4Q-05

深層学習モデルによる音響ベクトル系列から文書ベクトル系列へのメディア横断マッピング

五藤 智晃† 西村 良太‡ 北 研二‡

徳島大学理工学部† 徳島大学大学院社会産業理工学研究部‡

1. はじめに

近年、情報技術の発展により、テキスト、画像、音楽等の様々なメディアデータがデジタル化され、多くの人が様々な形態の情報を容易に得られるようになったしかし，情報量が増大かつ複雑化し、ユーザが求める情報を正確に検索し、得ることが難しくなってきている．

本研究の目的は、楽曲（音響）から文書（歌詞）を検索することである．「楽曲」と「文書」といった異なるメディア間の検索を可能にするために、深層学習モデルである RNN-LSTM を用いてメディアを横断するマッピングの実現を行う．

2. 収集データ

2.1 楽曲データ

本研究では、音楽データと歌詞データの２つを使用する．楽曲は，日本のレコード会社 13 社が共同企画・作成した“青春歌年鑑”を使用し，この中から 672 曲を利用する．

2.2 音楽データ：chroma ベクトル

音楽データは長さを均一にするために、20 秒分の区間を１つのデータとして，各楽曲から最大 10 区間を抽出した．これにより、楽曲データは約 6500 個作成された（学習用データ約 5500 個，テスト用データ約 1000 個）．これらのデータから音楽特徴量である chroma ベクトル[1]を抽出する． chroma ベクトルの抽出には“LibROSA”[2]を用い、時間フレームごとに 12 次元のベクトルとして抽出される．これは，ある時刻における音を 12 音の音階で表現したものである．このベクトルが，一秒間に 86 個，20 秒の楽曲データから，1720 個抽出された．

2.3 歌詞データ：文書ベクトル

各楽曲の歌詞データに対しては、“MeCab [3]”を用いて形態素解析を行い、歌詞を形態素に分割した．本研究では、各楽曲から未知語を除いた 145 単語を抜き出し、“word2vec[4]” を用いて 200 次元の単語ベクトルを作成した．145 単語に満たなかった場合には、0 ベクトルを追加している．word2vec とは、各単語の意味をベクトル表現化する手法である．これにより、単語の意味の近さが，距離の近さとして計算可能である．

3. メディア横断マッピングと文書ベクトル空間の学習

本提案手法では，楽曲から文章を検索するために，ある曲の歌詞情報を持つ文書ベクトル空間の１点に，同曲の音響情報を持つ音響ベクトルを非線形変換している．そのため，最初に歌詞情報を十分に反映した，文書ベクトル空間を作るために文書モデルの構築、学習を行った。

2．3 節で抽出した各単語ベクトルから文書ベクトルを作成するために、文書モデルの学習を行った．文書モデルの構成は，Sequence-to-sequence 型の Autoencoder となっており、モデルへの入力と出力が同じものになるように学習を行う．Autoencoder は，情報を圧縮する Encoder 部と，圧縮された情報を復元する Decoder 部に分かれており，うまく学習が行われると，Encoder 部の出力が，入力された情報の特徴を表すベクトル（埋め込みベクトル）として利用可能となる．入力層の活性化関数には tanh を用いており LSTM 層の活性化関数は恒等関数を用いている．

4. 音響ベクトル空間から文書ベクトル空間への RNN-LSTM を用いたメディア横断

本研究では、2．3 節で抽出した各単語ベクトルを入力とし，文書ベクトルを出力するモデル（文書モデル）と、楽曲データを文書ベクトルへ結び付けるための音響モデルの構築を行った．各モデルの構成と提案手法の構成を図 1 に示す．文書モデル部分では、3 節で事前に学習したモデルを用いており，単語ベクトルが入力される．文書モデル内では，NN 層 600 ユニット、LSTM 層 600 ユニット、NN 層 600 ユニットと順に処理され、最後に 200 次元のベクトルが出力される．これを文書ベクトルとする．出力層の活性化関数には

Mapping of acoustic vector sequence and document vector sequence

† Tomoaki GOTO, Tokushima University

‡ Ryota NISHIMURA, Tokushima University

‡ Kenji KITA, Tokushima University

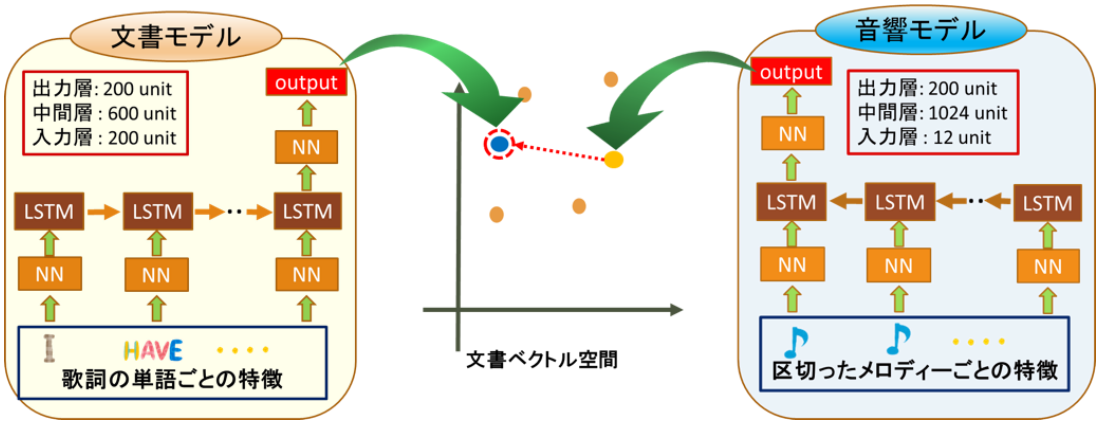

図１　提案手法の概要図

恒等関数を用いている．

音響モデルでは、入力層と出力層が 200 ユニット、NN 層（入力側，出力側）と LSTM 層が 1024 ユニットなっている．この音響モデルに 2．2 節で抽出した chroma ベクトルを時系列順に入力していき、その楽曲と対応した文書ベクトルを出力できるように学習を行う．

5.　実験結果

モデルの精度を確認するため音響モデルに chroma ベクトルを入力し、文書ベクトルを出力させた．音響モデルから出力した文書ベクトルとそれに対応する正解が 1 番目に近いものであったか、正解が 2 番目に近いものであったか、というように、モデル出力が正解に対して何番目に近かったかを調査した．この結果を図 2 示す．図はヒストグラムであり、全楽曲における各モデルから出力した文書ベクトルの距離を算出し、入力した楽曲データに対応する歌詞がでた順位（ランク）を数え上げたものである．横軸はモデルの出力結果のランクを、縦軸は該当する楽曲数を表している．今回テストに用いたモデルは学習 epoch 数が 250 回のモデルである．図 2 では、ランク上位のものがより多くカウントされており、マッピングがうまくできていると考えられる．ここでは本提案手法による学習が可能であるか検討するためにクローズドテストの結果を示している．オープンテストの結果と汎化性能の検討は今後の課題である．

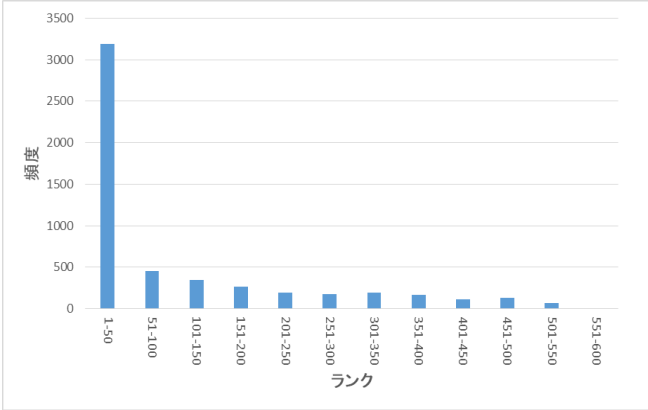

図 2 マッピング結果のヒスとグラム

6.　結論

本研究では、音響特徴量と文書ベクトルとの対応付けを可能とするモデルの構築を行った．提案手法では、時系列データの学習に RNN-LSTM を用いてクロスメディアでのマッピングを可能としている．構築したモデルを用いて評価実験を行ったところクローズドテストではうまくマッピングされることが確認された．今後はオープンテストでの調査を行い，汎化性能の検討を行う．

参考文献

[1] M. A. Bartsch and G. H. Wakefield: “Audio thumbnailing of popular music using chroma-based representations,” IEEE Transactions on Multimedia, Vol.7, No.1, pp. 96-104, 2005.
[2] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, O. Nieto. “librosa: Audio and Music Signal Analysis in Python”,Proc. of the 14th Python in Science Conf (SCIPY), pp. 18-25, 2015.
[3] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/.
[4] T. Mikolov, K. Chen, G. Corrado, J. Dean. “Efficient Estimation of Word Representations in Vector Space”, ICLR, 2013.