抽出する手法を提案する．HMMを用いたのは，時間方向に展開したRNNの構造がHMMと類似しており，HMMが観測シンボルに対する尤度算出を行うように，RNNで単語のパターンに対する重要度を算出すれば，容易に分類パターンが抽出できると考えたからである．なお，HMMとRNNを組合せて新しいモデルを構築する研究[45]もあるが，本章ではあくまで既存のRNNモデルの学習ネットワークの解釈を行うために，HMMの構造を利用しているだけであり，新規の深層学習モデルを構築しているわけではない．

4.1 RNNモデルの構造

RNNの構造（ここでは簡略化のため，各層を構成するノードを1とする）を図21にを示す．RNNの特徴として，入力がタイムステップごとに変化することが挙げられる．これは，例えば株価の変動数値のように，時間によって値が変化するデータの学習を行うことができることを表す．RNNの基本的構成は，中間層が1層（1層以上の場合もある）のディープニューラルネットワークに，一時記憶層を追加した構造となっている．一時記憶層の役割は，中間層出力を記憶し，次のタイムステップ時に，入力とともに中間層へ入力する処理を行う．この処理を再帰的に

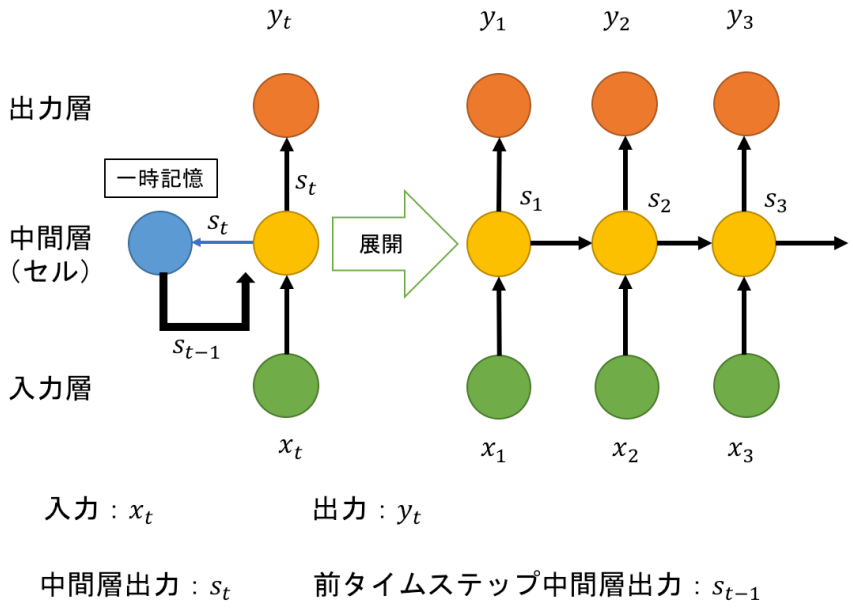

図 21　RNN

繰り返すため，中間層出力s_tは式(15)となり，出力y_tは式(16)と表せる．なお，$\boldsymbol{U}$は入力層と中間層の間の重み行列を，$\boldsymbol{W}$は一時記憶層と中間層の間の重み行列を，そして$\boldsymbol{V}$は中間層と出力層の間の重み行列を示している．また，fは中間層の活性化関数を，gは出力層の活性化関数を示している．図21の右の展開図は，この再帰的処理をタイムステップごとに並べていった場合のモデル構造である．なお，誤差逆伝播法は中間層から一時記憶層方向へも誤差が伝播する．図21の展開図を例にすれば，ちょうど矢印を逆に辿っていくように，誤差伝播を行って重みを更新していく．

$$s_t = f(\boldsymbol{U}x_t + \boldsymbol{W}s_{t-1}) \tag{10}$$

$$y_t = g(\boldsymbol{V}s_t) \tag{11}$$

4.2 RNNを用いたテキストベースの分類パターン解釈支援システムの構成

提案システムの構成を図22に示す．提案システムでは，最初に正解ラベルを付与したテキスト集合を学習データとし，RNNにて分類する学習を行う．なお，今回の実験においては，学習時の精度の向上を目的としてRNNの発展系であるLSTMモデルを使用している．次に，学習済みの重み付きネットワークをHMMに変換し，学習に用いたテキスト集合（原文）に出現した単語出現パターンの尤度を算出する．最後に，尤度の高い単語出現パターンを分類パターンとしてインタフェースに表示し，ユーザがその分類パターンの解釈を行う．この時，ユーザは分類パターンをいくつ表示するか任意で設定できる．また，システムでは分類パターンの意味を理解しやすくするために，ユーザが原文の内容を参照できる原文表示機能を備えている．

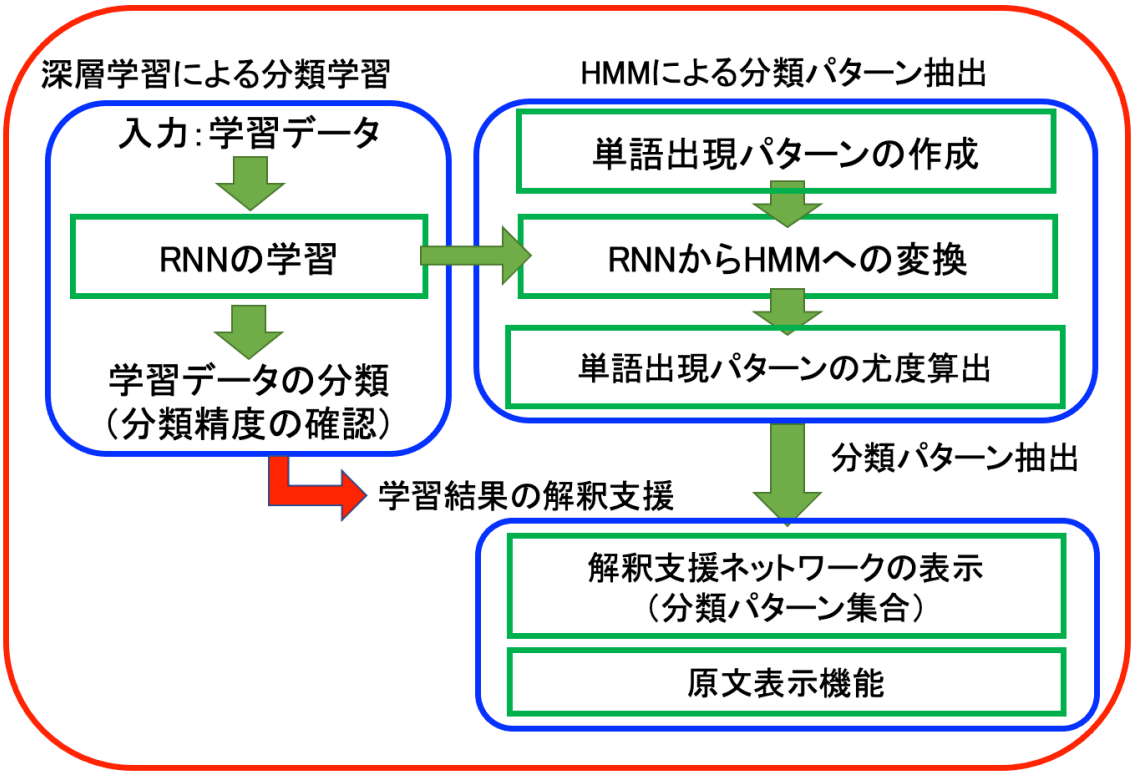

図 22 システムの構成

4.3　深層学習による学習ネットワークの形成

4.3.1　文中の単語のベクトル化

RNNで学習を行う前に，テキストデータは文中の単語を抽出してあと，One hot法[43]と呼ばれる手法に従い単語ベクトルの羅列に直される．One hot法とは，まず，全文書中に登場する単語の総種類数をベクトルの全要素数とする．そして，単語の種類ごとに1要素だけ値を1にし，残りの値を0とした単語ベクトルを作成する．この時，1にする要素は単語種類ごとに異なるようにする．あとは，文中の各単語を単語ベクトルに置き換え，入力データとする．例として「犬　の　散歩」「猫　の　散歩」という文があった場合，単語の総種類数は「犬，の，散歩，猫」の4種類であるため，単語ベクトルの要素数は4となる．そして，単語の出現順に単語ベクトルを「犬」=[1,0,0,0],「の」=[0,1,0,0],「散歩」=[0,0,1,0],「猫」=[0,0,0,1]と決める．よって，最初の文は「[1,0,0,0]，[0,1,0,0]，[0,0,1,0]」となり，2つ目の文は「[0,0,0,1]，[0,1,0,0]，[0,0,1,0]」と表現することができる．One hot法を用いた理由は，この手法では単語を容易にベクトル化でき，さらに単語の種類ごとに1となる要素が異なるため，4.4.2項で述べる，重要ノードのノード情報の決定が容易になる点が挙げられる．

なお，本研究では，単語抽出にはテキストマイニングのフリーソフトウェアであるTETDMの単語抽出機能を用いて単語のベクトル化を行っている．また，TETDMは，形態素解析ツールのIgoの形態素解析の結果を利用している．

4.3.2　学習によるネットワークの重み付け

One hot法によって単語ベクトルの羅列に変換され，さらにラベル付け（分類先を表す番号を付与）されたテキストデータは，RNNにて，それぞれの分類先（ラベル）を導くネットワーク構造を構成するよう，重み付けがされていく．その様子を図23に示す．入力文章が「犬　の　散歩（[1,0,0,0]，[0,1,0,0]，[0,0,1,0]）」であり，分類先がA，Bの場合，タイムステップ1では「犬」を表す[1,0,0,0]が入力層に入力され，学習によってエッジに重みがついていく．その学習された情報は一時的に記憶され，タイムステップ2で「の」を表す[0,1,0,0]が入力層に入力されると同時に中間層へ入力される．そして，それらの学習された情報はタイムステップ3で「散歩」を示す[0,0,1,0]が入力されると同時に中間層へ入力され，全て単語が入力されたことにより，最終的な分類先が出力される．図23のように，RNNでの

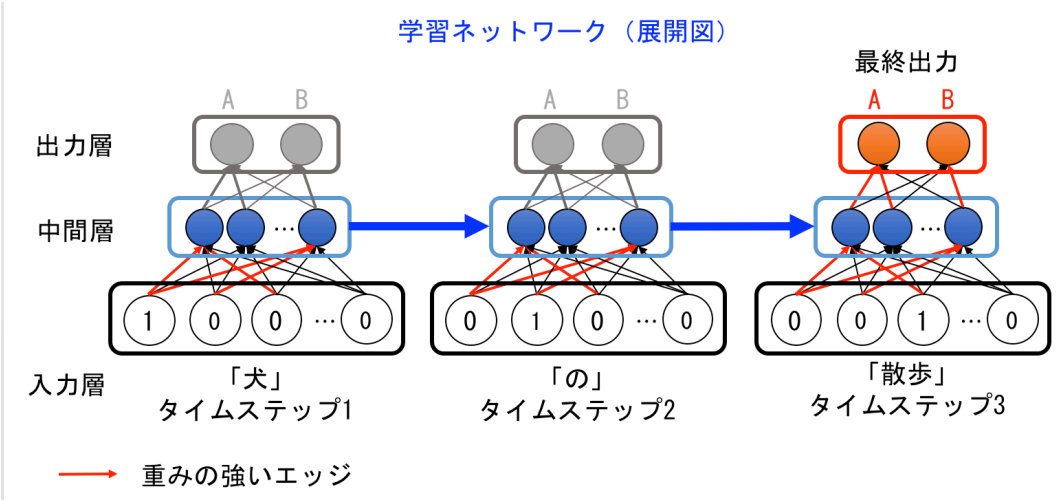

図 23　RNNの学習ネットワークと学習の様子

分類時は，最後の単語が入力されたタイミングで，出力層から出力される．

なお，今回の研究では，RNNの実装には，DNNの時と同じくDL4jを使用している．

4.4　学習ネットワークからの分類パターンの抽出処理

4.4.1　単語出現パターンの作成

システムでは，分類パターンを実際の文章に近づけることで解釈性を向上させるため，HMMへ与える観測系列（詳細は節4.4.2参照）として，RNNの学習に使用したテキスト集合に実際に出現した単語の出現パターンを使用する．この時，使用する単語の条件として以下を満たす単語出現パターン全てを候補とする．なお，単語出現パターンの長さ（単語数）はユーザが任意に設定できる．ただし，パターンごとに個別に設定することはできず，全てのパターンの長さは同一に設定される．

- 単語出現パターン中の単語は，原文中の名詞・動詞・形容詞（実験時は形容詞を省く場合もある）とする
- 単語出現パターン中の単語は，単語が出現している文章数（文章頻度）が1%以上となる単語のみとする
- 単語出現パターン中の単語の順序は実際の原文中の単語の出現順序に基づくものとする

このような条件を設けたのは，まず，本研究では，分類パターンの中でも典型的な（適用されるデータ数が多い）パターンの解釈を促すことを目指している．そのため，助詞や使用頻度の低い単語から成る分類パターンを抽出しても，典型的なパターンの解釈に繋げることは難しい．また，原文に出現しない単語の並びのパターンを解釈した場合，間違った解釈になる可能性が存在する．そこで，単語出現パターンとして使用する単語は名詞，動詞，形容詞とし，さらに文章頻度で一定の閾値以上の単語のみを用いる．そして，単語出現パターンの単語の順序も，実際の単語の時系列に基づくものとした．

4.4.2 重みネットワークのHMMへの変換手法

HMMとは，状態と観測シンボル（出力）の二過程で，状態が確率的に推移する場合，確率的に観測シンボルを出力する非決定性の有限状態オートマトンモデルとされる．HMMでは，観測シンボルの変化に対して，それがどの程度もっともらしいかという値である，尤度を算出できる．一方で，RNNについて，各層間の重みを確率分布とし，再帰処理による中間層ノードの発火の移り変わりを時間による状態の変化と捉えれば，RNNの構造を図24に示すように，ひとつのHMMとして扱うことができるようになる．そこで，RNNをHMMとして扱うことで，ある単語の時系列パターンを有する分類パターンについて，RNNの分類にどの程度寄与するかを，尤度として算出できるのではないかと考えた．

提案システムでは，単語出現パターンの尤度推定において，RNNの学習で得られた重み付きネットワークをHMMに変換し，単語の時系列パターンに対する尤度を算出する．変換方法として，まず，RNNの入力層ノードをHMMの観測シンボル集合，中間層ノードを状態集合$\boldsymbol{S}$とする．同様に中間層の（再帰的処理による）時系列間の重み集合$\boldsymbol{W_r}$を状態遷移確率$\boldsymbol{A}$，入力層中間層間の重み集合$\boldsymbol{W_i}$をシンボル出力確率$\boldsymbol{B}$とする．最後に，中間層出力層間の重み集合$\boldsymbol{W_o}$を初期状態確率$\boldsymbol{\pi}$とする（この時，$\boldsymbol{\pi}$はその時選択する分類先によって変わる）．

ただし，RNNの重み集合は確率の条件を満たしていない．そこで，ある中間層間の重みベクトル$\boldsymbol{w}$を構成している重み $w_i (1 \leq i \leq N)$について（Nは重みベクトル$\boldsymbol{w}$の要素数），重みが負の値であれば0にしたw_i'を用いて（式(12)），重みの合計が1になるように正規化した値w_i''を用いる（式(13)）．

$$w_i' = \max\{0, w_i\} \tag{12}$$

$$w_i'' = \frac{w_i'}{\sum_{s=1}^{N} w_s'} \tag{13}$$

以上より，RNNの各層間の重み集合について，重みベクトルの各重みを，式(??)を用いて合計が1になるように正規化することで，RNNの重み集合をHMMの状態遷移確率やシンボル出力確率として扱う．

以上より，RNNの各層間の重み集合をHMMの各確率とみなすことで，RNNからHMMの変換が可能となる．

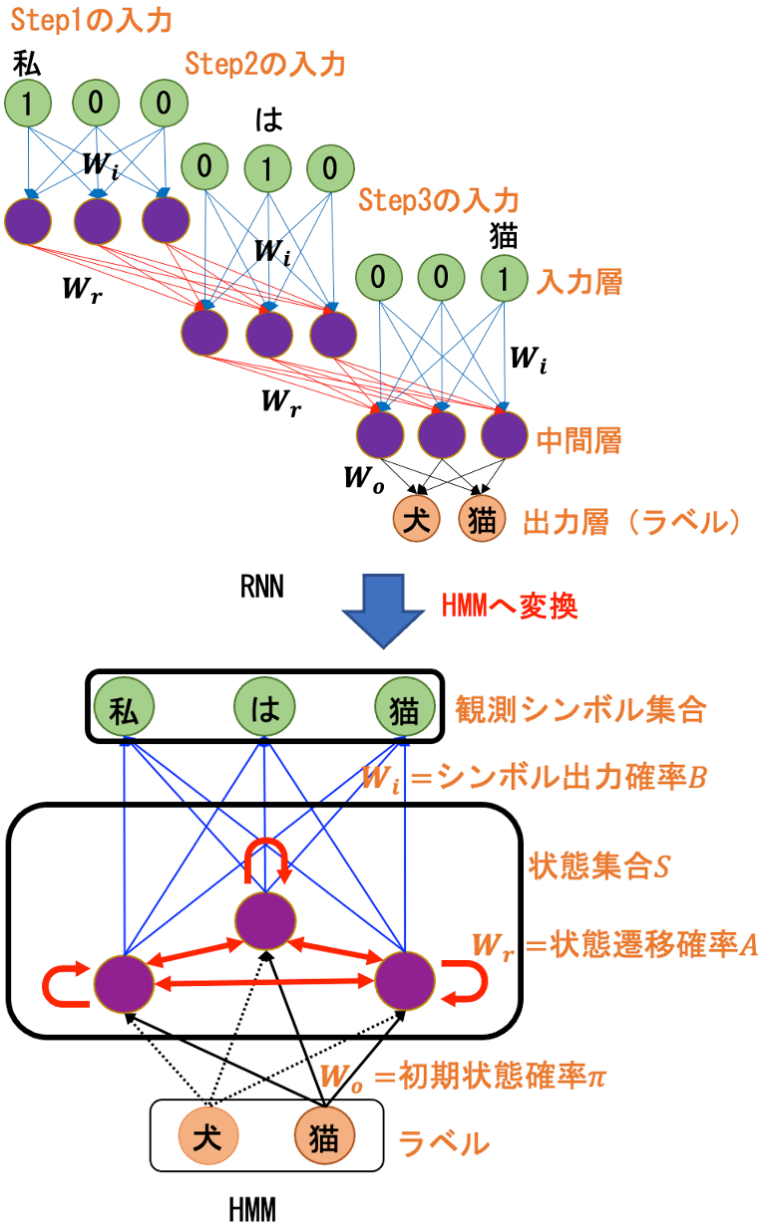

図 24　RNNとHMMの相互関係

4.4.3 単語出現パターンの尤度算出

4.4.1節で作成した単語出現パターン集合について，尤度の算出方法を述べる．4.4.2節でHMMに変換したRNNの重み付きネットワークに対して，観測シンボルによる観測系列（前述した単語出現パターン）を$\boldsymbol{O} = \{o_1, o_2, ..., o_T\}$（$T$は観測系列の長さ，つまり単語出現パターンの長さ），状態数（中間層ノード数）をN（状態番号はi, j）と置くと，状態遷移確率$\boldsymbol{A}$は式(14)，シンボル出力確率$\boldsymbol{B}$は式(15)，初期状態確率$\boldsymbol{\pi}$は式(16)となる．

$$\boldsymbol{A} = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\} (1 \leq i, j \leq N) \tag{14}$$

$$\boldsymbol{B} = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} \\ (1 \leq i, j \leq N, 1 \leq t \leq T) \tag{15}$$

$$\boldsymbol{\pi} = \{\pi_i | \pi_i = P(s_0 = i)\} (1 \leq i \leq N) \tag{16}$$

この時，ある分類先xに対して，単語出現パターン$\boldsymbol{O}$が存在する時，初期状態確率を$\boldsymbol{\pi}_x$と表すと，尤度$P(O|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B})$は，式(17)で算出される．

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) &= \sum_{all \boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) P(\boldsymbol{O}|\boldsymbol{S}, \boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) \\ &= \sum_{all s_0 ... s_T} \pi_{x s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\ &\qquad ... \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \end{aligned} \tag{17}$$

最後に，全ての単語出現パターンについて式(17)で尤度を算出し，尤度の高い順に，単語出現パターンを分類に寄与する分類パターンとして抽出する．

4.5　学習ネットワークからの分類パターンの可視化処理

4.5.1　解釈支援ネットワークの形成

解釈支援機能では，前節で抽出した，分類先に強く結びつく分類パターンの集合を，解釈支援ネットワークとして表示する．解釈支援ネットワークでは，分類パターン上の単語と単語間の時系列関係を理解しやすくするために，単語はオレンジ色のノード，単語の時系列関係はノード間の青い矢印として表示する．この時，尤度の大きさを矢印の太さで表す．さらに，双方向に矢印が存在するノード同士は，時系列関係が薄いとみなし，緑色のエリアにひとつのグループとして表示する．また，どの分類パターンがどの分類先に属しているのかを表すため，分類先名を表示した紫色のノードと分類パターンの最後の単語ノードを繋ぐ赤い矢印を表示する．

例として，5種類の和菓子の作り方に関するテキスト集合[11]から表示される解釈支援ネットワークを図25に示す．ユーザーは初めに，インタフェース下部の，解釈を行いたい分類先名（ここでは「饅頭･大福」）が表示されているノードを選択する．システムは，まず，選択した分類先名について，ユーザの任意の数で，尤度順に分類パターンを抽出する．抽出した分類パターンを表6に示す．次に，抽出した分類パターンの単語をノード，単語間の時系列関係を矢印とした解釈支援ネットワークが表示される．最後に，ユーザは解釈支援ネットワークを眺めることで，どのような単語や単語の時系列関係が，選択した分類先に寄与しているのかパターンを見つけ，解釈を行う．

表 6　抽出分類パターン例

尤度順位	抽出分類パターン
1位	「生クリーム」→「冷凍」→「片栗粉」
2位	「生クリーム」→「片栗粉」→「冷凍」
3位	「イチゴ」→「白あん」→「片栗粉」
4位	「刷毛」→「片栗粉」→「冷凍」
5位	「チン」→「白あん」→「イチゴ」

ここで，解釈支援ネットワークで表示されている選択した分類先の分類パターンは，あくまでその他の分類先と比較した場合の特徴を示しており，世間一般的

[11] クックパッド(URL:https://cookpad.com)から収集

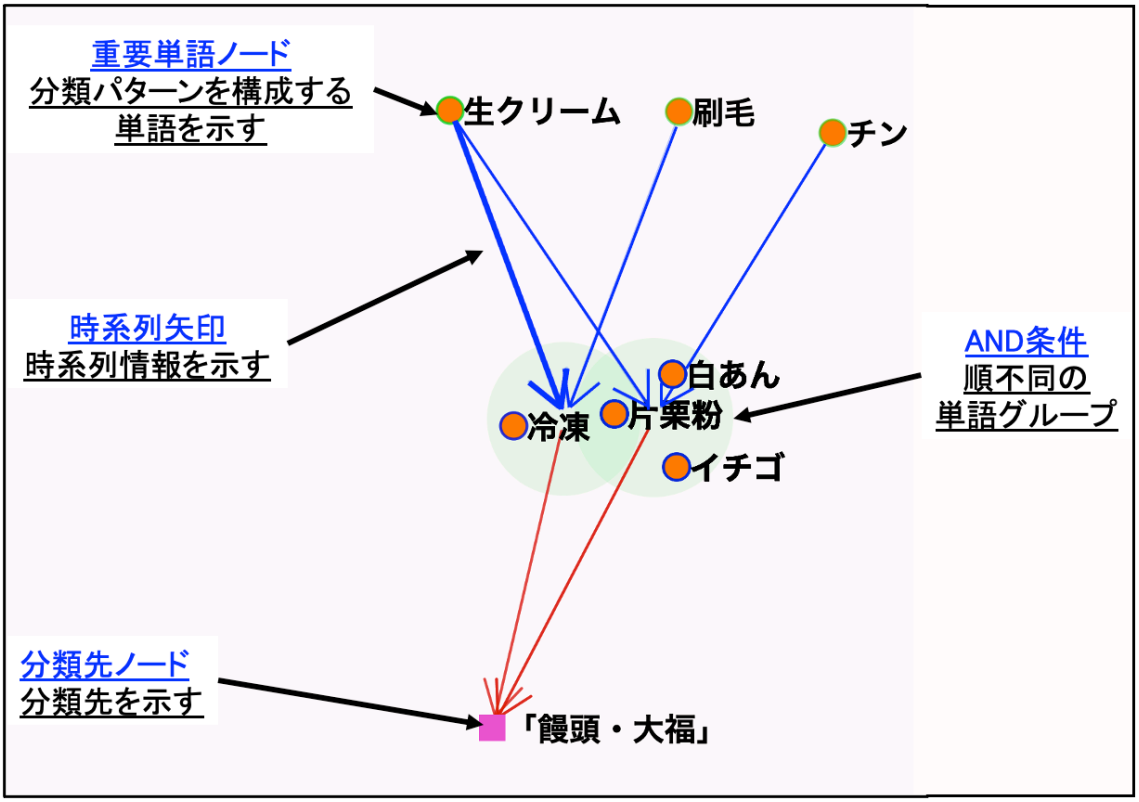

図 25　システムの画面例

な特徴が表示されているわけではないことを注意しておく．例えば，図25の例では,「生クリーム」→「冷凍･片栗粉」というパターンを見つけた場合,「生クリームを冷凍した片栗粉と混ぜる」もしくは,「生クリームを片栗粉と混ぜて冷凍する」等の作り方が，他の4つの和菓子と違う点であると解釈する．

4.5.2　原文表示機能

分類パターンの解釈に向けて，単語情報だけでは，その単語が実際にどのような文脈で使われていたのかを把握することは難しい．そのため，原文表示機能により，分類パターン中の単語群が，実際に学習に用いたテキスト内でどのように使われているかを表示する．

ユーザは，解釈支援ネットワーク上で単語を選択することで，原文の中でその単語を含む文章が表示され，参照することができる．ただし，見やすさを考慮して，表示されるのは1つの文章につき，選択した単語の前10単語，後10単語までの区間とした．また，単語は最大2種類まで選択でき，その場合は単語間の文章は全て表示される．図26に5種類の和菓子の作り方について分類先「饅頭･大福」の

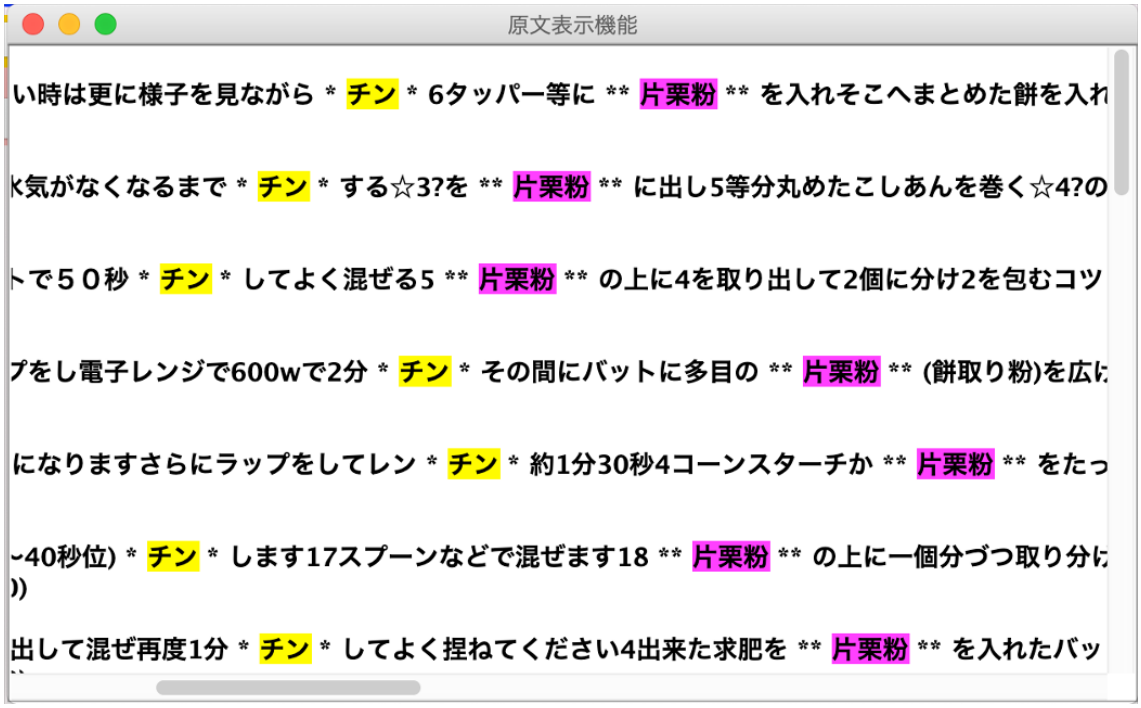

図 26　原文表示例（「饅頭・大福」についてのテキストに対して単語「チン」と「片栗粉」を選択）

テキストを使用し，単語「チン」と「片栗粉」を順番に選択した時の分類パターンの原文表示例を示す．

4.6　システムの使用例

システムの使用例として，5種類のお菓子「マカロン」,「ティラミス」,「チーズケーキ」,「アップルパイ」,「スイートポテト」の作り方（クックパッドから収集）について書かれたテキストデータを用いて,「マカロン」の分類パターンを抽出する手順を示す．RNNの学習ネットワークは，入力層のノード数（総単語種類数）が3,134個，中間層のノード数は50個，出力層のノード数が5個となっている．

利用者はまず，自分が分類パターンの抽出を行いたい分類先を選択する必要がある．今回のように1つの分類先の分類パターンの抽出を行いたい場合は，目的の分類先を選択すれば良い．最初にシステムを立ち上げると，最初に全ての分類先で，図27のように，分類先のノードと分類パターンが表示される．今回は5つの分類先のうち，目的の「マカロン」を選択する．すると，システム上では図28に示すように,「マカロン」の分類パターンが表示される．なお，分類パターン長などの初期値は分類パターン長が3，分類パターン数が1となっているため，今回は分類パターン数選択機能（図28の①）を使用して，分類パターン数を5とする．各調整を終えたシステムの表示画面を図29に示す．

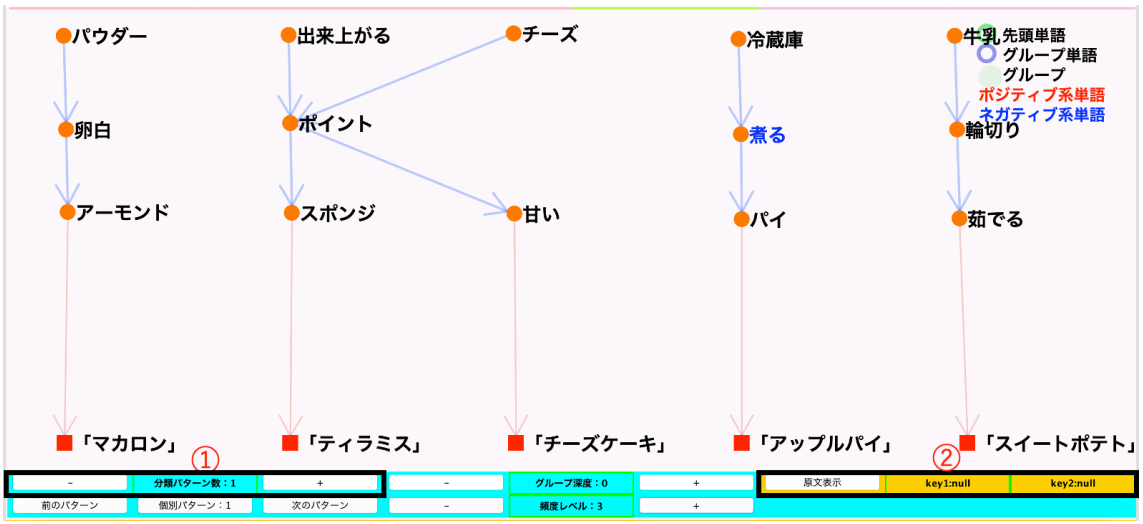

図 27　システムの初期表示（分類パターン数は1）

ここから，分類パターンの解釈を行っていく．まず重要度が高い順（矢印の太さが太い分類パターン）に分類パターンを見ていくと，最も重要な分類パターンから得られる特徴は,「パウダー」→「卵白」→「アーモンド」である．なお，ここでは単語の順序を「→」記号で表すこととする．ただ,「パウダー」→「卵白」→「アーモンド」だけでは，分類先「マカロン」に対する意味がわかりにくいの

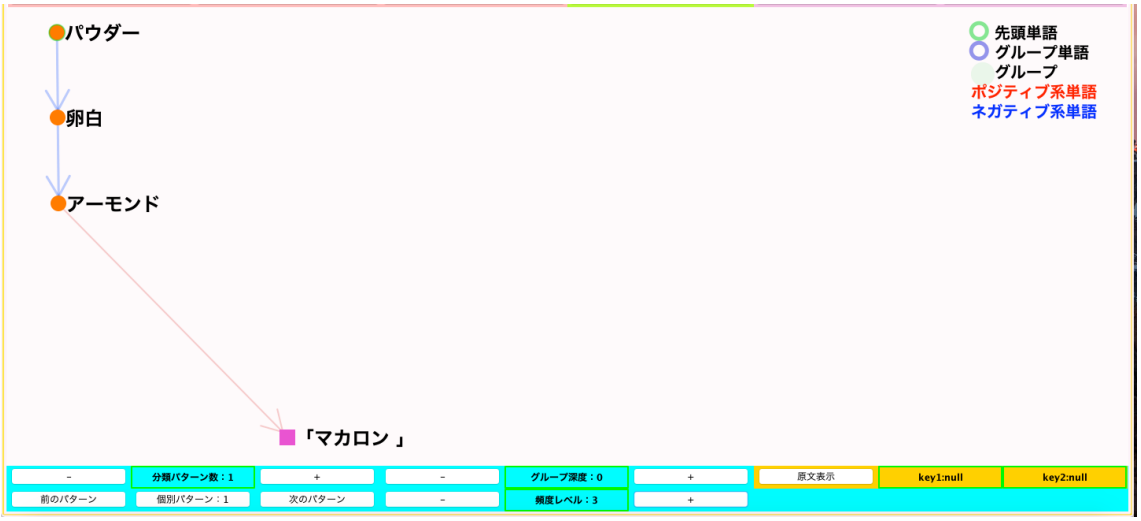

図 28　「マカロン」選択時のシステム表示（分類パターン数は1）

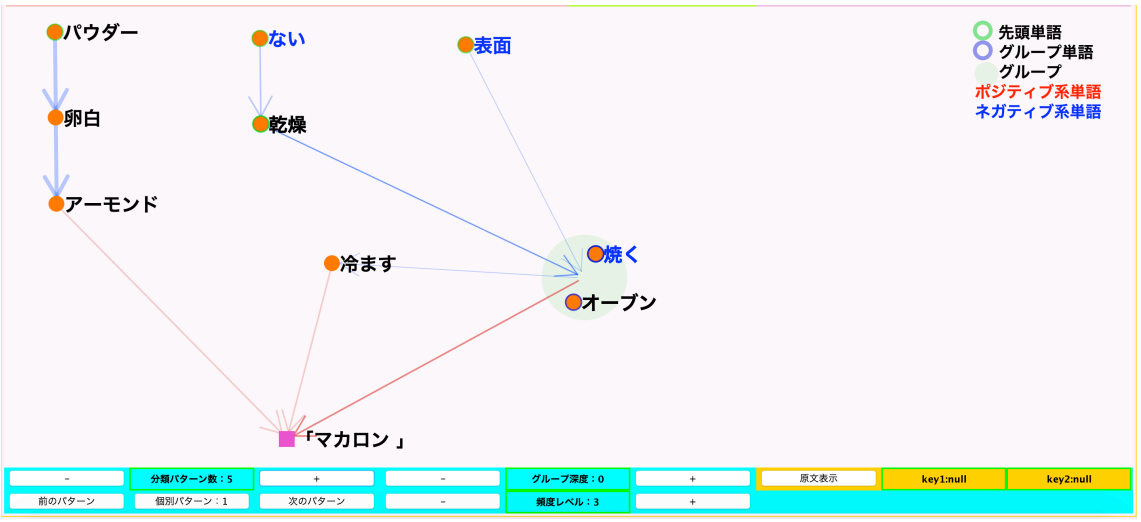

図 29　「マカロン」選択時のシステム表示（分類パターン数は5）

で，この場合は原文表示機能（図28の②）を使用する．システム画面上で,「パウダー」→「卵白」の順に単語をクリックすると，図28の②に選択した単語セットされ（図30）,「原文表示ボタン」を押すと,「パウダー」→「卵白」が出現している原文の一部が表示される（図31）．図31より,「パウダー」→「卵白」は，テキストの「アーモンド**パウダー**と**卵白**を混ぜる」という部分が抽出されて表示されたことがわかり，ここから，アーモンドパウダーと卵白を混ぜることが,「マカロン」の作り方にとって重要であると解釈できる．これは分類先「マカロン」の作り方への理解につながる．このように，システムで分類パターンを抽出し，意味を理解していくことで，分類先のテキストの内容を理解することができる．

図 30　「パウダー」→「卵白」選択時の原文表示ボタン

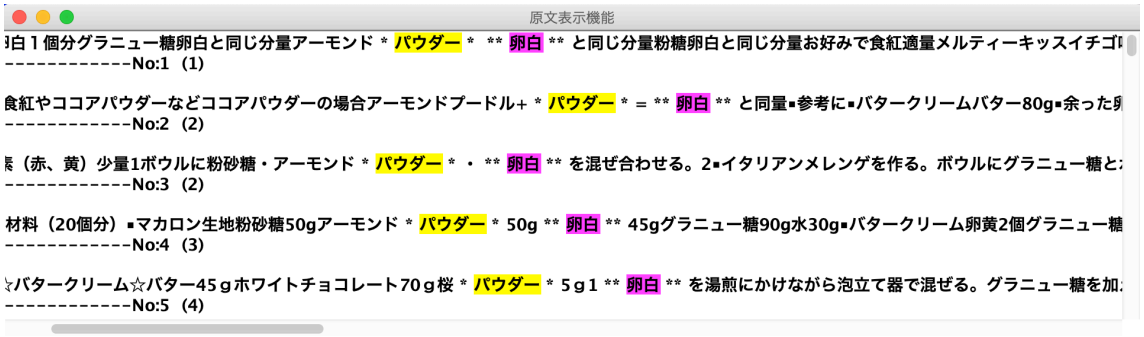

図 31　「パウダー」→「卵白」選択時選択時の原文表示

4.7　RNNの重みネットワークを用いたテキスト分類パターンの解釈支援システムの有効性の検証実験

本章では，深層学習の深い知見を有さない被験者が提案システムの出力する単語の出現パターンをもとに，分類パターンの解釈を行うことができるかを検証した実験について述べる．

4.7.1　使用テキストデータと学習モデル

表7に，本実験で用いた課題ごとの深層学習データを示す．学習データは，課題「キャラセリフ」においては，Twitterの「ツンデレbot」「デレデレbot」「キャラセリフbot」からそれぞれ「ツンデレ」「デレデレ」「ノーマル」のキャラクターの特徴を持つセリフを500件ずつ，計1,500件を利用した[12]．課題「家電レビュー」においては，amazon[13]の「人気家電製品」の上位50種類より，「役に立つ」（星4以上[14]かつ，「役に立つ人数」が10以上）レビュー，「役に立たない」（星4以上かつ，「役に立つ人数」が0）レビュー，「低評価」（星2以下）レビューをそれぞれ1,036件ずつ，計3,108件を利用した．課題「ゲームレビュー」においては，amazon[4]の「人気ゲームソフト」の上位100種類より，「役に立つ」（星4以上かつ，「役に立つ人数」が10

[12]使用したbotについては，「ツンデレ Twitter bot」「デレデレ Twitter bot」「キャラセリフ Twitter bot」で検索した際の上位のbotアカウント（データ取得日：2020年7月10日）を使用．

[13]amazon: (URL) https://www.amazon.co.jp （データ取得日：2020年7月14日）

[14]「星4以上」のレビューを対象としたのは，同じ高評価のレビューの中でも，意味のあるレビューとそうでないレビューが混在していると考えられ，それらを区別する学習の結果についての解釈を与えてもらうことを意図したこと，ならびに星の数が多いレビューは，その文章量も多い傾向があったことによる．

表 7　課題ごとの深層学習データ

	キャラ セリフ	家電 レビュー	ゲーム レビュー
学習テキスト数	1,500	3,108	4,419
1テキストの文字数 (平均)	40	244	455
入力層ノード	510	916	1809
中間層ノード	10	10	15
出力層ノード	3	3	3
分類精度	98.7%	99.2%	96.7%

以上）レビュー,「役に立たない」（星4以上かつ,「役に立つ人数」が0）レビュー,「低評価」（星2以下）レビューをそれぞれ1,473件ずつ，計4,419件を利用した[15].

学習はLSTMによって行い，中間層は1層とした．中間層のノード数は，分類精度が95%を下回らない範囲で，ノード数を減らす操作を行った．学習率は0.1，l1ノルム係数，l2ノルム係数はともに0.0001，学習回数は50回で学習を行った．

4.7.2　実験手順

実験は，表9に示す3つの課題に対して，各課題ごとに指定する「出力ラベル」に分類される文章の分類パターンの解釈を，行ってもらった．解釈の際には，被験者が解釈をしやすいように，また，解釈結果の解析を容易にするため，課題ごとに解釈の目的を設定した．実験は深層学習についての深い見識がない16名の大学生，大学院生に対して行い，提案システムを用いるグループと，比較システムを用いるグループに8名ずつの2つに分けて行った[16]．実験では，学習結果を言葉で表して説明する提案システムとの比較として，分類先の特徴を得る方法として，分類先カテゴリに特有の単語を抽出するTFIDFを比較システムとして用いた．提案システムを用いるグループでは，提案システムを用いて，分類に寄与する単語（単語単体，組合せ，時系列順序）を見つけてもらった．比較システムを用いるグルー

[15]本実験で用いたテキストデータについて，単語数が1，2個だけなど極端に短いテキストや，過度に不自然な日本語のテキストはあらかじめ除いた．また，本実験は存在するテキストデータのありのままの情報の解釈できるかを確認することを目的としているため，例えばレビューテキストにおいて，その内容の正しさは問わないものとした．

[16]また，本実験では，被験者の性格や考え方の影響が大きいと思われる解釈の質は問わず，被験者の与える解釈が，元のテキストデータに当てはまるか否かの妥当性のみに焦点をあてたため，メンバーを変えることなく実験を行った．

表 8　被験者の解釈の分類結果と分類理由の例

分類結果	解釈例	分類理由
妥当解釈	「勘違いしないでよね」という枕詞を置いた後に，今までの会話を否定するようなことを言うのが特徴	原文中に「勘違いしないでよね」と相手を否定する内容の記述が確認できたため
	カーペットの場合のノズルの性能が述べられているのが特徴と考えられる	原文中に「カーペット用のノズルについて」の記述が確認できたため
不明解釈	ツンデレという属性は，恋愛関係に強く関連する属性だろうと推測できる	原文中で恋愛に関係する文章がどれか断定しにくいため
	マイナス的な発言をしない傾向がある	原文中で「マイナスな発言」の有無が断定しにくいため
不当解釈	他の種類の掃除機に比べ，ロボット掃除機があまり評価が高く無いと考えられるなど	原文中でロボット掃除機の評価が低いことを示す記述は見られなかったため
	製品の仕様と評価について書いているのが特徴と考えられる	全ての原文に当てはまり，解釈の目的を達成していないため

プでは，比較システムとして，指定する出力ラベルに特有の単語を式(18)のTFIDF値により抽出してリスト形式で提示するシステムを用意し，これらの単語を元に，分類に寄与する単語を見つけてもらった．また比較システムにおいても，提案システムの原文表示機能を利用できるようにした．

なお，被験者がすでに実験対象となるテキストに対しての事前知識を有する可能性があるが，回答に際しては，被験者が考える解釈が，実際に分類先のテキストに当てはまるか否か，その妥当性を確認する必要があり，事前知識の有無が実験結果に与える影響は少ないと考えた．

$$\text{ある単語}i\text{の}TFIDF_i = \text{単語}i\text{の文章頻度} \times(\log(\frac{\text{出力ラベル数}}{\text{単語}i\text{の}DF\text{値}}) + 1) \tag{18}$$

実験手順について，以下のステップで両グループの被験者に解釈を行ってもらった．その際，提案システムで表示される分類パターン数は，単語数3個から構成される分類パターンを尤度の高い順に5つとした．また，比較システムでの表示単語数も提案システムに合わせて15個とした．

表 9　被験者に与えた実験課題と解釈目的

課題名	内容	解釈の目的
課題1「キャラセリフ」: 出力ラベル「ツンデレ」	アニメや漫画に出てくる特有の特徴を持つキャラクターのセリフの分類:「ツンデレ」の特徴を持つキャラクターのセリフの特徴を解釈してもらう.	あなたが小説家になったとして，小説に出すための「ツンデレ」キャラに特有の言葉の使い方のパターンを見つけて，その解釈を与えてください.
課題2「家電レビュー」: 出力ラベル「役に立つ」	Amazonの人気家電についてのレビューの分類：「このレビューが役に立った」の人数が多いレビューの特徴を解釈してもらう.	あなたは家電紹介の記者になったとして，人気家電製品について,「人の役に立つレビュー」特有の言葉の使い方のパターンを見つけて，その解釈を与えてください.
課題3「ゲームレビュー」: 出力ラベル「役に立つ」	Amazonの人気ゲームソフトについてのレビューの分類:「このレビューが役に立った」の人数が多いレビューの特徴を解釈してもらう.	あなたはゲームソフト紹介の記者になったとして，「人の役に立つレビュー」特有の言葉の使い方のパターンを見つけて，その解釈を与えてください.

手順１　解釈対象の出力ラベルを選択する:「キャラセリフ」課題では「ツンデレ」に分類されるキャラのセリフを対象とした.「家電レビュー」課題と「ゲームレビュー」課題では評価が4以上で「役に立つ」が10以上のレビューを対象とした.

手順２　それぞれの選択した出力ラベルに対応した「解釈の目的」を読んで内容を理解する.

手順３　選択した出力について,「解釈支援ネットワーク」を表示させて，出力に寄与すると思われる特徴（単語単体や組合せ，時系列順序等）を10個見つける.

手順４　注目した特徴に対して,「解釈の目的」に従った解釈を，原文表示機能を用いながら考案してもらう.

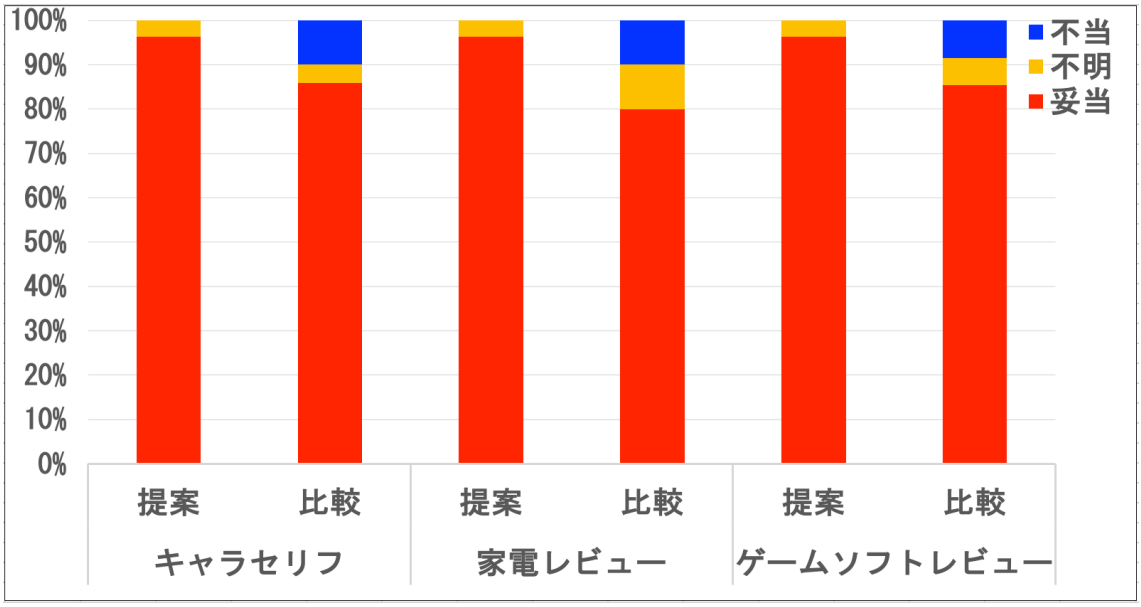

図 32　被験者の解釈の妥当性の内訳（被験者平均）

4.7.3　結果と考察

まず，被験者により記述された解釈の妥当性の内訳（被験者平均）を図32に示す．ただし，解釈の妥当性の内訳は，以下に定義する内容をもとに，著者の1名が分類を行った．

- 妥当な解釈（妥当）：内容の正しさが原文から確認でき,「解釈の目的」にも合っている．
- 妥当か判断できない解釈（不明）：内容の意図がはっきりせず，妥当か妥当でないかが判断できない．
- 妥当でない解釈（不当）：解釈の内容に誤りが確認できたか．または「解釈の目的」に合った内容ではない．

解釈の妥当性による分類は，以下の手順に基づいて機械的に行い，見落としを避けるため，この作業を時間を空けて複数回繰り返して行った．

1) 解釈の内容が，表9の「解釈目的」に合っているかを確認する．明らかに目的に合っていない解釈は「不当解釈」に分類する．

2) 解釈を導く際に注目した特徴（単語）が出現している原文集合（ORG）を対象とする．この原文が存在しない解釈は「不当解釈」に分類する．

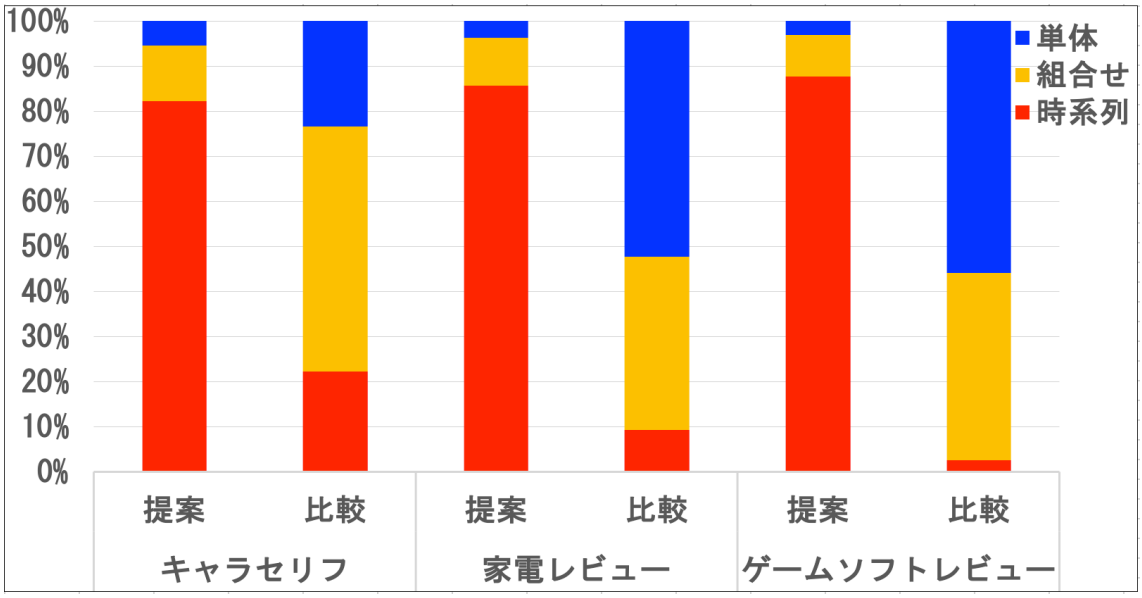

図 33　被験者別の不明，不当解釈個数

3) ORGに解釈の内容が含まれていれば「妥当解釈」に，含まれていなければ「不当解釈」に分類する．

4) 解釈の意味が理解できない場合，または複数の意味が考えられる場合など，3)で解釈の内容がORGに含まれるか否か明確に判断できない場合は「不明解釈」に分類する．

また，表8に実際に上記の分類手順で「妥当解釈」,「不当解釈」,「不明解釈」それぞれに分類された解釈の例と理由と示す．

図32の結果から，提案システムでは，97%以上の解釈が妥当な解釈に分類され，その正しさが確認できた．特に，比較システムで10%近く存在している妥当でない解釈については，提案システムの結果ではひとつも見られなかった．また，妥当か判断できない解釈についても，比較システムでは全体の5%から10%程度に含まれていたが，提案システムでは全体の3%以下であった．このことから，提案システムではより意図が明確で妥当な内容の解釈が行われていたと言える..

また，被験者ごとの「妥当でない解釈（不当解釈)」と「妥当かどうか判断できない解釈（不明解釈)」の数を図33に示す．図33のAからHは提案システムの被験者8名，IからPは比較システムの被験者8名を表す．

図33より,「不明解釈」を行った被験者の数は，提案システムで5人，比較システムで7人と大きな差はなかった一方で,「不当解釈」を行った被験者は提案システ

表 10　被験者の解釈趣旨の傾向（Aは提案システム，Bは比較システムの解釈）

課題	解釈の趣旨（件数）	注目した特徴例	解釈例
A:キャラセリフ	キャラ特有の表現（32件）	時系列：「[ない]→[嫌い]」など	「好きじゃ[ない]が[嫌い]でも[ない]と，相手への好意を示す際に曖昧な表現をする」など
A:家電レビュー	製品の付属品などの詳細（22件）	時系列：「[付属]の後に [充 電], [パック], [ノズル]という言葉が続いている」など	「[付属]品についての詳しい情報が役立つ場合が多いと考えられる」など
A:ゲームレビュー	ゲームの興味を惹かれる部分（24件）	時系列：　「[史上], [オープン]と続いている」など	「[史上]最高と書くことで面白さが伝わりさらに最近人気の[オープン]ワールドゲームという情報を入れることで興味を持たせられると考えられる」など
B:キャラセリフ	キャラ特有の表現（34件）	単体:「[ない]が上位に上がっている」など	「好きじゃ[ない]のように言葉を否定するのが特徴と考えられる」など
B:家電レビュー	製品自体について（36件）	単体:「[明るい]の単語が頻度が高い」など	「ライトの[明る]さに関しての記事が明確に書かれているものが多い傾向にある」など
B:ゲームレビュー	ユーザの求めるジャンル（21件）	単体：「[ファンタジー]がジャンルとして出現している」など	「[ファンタジー]性をゲームに求めているユーザーが多いと考えられる」など

ムで0人に対し，比較システムで8人全員となり，1人を除いて複数の「不当解釈」を与えていたことがわかる．そのため，個人差によらず提案システムを用いた方が，より妥当な解釈を与えられたことが確認できる．

次に，被験者によって与えられた，解釈に当てはまる原文の割合（被験者平均）[17]を，図34に示す．課題「キャラセリフ」では結果がほぼ同じだったが，課題「家電レビュー」と課題「ゲームレビュー」については，提案システムの方がより多くの原文に当てはまる解釈を導くことができていた．特に，課題「ゲームレビュー」では提案システムが比較システムを30%近く上回っており，提案システムで表示される解釈支援ネットワークでは，より原文の広い範囲に当てはまる典型的な解釈を導くことができていたと言える．

図35に被験者がどの特徴（単語単体や組合せ，時系列順序等）に注目して解釈を行ったかの内訳（被験者平均）を示す．ただし，注目した特徴の内訳は，以下に定義する内容をもとに，著者の1名が分類を行った．

- 単語単体：1つの単語から1つの解釈を行っている．

[17]集計対象は，先の解釈の妥当性の内訳で「妥当な解釈」に分類されたもののみとし，被験者ごとにそれらの解釈の内容と合致する記述を含む原文の和集合の個数を，課題ごとの原文数（課題「キャラセリフ」では500件,「家電レビュー」では1,036件,「ゲームレビュー」では1,473件）で除した値を，解釈が当てはまる原文の割合としている．

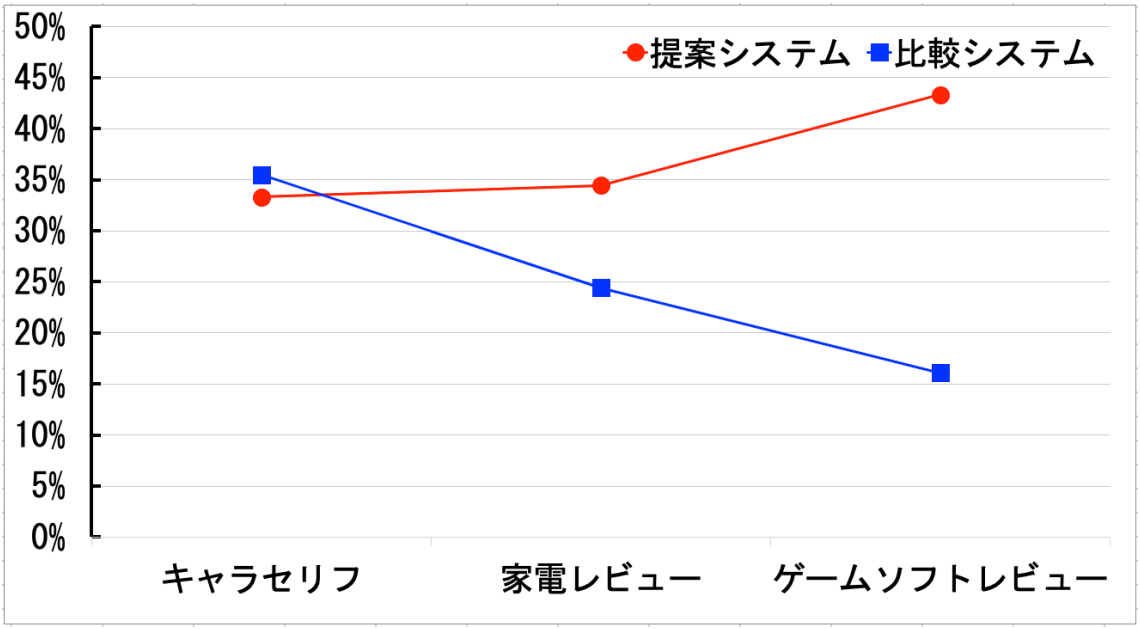

図 34　被験者の解釈が当てはまる原文の割合（被験者平均）

- 組合せ：時系列関係を特に考慮せず，複数の単語から 1 つの解釈を行っている．
- 時系列：時系列関係を考慮して，複数の単語から 1 つの解釈を行っている．

図35の結果から，提案システムでは，単語の時系列関係に注目した解釈が80%以上の割合で行われていた．また，比較システムでは，単語の時系列関係に注目した解釈は10%ほどで，残りは単語単体と単語の組合せが同程度の割合となっていた．これは，提案システムの解釈支援ネットワークでは単語の時系列情報がわかりやすかったため，被験者も単語の時系列に注目した解釈が行いやすかったためと考えられる．反対に，比較システムでは，TFIDF上位の特徴的な単語が表示されてはいたが，その単語 1 つ 1 つの繋がりが不明で，単語 1 個から解釈するか，似たような意味の単語を組合せて解釈する場合が多かったためと考えられる．このことから，提案システムでは単語の時系列を考慮した，典型的な解釈が行われていたと言える．

最後に，表10に，提案システムと比較システムで，課題ごとに解釈の傾向として，特に多かった解釈の趣旨と例を示す．なお，[]でかこっている単語は，提案システムの解釈支援ネットワーク上や比較システムの単語リスト上で実際に表示されていた単語を示している．また，注目した特徴例では，注目した特徴が時系列かどうかの分類も表記している．

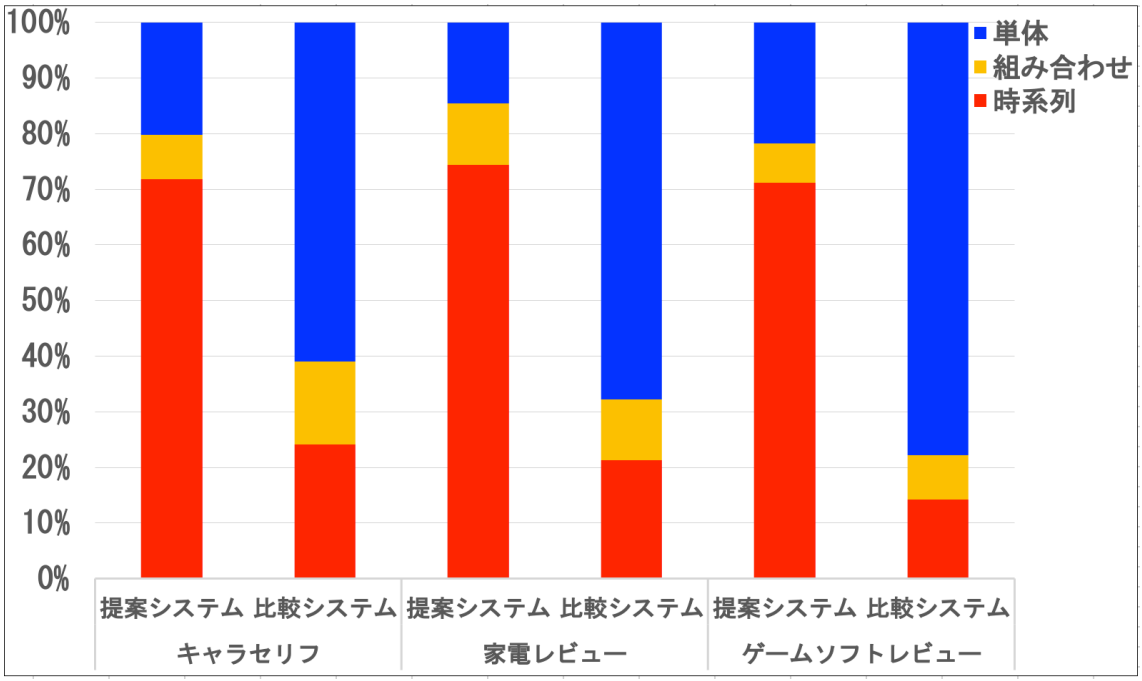

図 35　被験者が注目した特徴の内訳（被験者平均）

表10より，課題「家電レビュー」と「ゲームレビュー」について，提案システムでは製品本体やゲームの内容ではなく，製品の何に注目したレビューがいいのかや，人が興味を持ちそうな内容に関する記述など，製品やゲームの詳細を考察するような解釈が多いことが確認できた．また，これらのほとんどは時系列の特徴に注目した解釈であり，時系列の特徴に注目することで，文章全体を考察するような解釈ができると考えられる．逆に，比較システムでは，単語単体から，製品本体の特徴やゲームの内容（ジャンル）についての解釈が多い結果となった．このため，比較システムにおいて，一部の製品やゲームにしか当てはまらない解釈が多く，解釈が当てはまる原文の割合が下がってしまったと考えられる．

一方で，課題「キャラセリフ」において，提案システムでは時系列のパターン，比較システムでは単語単体のパターンに注目した解釈が多かったが，どちらもキャラの特有の表現について記述が多い傾向が確認できた．これは，課題「キャラセリフ」の平均文字数が40字程度（単語数にして20単語程度）と短く，原文表示機能で単語単体と複数の単語，どちらを選択した場合でもほぼ全文が参照できてしまうことから，結果として両グループで同じような解釈になったと考えられる．また，課題「キャラセリフ」の解釈が当てはまる原文の割合が，両グループで同程度になった結果についても，同じ理由が考えられる．

以上をまとめると，提案システムでは，比較システムより，正解率の高い，広い範囲の原文に当てはまるような，典型的で妥当な解釈が導き出せることが確認できた．これは，特に複数の単語の時系列関係に注目して解釈が行えることが要因と言える．さらに，課題「キャラセリフ」のようにテキストの文字数が短い文章の場合でも，提案システムでは，TFIDF値の高い単語を参照するのと同じ水準で，典型的な解釈が導き出せるとことが確認できた．

5. 結言

まず本研究の目的を振り返ると，深層学習の学習結果について，その分類根拠を表す分類パターンの解釈を支援するシステムの構築を行った．また，その分類パターンについて，学習済みの深層学習の学習ネットワークの中身，すなわちネットワークを構築している層やノードで一体何が学習されたのかを，学習によってノード間に付与された重みの値を参照することで，学習された特徴の集まりである分類パターンとして抽出する手法を提案した．そこで，まずは3章で最も基本的な深層学習モデルである，DNNについて，提案手法を用いて分類パターンを抽出し，その解釈を支援するためのインタフェースを持つシステムを構築した．DNNに対して提案手法を適用したシステムの有効性の検証実験では，まず本当にこの手法で抽出された分類パターンの解釈が，分類根拠の理解に役立つのかを検証するため，被験者に文章集合に対して提案システムを用いて分類パターンの解釈を行ってもらい，その解釈内容が学習に用いた文章集合の内容に対して妥当であるかを調査した．その結果，DNNに対して提案手法を適用したシステムでは，分類パターンの解釈の90%近くが学習に用いた文章集合の内容に対して妥当であるとわかり，それはすなわち，DNNが学習した分類の特徴，つまり最初の目的である分類根拠が，システム利用者に理解できたと結論付けた．

本研究の提案手法がDNNには有効とわかったところで，4章では，実際に自然言語処理分野で広く使われている，文章中の単語の時系列情報を学習できるRNNに提案手法を適用したシステムの構築を行った．RNNでは，時系列順に中間層を展開すると，DNNと同じ形になるため，DNNの時と同様に重みを参照することで，時間ごとの中間層が学習した特徴（単語情報）を抽出し，時系列情報を含んだ分類パターンを抽出できると予想した．さらに，そのRNNの展開した形が，従来から時系列パターンに対する尤度の推定などに活用されてきたHMMと類似している点に注目し，RNNをひとつのHMMと見なすことで，単語の時系列パターンに対する尤度を算出し，容易に時系列情報を含んだ分類パターンの抽出が可能となった．RNNに対して提案手法を適用したシステムの有効性の検証実験では，DNNの時と同様に，被験者に文章集合に対して提案システムを用いて分類パターンの解釈を行ってもらい，その解釈内容が学習に用いた文章集合の内容に対して妥当であるかを調査した．その結果，分類パターンの解釈の95%近くが学習に用いた文章集合の内容に対して妥当であると確認された．また，時系列情報が含まれてい

ることが解釈にどう有効かを調査するため，TFIDF値を用いて文章中の特徴的な単語をリストで表示するシステムを比較のために用意し，その比較システムによって行われた解釈と提案システムの解釈を比較した．その結果，比較システムでは10%ほど存在した妥当ではない解釈が，提案システムでは0%だったことがわかり，この結果から，分類パターンには文章中の特徴の時系列情報を含ませた方が，より解釈の精度が向上すると結論付けた．

以上の考察により，本研究で提案する重みによって深層学習の学習ネットワークが学習した内容を分類パターンとして抽出する手法と，抽出された分類パターンの解釈を支援するインタフェースを備えたシステムの構築は，深層学習の分類根拠を理解するために有効であり，その提案手法はDNN等の単純なモデルから，RNNなどの発展系のモデル活用できると結論づけられる．また，RNNの発展系であるLSTMや，実際に実験は行っていないが同じくDNNの発展系であるCNNなどにも本研究の提案手法が応用可能であると想定され，深層学習の分類根拠の理解のために，本研究は大きな助けになると考えている．今後の研究では，自然言語処理分野の最新の研究で多く使われているTransformer[15]やBERT[46]，GPT-3[47]などの深層学習モデルに対しても，本研究の波及効果の充実を進めていきたい．

なお，実験に使用したデータについては，Twitter等の個人情報が含まれるため，一般的な公開は行わないこととする．ただし，研究目的での利用を検討する場合は，次の連絡先にて，個別に対応を行う．Email: oh23mandou@ec.usp.ac.jp

謝辞

本研究を進めるにあたり，まず，指導教官主査である，滋賀県立大学工学部 砂山渡先生には，大学生時代の卒論研究指導から始まり，修士論文，博士論文作成まで，非常に長い期間，研究に向かう姿勢や研究に関する困難克服のための具体的な方策までていねいに教えていただきました．他にも，学会での研究発表の心得や学会での各種手続き等，砂山先生がいらっしゃらなければ今の私はここまでこれなかったと思います．本当に心からの深い感謝を申し上げます．さらに，今回，副査を担当していただいた九州大学 マス・フォア・インダストリ研究所 河原吉伸先生にも，他大学生でありながら頻繁に研究室を訪れた私に，通常業務もありながら，最大限，研究の進め方や不足していた知識の講義を行っていただき，非常に親身になって相談にのっていただきましたこと，深くお礼申し上げます．

また，河原吉伸先生のお口添えで理化学研究所の仕事を紹介していただき，短期間でしたが，様々な知識や経験を積むことができました．また，理化学研究所在籍中も熱心にご指導いただき，研究を進める大きな力となったこと，重ねて感謝いたします．

大分大学理工学部 畑中裕司先生，そして京都産業大学理学部 小郷原一智先生には，畑中裕司先生と小郷原一智先生が滋賀県立大学工学部所属だった大学学部学生，修士学生時代の私に，研究の楽しさと難しさを教えてくださいましたこと，大変お礼申し上げます．研究開発に向かう姿勢を厳しくご指導くださるとともに，研究室での生活や新しいものを作り上げる喜びを教えてくださったこと，また，特に畑中裕司先生には，ご多忙の中，今回の副査を担当していただいたこと，感謝の念に堪えません．

滋賀県立大学工学部 南川久人先生，同じく同大学工学部 酒井道先生には，大学学部学生時代より多大なるご指導をいただいてまいりました．ここに深く感謝いたします．また，南川久人先生，酒井道先生ともにご多忙の中，今回の副査を担当していただき，重ねてお礼申しあげます．さらに，卒業論文，修士論文，博士論文において必要な実験に惜しみなく協力していただいた研究室のメンバーにも，心からの感謝とお礼を申し上げます．

最後に，これまで自分の思う道を進むことに対し，温かく見守りそして辛抱強く支援してくださった両親に対して深い感謝の意を表して謝辞と致します．

参考文献

[1] R Frank, The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, Vol.65, No.6, pp.386-408, 1958.

[2] M. Marvin and A. P. Seymour, Perceptrons, MITPress, 1969.

[3] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors, Nature, Vol.323, pp.533-536, 1986.

[4] G. E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput, Vol.18, No.7, pp.1527-1554, 2006.

[5] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS), 2012.

[6] L. Sutskever, O. Vinyals, and Q. Le, Sequence to Sequence Learning with Neural Networks, Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Vol.2, pp.3104-3112, 2014.

[7] 久保 陽太郎, 音声認識のための深層学習, 人工知能:人工知能学会誌, Vol.29, No.1, pp.62-71, 2014.

[8] S. Fernández, A. Graves, J. Schmidhuber, An Application of Recurrent Neural Networks to Discriminative Keyword Spotting, Proceedings of the 17th International Conference on Artificial Neural Networks (ICANN), pp.220-229, 2007.

[9] 立石雅彦, 山崎晴明, 手書き数字認識における改装ニューラルネットワークの中間層に関する考察, 情報処理学会論文誌, Vol.30, No.10, pp.1281-1288, 1989.

[10] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, Multimodal Emoji Prediction, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.2, pp.679-686, 2018.

[11] 石晶, 李志豪, 本吉俊之, 大西直, 森裕紀, 尾形哲也氏, End-to-End自動運転モデル改善のための画像認識サブタスクの設計と評価, 第33回人工知能学会全国大会, 1L2-J-11-01, 2019.

[12] 妻広明, 人工知能による運転支援・自動運転技術　現状と課題, 計測と制御, Vol.54, No.11, pp.808-815, 2015.

[13] 国土交通省, 物流の現状とドローン物流の主な取組, https://www.mlit.go.jp/common/001282862.pdf, (2022年6月1日確認).

[14] 本間 経康, 張 暁勇, 鈴木 真太郎, 魚住 洋佑, 市地 慶, 柳垣 聡, 高根 侑美, 川住 祐介, 石橋 忠司, 吉澤 誠, 深層学習：医療ビッグデータと診断支援システム, 生体医工学, Vol.55, No.3, pp.228-228, 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, In the Annual Conference on Neural Information Processing Systems (NIPS), 2017.

[16] ベルテクス AI INNOVATION SOLUTION チーム, 海外での文章自動生成AIのSEO活用事例, https://www.vertex-itb.com/single-post/ainlg-seo, (2022年6月1日確認).

[17] 伊藤 暢, 横井 恵人, 榎 藍香, 山下 芳樹, 浅野 夏美, 人事採用における納得阻害要因の解明に向けたサーベイ実験, 第35回人工知能学会全国大会, 4H2-GS-11c-04, 2021.

[18] 鍾 淑玲, 台湾コンビニのデジタル・イノベーション, 流通, Vol.2020, No.46, pp.29-44, 2020.

[19] 大日本住友製薬とExscientia Ltd.の共同研究 人工知能(AI)を活用して創製された新薬候補化合物のフェーズ1 試験を開始, https://www.ds-pharma.co.jp/ir/news/pdf/ne20200130_2.pdf, (2022年6月1日確認).

[20] 内閣府, 人工知能技術戦略実行計画(案), https://www8.cao.go.jp/cstp/tyousakai/jinkochino/keikaku.pdf, (2022年6月1日確認).

[21] M. Daniluk, T. Rocktaschel, J. Welbl, S. Riedel, Frustratingly Short Attention Spans in Neural Language, ICLR, 2017.

[22] 藤吉 弘亘, 山下 隆義, 平川 翼, アテンションマップによる深層学習モデルの判断根拠の可視化, Medical Imaging Technology, Vol.39, No.3, pp.110-116, 2021.

[23] M. Luong, H. Pham, C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.1412-1421, 2015.

[24] J. Gehring, M.Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning, arXiv preprint arXiv:1705.03122v2, 2017.

[25] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, S. Behnke, Interpretable and fine-grained visual explanations for convolutional neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.9097-9107, 2019.

[26] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, Proceedings of the 2020 Conference on Fairness Accountability and Transparency, pp.629-639, 2020.

[27] 松倉健太郎, 村田 昇, ニューラルネットワークの中間層における独立な特徴量の抽出, 電子情報通信学会技術研究報告, Vol.106, No.102, pp.63-67, 2006.

[28] M. Tsang, D. Cheng, Y. Liu, DETECTING STATISTICAL INTERACTIONS FROM NEURAL NETWORK WEIGHTS, ICLR, 2018.

[29] D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA), 2017.

[30] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?, IEEE Computational Intelligence Magazine, Vol.14, No.1, pp.69-81, 2019.

[31] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, A. K. Pradhan, Explanations and expectations: Trust building in automated vehicles, Companion of the ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp.119-120, 2018.

[32] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, M. Sebag, Learning functional causal models with generative neural networks, Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp.39-80, 2018.

[33] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp.6276-6282, 2019.

[34] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Transactions on Neural Networks and Learning Systems, Vol.30, No.9 pp.2805-2824, 2019.

[35] G. Audemard, F. Koriche, P. Marquis, On Tractable XAI Queries based on Compiled Representations, KR Proceedings 2020 Special Session on KR and Machine Learning, pp.838-849, 2020.

[36] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, Interpreting CNNs via decision trees, IEEE Conference on Computer Vision and Pattern Recognition, pp. 6261-6270, 2019.

[37] ボレガラ ダヌシカ, 自然言語処理のための深層学習, 人工知能:人工知能学会誌, Vol.29, No.2, pp.195-201, 2014.

[38] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ICLR, 2013.

[39] X. Glorot, A. Bordes, Y. Bengio, Deep Sparse Rectifier Neural Networks, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11), pp.315-323, 2015.

[40] 山本 健二, 小圷 成一, 岡本 卓, 平田 廣則, パルスニューラルネットワークのための学習率最適化を用いた誤差逆伝播学習法, 電気学会論文誌C（電子・情報・システム部門誌）, Vol.128, No.7, p.1137-1142, 2008.

[41] 麻生英樹, 多層ニューラルネットワークによる深層学習の学習, 人工知能学会誌, Vol.28, No.4, pp.649-659, 2013.

[42] 巣篭 悠輔, Deep Learning Java プログラミング 深層学習の理論と実装, 株式会社インプレス, 2016.

[43] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning, In Neural Information Processing Systems.

[44] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, Eurospeech, Vol.99 No.1, pp.2374-2350, 1999.

[45] 平岡達也, 高瀬翔, 内海慶, 欅惇志, 岡崎直観, RNNにより高次の依存を考慮したニューラル隠れマルコフモデル, 言語処理学会第26回年次大会発表論文集, pp.1332-1335, 2020.

[46] J. Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.2, 2019.

[47] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei, Language Models are Few-Shot Learners, Advances in Neural Information Processing Systems 33 (NIPS), 2020.

研究業績

学術論文誌

1. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 深層学習ネットワークへのHMM適用によるテキストベースの分類パターン解釈支援, 日本知能情報ファジィ学会誌, Vol.34, No.1, 2022（掲載予定）.

2. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, テキストベースの深層学習における分類パターンの解釈支援, 日本知能情報ファジィ学会誌, Vol.31, No.4, pp.779-787, 2019.

国際会議

1. Masayuki Ando, Yoshinobu Kawahara, Wataru Sunayama, Yuji Hatanaka, Interpretation Support System for Classification Patterns Using HMM in Deep Learning with Texts, In Proceedings of the Fourteenth International Conference on Advances in Computer-Human Interactions (ACHI 2021), Nice (France), pp.64-70, 2021.

国内口頭発表

1. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 深層学習ネットワークへのHMM適用による分類パターン解釈支援, 第27回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.1-6, 2021.

2. 安藤雅行, 砂山渡, 畑中裕司, HMMを利用した深層学習ネットワークからの分類パターンの解釈支援システム, 第35回人工知能学会全国大会, 4C3-OS-1a-04, 2021.

3. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, HMMを利用した深層学習ネットワークからの分類パターンの抽出と解釈, 第34回人工知能学会全国大会, 3K1-OS-5a-02, 2020.

4. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, HMMを利用した深層学習ネットワークからの分類パターンの抽出と可視化, 第23回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.44-49, 2019.

5. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 深層学習における学習ネットワークからの分類パターンの抽出, 第33回人工知能学会全国大会, 4G2-OS-8a-04, 2019.

6. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 深層学習における学習ネットワークからの分類パターンの抽出, 第21回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.49-54, 2019.

7. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 深層学習における分類パターンの解釈支援, 第20回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.9-14, 2018.

8. 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, 小郷原一智, ディープラーニングにおける分類パターンの意味付け支援, 第31回人工知能学会全国大会, 2M2-OS-34a-2, 2017.