

お客様の声などのテキストデータの総合的な分析手法

鵜飼 正人* 小川 知行* 横内 俊秀*
影山 椋* 中挾 晃介** 遠藤 広晴***

Comprehensive Analysis Method for Text Data such as Customer Feedback

Masato UKAI Tomoyuki OGAWA Toshihide YOKOUCHI
Ryo KAGEYAMA Kosuke NAKABASAMI Hiroharu ENDO

We have developed a neural network that automatically classifies “customer feedback” into 22 categories such as insufficient cooling and excessive broadcasting. We conducted recognition experiments with the voices of about 6000 customers, and automatic classification can be performed with a correct answer rate of more than 90%. Furthermore, we have developed a tool that utilizes various data such as vehicle information recording device data and outside temperature. Using this tool, we get useful findings such as the relationship with “customer feedback”.

キーワード：自然言語処理，トピック分類，深層学習，ビッグデータ解析，車両情報

1. はじめに

年間千数百件寄せられる「お客様の声」については、人手と時間をかけてサービス向上に活用しているが、さらなる活用を目指して分析を自動化・詳細化する手法を検討した。作業の省力化に資するものとして、自然言語処理技術を用いてお客様の声のテキストを「冷房不足」や「放送過剰」などのカテゴリに分類するニューラルネットワークモデルを開発した。

さらに車両不具合原因の特定を迅速化するなどの業務改善をめざして、お客様の声から車両を特定するアルゴリズムや、車両情報記録装置のデータや気象データといった多様なデータを、図1に示すように組み合わせる活用するツールを開発したので報告する。例えば本ツールより「冷房不足」の声が発生した当該車両を特定し、車両モニタデータと突き合わせることでお客様が不満と感じる温熱指標を把握するといった分析が可能となる。

2. お客様の声テキストの特徴

2.1 「言語」処理の難しさ

人間が日常的に使っている言葉を解析する上では、自然言語処理の技術が必要である。画像や音声と違い文字は「物理的な実数値」ではなく、人間が恣意的に定義した「記号」であるため、そのままの形ではニューラルネットワークで処理できないという障壁があった。それに対して、文字や単語をベクトルや行列で計算できる数値に変換する「分散表現」という考え方が提案された。分散表現の方法は色々あるが、表1の例に示すように、単語を高次元の実数ベクトルで表現する技術である。

表1 単語の分散表現の例

単語	大きさ	速さ	涼しさ	快適さ	...
新快速	0.75	0.92	0.35	0.85	...
冷房	0.05	0.10	0.98	0.85	...
国立駅	0.95	0.10	0.15	0.55	...

2.2 自然言語処理

自然言語処理は大きく単語分割、構文解析、意味解析の処理からなる。単語分割で文章は名詞や助詞などの単位に分解することができるが、それ以上細分化してしまうと意味を成さない最小単位のことを「形態素」と呼ぶ。文章をこの形態素に分割する処理を「形態素解析」と呼ぶ。形態素に分割するのは、以下の理由による。

- ・文よりも単語の方が内容を正確に捉えられる
- ・形態素解析器が文を単語レベルで扱うようになっている
- ・処理する情報を「冷房」、「温度」などの名詞、「暑い」、「寒い」などの形容詞といった文章分類に有効なものだけに絞りたい
- ・「は」、「です」、「ます」などの助詞や助動詞は自動解析としての文章理解には役立たないので、これらのノイズを除去したい

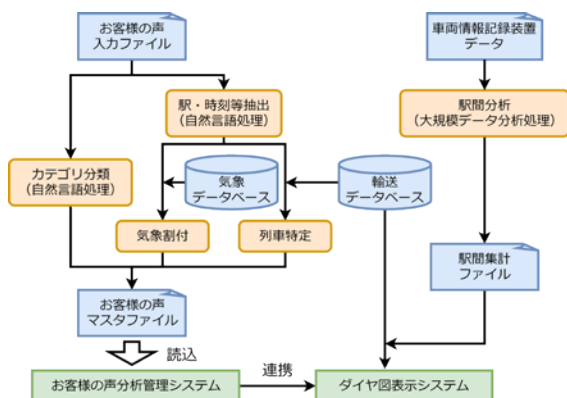


図1 多様なデータを活用したお客様の声の処理

* 信号・情報技術研究部 画像・IT研究室
** 信号・情報技術研究部 運転システム研究室
*** 人間科学研究部 人間工学研究室

一方で、日本語は英語と異なり、単語の区切りが明らかでないためこの分割処理は容易ではない。一般に形態素解析は形態素解析器と呼ばれるエンジンで実行する。今回は解析器の中でも速度と精度のバランスが良い MeCab¹⁾ を使用した。MeCab による解析例として、「東京駅 10 時発の新快速電車の車内が暑くて汗をかいた」を入力した結果を図 2 に示すが、適切に品詞分解されている。ここで着目すべき点は、名詞、形容詞の「車内」「暑い」から冷房（不足）と判定できる点である。

通常解析器は登録されている「辞書」を参照しながら形態素を抽出するが、目的に合った鉄道専用辞書を整備することは分類性能向上に有効である。例えば一般の辞書では、「新快速」は「新」と「快速」の二つの名詞に分割されるが、「新快速」という語を優先度を高くして登録すると、これを一語で抽出することができる。

東京駅 10 時発の新快速電車の車内が暑くて汗をかいた	
東京	名詞 固有名詞 地域 一般** 東京 トウキョウ トーキョー
駅	名詞 接尾 地域 *** 駅 エキ エキ
10	名詞 数 *** **
時	名詞 接尾 助数詞 *** 時 ジジ
発	名詞 接尾 一般 *** 発 ハツ ハツ
の	助詞 連体化 *** ** の ノ
新	名詞 接尾 一般 *** 新 シン シン
快	名詞 接尾 一般 *** 快 カイ カイ
速	名詞 接尾 一般 *** 速 ソク ソク
電	名詞 接尾 一般 *** 電 デン デン
車	名詞 接尾 一般 *** 車 シャ シャ
内	名詞 接尾 一般 *** 車内 シヤナイ シヤナイ
が	助詞 格助詞 一般 *** が ガ
暑	形容詞 自立 ** 形容詞 アウオ段 連用テ接続 暑い アツク
く	助詞 接続助詞 *** ** て テ
汗	名詞 一般 *** 汗 アセ アセ
を	助詞 格助詞 一般 *** を ヲ
か	動詞 自立 ** 五段 カ行イ音便 連用テ接続 かく カイ カイ
い	動詞 自立 ** 五段 カ行イ音便 連用テ接続 かく カイ カイ
た	助動詞 *** 特殊 タ 基本形 た タ

図 2 MeCab による形態素解析結果

3. 自然言語処理によるカテゴリ分類

3.1 お客様の声分類の内訳

お客様の声のカテゴリ分けの方法は鉄道事業者によって様々であるが、今回使用のお客様の声は、担当部門で 22 のカテゴリに振り分けられてシステムに登録されている。分類ラベルの名称とカテゴリ毎の件数を図 3、4 に示す。年間の件数は平均して 1300 件程度、そのうち 30% 強が冷房や暖房などの空調に関する声となっている。

3.2 自然言語前処理

テキストデータの分類器を作成するためには、学習用のデータセットに相当するコーパスを作成する必要がある。コーパスとは処理しやすいように文を構造化したもので、今回は Bag-of-Words 形式のコーパスを作成した。

文書の分散表現（ベクトルで表現する）の一つの手法に Bag-of-Words がある。図 5 に Bag-of-Words の考え方を示す。Bag-of-Words では各単語をベクトルの 1 自由度に対応させ、文書中に表れる単語の出現頻度をベクトル化する。また文書分類であれば、文書に分類ラベルを付与した形式のものを作成する必要がある。今回この

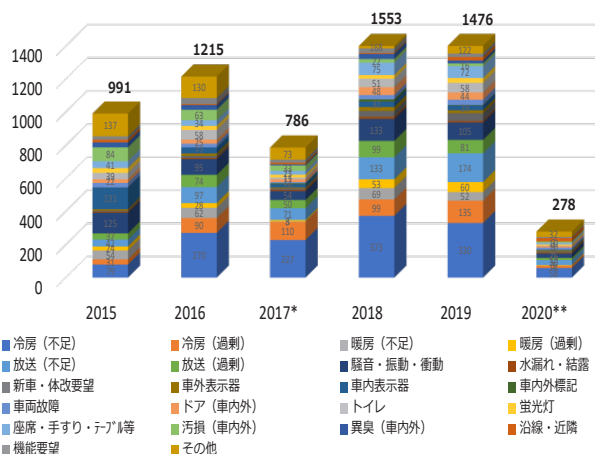


図 3 年度別のお客様の声データ数

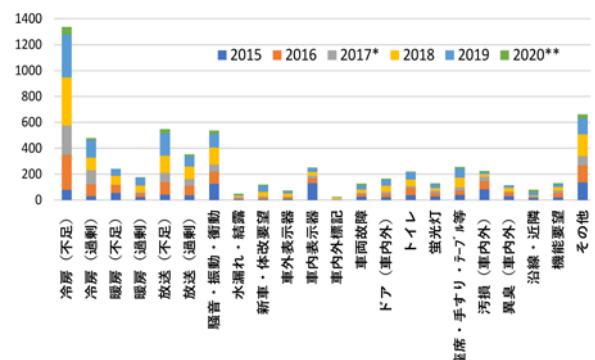


図 4 カテゴリ別のデータ数

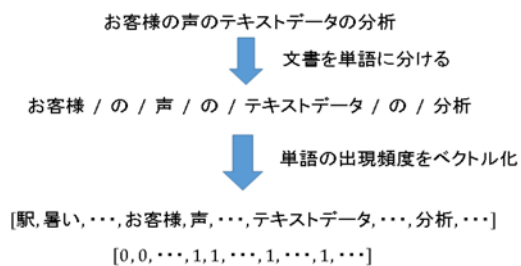


図 5 Bag-of-Words の概要

コーパスは CSV 形式とし、文書を 1 カラム目に、文書ラベルを 2 カラム目とするフォーマットで作成した。

3.3 MLP による自動分類

ニューラルネットワークの一種である多層パーセプトロン (MLP) を適用して、お客様の声を冷房不足や放送過剰など 22 のカテゴリに自動分類する分類器を作成した。MLP によるカテゴリ分類のイメージを図 6 に、MLP の詳細を表 2 に、学習時のグラフを図 7 にそれぞれ示す。図からわかるように、目標と実際の出力の誤差を表す loss が順調に減少しており、およそ 40 回のイテレーションで実用上の目安となる「loss 値 ≤ 0.1」に達している。

今回設計したモデルでは、分類精度を評価しやすいように最終層の活性化関数には複数の出力値の合計が 1.0 になるような値を返す Softmax 関数を使用した。表 3 に

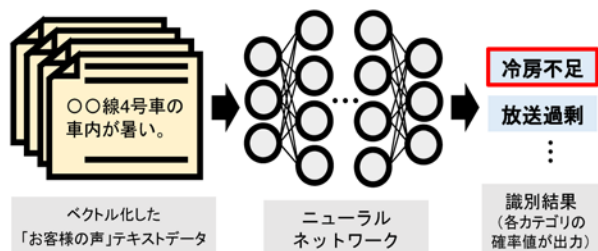


図6 カテゴリ確率を出力するニューラルネットワーク

表2 適用した MLP モデルの詳細

パラメータ	値	説明
Activation	ReLu	活性化関数の種類
max_iter	300	学習回数の上限
hidden_layer	100	隠れ層のユニット数
batch_size	Auto	ミニバッチサイズ
solver	adam	最適化のアルゴリズム
learning_rate_init	0.0005	初期学習率

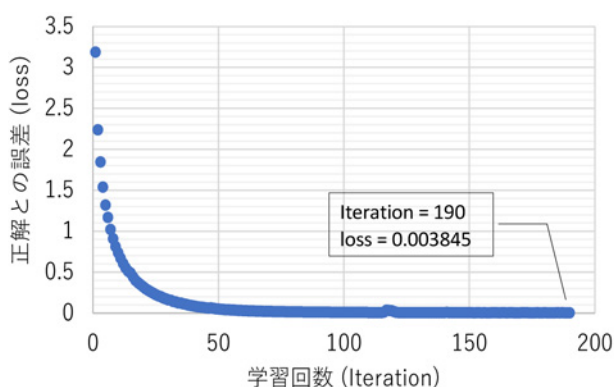


図7 適用した MLP の学習曲線

予測精度による降順で上位3位までの分類結果を示した。表中の括弧内の数字はそのラベル確率を表す。ここで色のついてない行は、AIが推定した分類ラベルの1位と正解が一致したことを示す。一方赤色の網掛けは、AIが推定した1位のラベルが正解ではなかったことを示す。結果的に何位のラベルが正解であったかを分かり易くするために、正解ラベルを黄色の網掛けで示した。

AIが推定した1位のラベルが正解とならなかったケースをみると、1位確率は0.9以下で、2位、3位もある程度高い数値となっていること、お客様が複数の話題

を話されている場合に間違える傾向があることがわかった。このようなマルチラベルへの対応として、複数の候補を併せて提示する方法が有効と考えた。

生成した学習モデルの精度を検証するために、Cross-Validation (交差検証) を行った。具体的には表4に示すように、6年分のデータ6299件の内5年分のデータ約5千数百件を使って学習したモデルを用いて、残り1年分のデータをテストした。学習とテストに使う年の組み合わせを変えながら計6パターンについて同様の作業を行い、得られた6つのモデルの正解率の平均を全体の正解率として評価した。

表4より、どの交差検証結果も、第1位が正解と一致した正解率は約67%、第3位まで含めると約86%の正解率となっていることから、モデルの正確性が示された。ただし表4の評価は、22個のカテゴリ全ての正解率を評価しているため、車両故障や水漏れのように、学習データが極端に少ないカテゴリは学習が十分でないため、これらの正解率の低さが全体の正解率を下げる形となった。お客様の声の多い空調系の声に絞って調べた結果を括弧内に示すが、2位までで91%、3位までを含めると95%の正解率が得られた。この結果からも、実用的なモデル作成のためには一つのカテゴリにつき最低でも300件程度は用意することが必要である。

表4 MLPによる交差検証結果 (%)
(T: 学習データ, V: テストデータ)

2015 991 件	2016 1215 件	2017 786 件	2018 1553 件	2019 1476 件	2020 278 件	第1位 正解率 (空調のみ)	第2位 正解率 (空調のみ)	第3位 正解率 (空調のみ)
V	T	T	T	T	T	57.3 (68.1)	74.2 (85.6)	81.1 (92.6)
T	V	T	T	T	T	68.1 (78.4)	81.6 (91.1)	87.0 (94.2)
T	T	V	T	T	T	74.8 (85.3)	85.6 (94.0)	90.6 (96.3)
T	T	T	V	T	T	68.4 (79.5)	81.6 (90.9)	86.9 (95.5)
T	T	T	T	V	T	69.4 (79.0)	83.4 (92.7)	88.6 (96.2)
T	T	T	T	T	V	64.7 (85.5)	75.2 (90.8)	83.1 (97.4)
平均						67.1 (79.3)	80.3 (90.9)	86.2 (95.4)

表3 MLPによるカテゴリ推定結果 (上位3位までの表示)

内容	分類(正解)	分類(推論)	1位	2位	3位
ラッシュ時の換気ならびにエアコン	冷房(不足)	冷房(不足)	冷房(不足)(0.999413)	冷房(過剰)(0.000398)	騒音・振動・衝撃(0.000154)
A駅～B駅間で自動放送の音	放送(過剰)	放送(過剰)	放送(過剰)(0.999883)	放送(不足)(0.000051)	騒音・振動・衝撃(0.000045)
C駅21:40発の列車に乗ってドア(車内外)	その他	その他	その他(0.448057)	車両故障(0.220252)	ドア(車内外)(0.208270)
ご回答ありがとうございます。D	暖房(不足)	暖房(不足)	暖房(不足)(0.982857)	冷房(過剰)(0.017143)	暖房(過剰)(0.000000)
JRの社員様こんにちは。いつも放送	その他	その他	その他(0.869661)	放送(過剰)(0.127872)	放送(不足)(0.002408)
XX系の列車に液晶ディスプレイ	その他	その他	その他(0.996568)	機能要望(0.002821)	新車・体改要望(0.000548)
2020年2月20日E駅19時21	放送(過剰)	放送(過剰)	放送(過剰)(1.000000)	騒音・振動・衝撃(0.000000)	放送(不足)(0.000000)
本日、F駅での事故で150分	暖房(不足)	暖房(不足)	暖房(不足)(0.996489)	暖房(過剰)(0.003511)	ドア(車内外)(0.000000)
ニュースで騒がれてるコロナ	その他	その他	冷房(不足)(0.688574)	その他(0.277208)	放送(不足)(0.002106)
車内の音声案内が小さかった	放送(不足)	放送(不足)	放送(不足)(0.949689)	放送(過剰)(0.023301)	暖房(不足)(0.012408)
JR以降に制作された電車で設	冷房(不足)	冷房(不足)	冷房(不足)(1.000000)	暖房(過剰)(0.000000)	その他(0.000000)
通勤等で、JRXY線を中心に利	座席・手すり・テーブル等	座席・手すり・テーブル等	座席・手すり・テーブル等(0.996)	機能要望(0.001746)	トイレ(0.001247)
2月22日 G駅 7時47分到着	放送(不足)	放送(不足)	放送(不足)(0.998357)	その他(0.001021)	暖房(不足)(0.000309)

3.4 文章のベクトル化手法の比較

本節では文章のベクトル化について考察する。文章のベクトル表現を得るため我々が最初に適用した Bag-of-Words という手法は、文書中の単語の出現頻度をベクトル化するシンプルなものであり、語順情報や文脈は考慮されない。一方、これとは違ったアプローチで単語や文章の分散表現を獲得する手法も提案されている。例えば Word2vec²⁾ では、文章中における単語の生起位置を学習し、各単語の分散表現を獲得する。また Doc2Vec³⁾ では、単語に加えて文脈を学習し、文章の分散表現を獲得する。今回空調に関するお客様の声データから、Bag-of-Words で得られた単語出現頻度のベクトルと Doc2Vec で得られた分散表現のベクトルを生成し、比較を行ったので結果を述べる。2016-2019年のデータから作成されたコーパスを用いて2015年のデータをベクトル化した。Bag-of-Words および Doc2Vec で得られたベクトルを大きさ1に正規化し、主成分分析による次元削減で3次元空間にプロット形式で可視化したものをそれぞれ図8(a)(b)に示す。プロットの色は文章のカテゴリ（冷房不足、冷房過剰、暖房不足、暖房過剰）を示す。図より、Bag-of-Words では全てのカテゴリのプロットが空間中の一箇所に凝集して分離できないのに対して、Doc2Vec の場合は、例えば「冷房不足」と「暖房不足」のプロットを分離可能な境界面を引けることが確認できる。このことから Doc2Vec では、Bag-of-Words と比べてより文書の性質に対応したベクトル表現を獲得できることが示された。

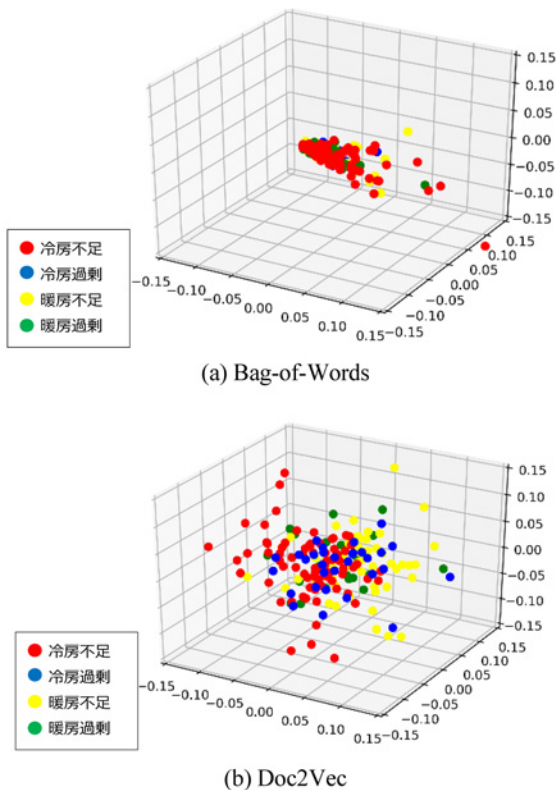


図8 文書ベクトルの可視化結果

4. 「お客様の声」と多様なデータとの結合

4.1 お客様の声マスタファイル

車両不具合の原因や予兆の把握といった更なる車両品質向上に向けて、5章で述べるお客様の声の可視化機能を備えた分析ツールを開発するため、お客様の声から時刻、駅、路線等のキーワードを自動抽出して列車を特定し、気象データ等の多様なデータと結合させたお客様の声マスタファイルを作成した。自動抽出したキーワードを表5に示す。McCabによる形態素解析を行い、キーワードごとに特定の語句を基準として前後の形態素の並びを参照した。例えば「発駅」や「着駅」であれば、駅名を基準の語句とした。さらに形態素の並び方が、事前に与えた特定のパターンに合致した場合、その語句を抽出した。例えば駅名の前に時刻、後ろに「から」「発」のような語句が含まれるパターンに合致する場合はその駅名を「発駅」として抽出した。

表5 自動抽出したキーワード

キーワード	用途
日付	列車特定、気象データ割り付けに使用
発時刻/着時刻	列車特定、気象データ割り付けに使用
発駅/着駅	列車特定、気象データ割り付けに使用
路線	列車特定に使用

4.2 列車特定アルゴリズム

抽出されたキーワードからお客様の声の中で言及されている列車を、計画ダイヤデータを利用して特定した。

「発駅」、「発時刻」、「着駅」、「着時刻」の4種類のデータの全てが抽出できていれば、列車を一意に特定できる可能性が高いが、実際には全てが抽出できるとは限らない。一方で、業務支援の観点で考えれば、該当する可能性のある列車を複数抽出できることが望ましい。

そこで、突き合わせの条件を、(A)「発駅」、「発時刻」が合致する列車を特定、(B)「着駅」、「着時刻」が合致する列車を特定、(C) (A)と(B)で特定された列車の和集合とした。また、抽出されたキーワードの「発時刻」と「着時刻」の単位が時分であるのに対し、計画ダイヤデータの発時刻と着時刻の単位は時分秒である。そのため、時刻が合致しているかを判定する際には、キーワードの発時刻と着時刻に60秒の時間幅を持たせている。

さらに、「発駅」と「着駅」がキーワード抽出できている場合には、特定列車の方面と、「発駅」と「着駅」から判断する方面とが合致している列車のみを出力した。

4.3 気象データ割り付け

お客様の声の発生状況を把握するため、気象データベースを参照して、発生時の気温・湿度・日射量・降水

量などの情報を得られるようにした。キーワード抽出により得られたお客様の声の発生地点・発生時刻に近い観測所のデータを割り付ける。ただし、発生地点に近い観測所の観測項目が少ないことがあるため、観測項目が多い観測所のデータも割り付けることとした。また、発生地点や発生時刻が定かでないお客様の声も存在するため、朝・昼・夜の3つの時刻における当該路線中心地の観測所のデータも割り付けることとした。

5. 車両品質向上に資する分析ツールの開発

5.1 お客様の声マスタファイルの可視化

お客様の声分析管理システムにより、お客様の声マスタファイル上のデータをプロットすることで、お客様の声の発生状況を可視化できるようにした。お客様の声を気温と湿度に対してプロットした例を図9に示す。気温10～20℃の範囲では、冷房と暖房に関するお客様の声の混在している。気温25℃を超えると冷房不足の声の方が多くなる。一方で、湿度とお客様の声との特段の関連性は認められないことが分かる。

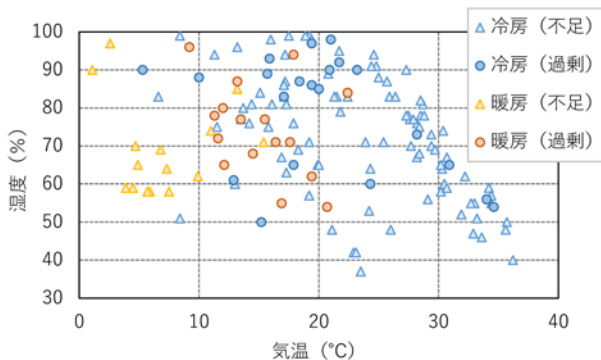


図9 気温と湿度に対するお客様の声の発生状況

5.2 車両モニタデータの活用

車両情報記録装置は、車両のデジタル伝送装置から取得した営業走行時における車両モニタデータを収集する装置である。携帯電話回線などを用いて、営業走行中のデータを地上サーバに伝送することでデータ取得を行う。データ収集のイメージを図10に示す。車両情報記録装置の大規模データ分析処理支援システム⁴⁾を用いて、駅間及び編成内で平均や最大・最小を取った値を算出して、ダイヤ図表示システムにて表示する。

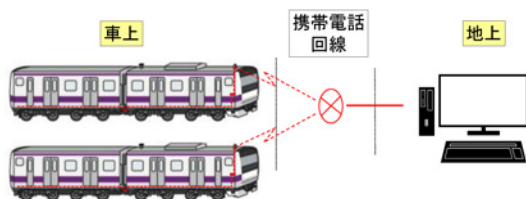


図10 車両情報記録装置によるデータ収集イメージ

5.3 ダイヤ図表示

今回対象とする路線では、車両情報記録装置を搭載した編成は少数であり、かつ限定運用されていないため、日々の車両情報記録装置搭載編成の運用を踏まえた分析が必要となる。このため、ダイヤ図表示システムを構築して、お客様の声と車両モニタデータの照合の支援を行えるようにした。

ダイヤ図表示システムでは、指定した路線及び日付に対してダイヤ図形式で情報を表示する(図11)。輸送データベースに基づき全列車の計画ダイヤが黒色で表示される。列車特定機能によって列車を特定できたお客様の声については、当該列車が灰色で強調表示され、マウスを近づけるとその内容がポップアップ表示される。また、車両モニタデータは、データが取得できた列車のみを対象に状態値が色表示される。表示する状態値の項目は様々であり、乗車率、車内温度、車内湿度、車内不快指数などが挙げられる。また、車両モニタデータの気温・湿度等の情報から温熱指標を表示することもできる。当該指標は、鉄道車両内での被験者実験により統計的に導出された、車内温熱環境に対して暑くて/寒くて/不満足と感じる乗客割合の予測値である⁵⁾。

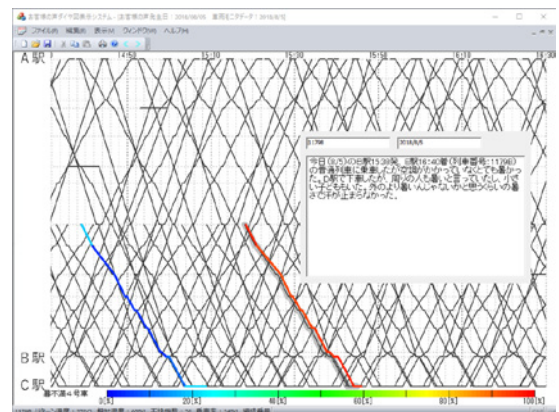


図11 お客様の声と車両モニタデータの対応事例

5.3.1 お客様の声と車両モニタデータの対応事例

図11では、車両情報記録装置を搭載した編成で空調機器の不具合のためお客様の声が発生した事例を示している。車両モニタデータの状態値として空調機器の不具合が発生した当該号車の温熱指標として暑くて不満足と感じる乗客割合の予測値を示した例である。お客様の声が発生している当該編成は、他の編成に比べて不満の指標が高くなっていることが確認できる。このように、お客様の声の発生原因を車両モニタデータと対照して分析することができる。

5.3.2 類似気象選択による分析事例

車両情報記録装置を搭載していない編成でお客様の声が発生した事例を図12に示す。状態値として温熱指標を用い、暑くて不満足と感じる乗客割合の予測値について編成内の平均値を示している。お客様の声が発生した

当日の列車は、車両情報記録装置を搭載していない編成のため、直接的に車両状況を把握することは難しい。そこで、このようなケースに対しては類似気象選択機能を用意した。類似気象選択機能では、気象データベースから対象路線に近い観測所の気温や湿度等のデータを取得して、当日の気象に近い日をリストアップして表示する。類似気象選択を行った事例を図13に示す。この事例では、順位5位で選定された日には、当該列車に車両情報記録装置搭載の編成が運用されており、情報が取得できる。X駅に近づくにつれて、不満足と感じる温熱指標が上昇していることがわかる。

このときの車両モニタデータの時間推移を分析した例を図14に示す。乗車率の増加によって、不満足と感じる温熱指標が上昇しているものと考えられる。このように、

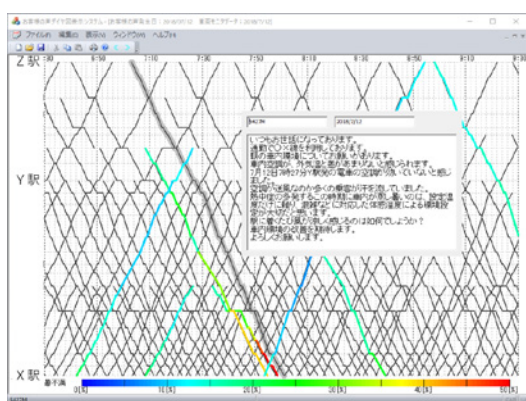


図12 お客様の声の発生日の表示例

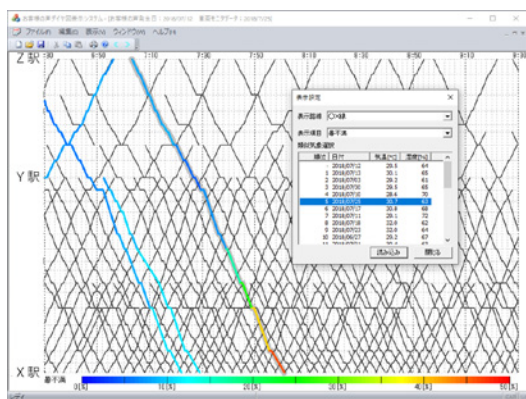


図13 類似気象日の表示例

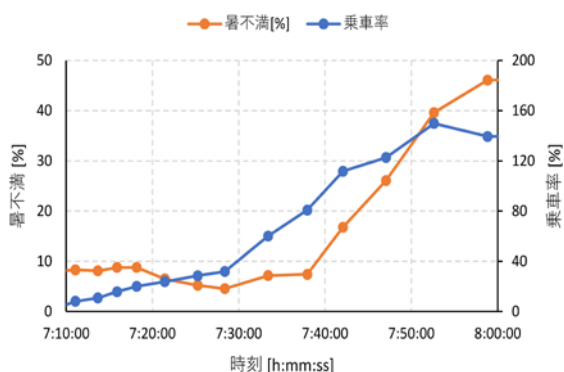


図14 車両モニタデータの時間推移の分析例

特定編成に限らずに、当該路線の当該列車で一般的に生じるお客様の声に対しては、類似気象選択機能によって、車両モニタデータとの対照による分析が可能である。

6. まとめ

「お客様の声」を冷房不足や放送過剰などのカテゴリに自動分類するニューラルネットワークモデルを開発した。複数の内容が含まれるマルチカテゴリの声にも対応できるように、判定結果を複数提示するようにした。6千数百件のお客様の声で検証実験を行った結果、凡そ90%以上の正解率で自動分類できることを確認した。さらに文書のベクトル化方法についても比較し考察を加えた。

開発したニューラルネットワークを業務に活用することを考えた場合、AIが1位と判定したカテゴリの確度が90%を超えた場合は、当該カテゴリが正解となっていることから、7割の声はモデルに分類を任せ、残りの3割を人が分類するとした場合、工数は7割削減が可能となる。残りの3割についても、3位までの候補から選ぶことで分類作業の負担軽減につながる。

次に、車両不具合の原因や予兆の把握といった更なる車両品質向上に向けて、列車特定に必要な情報とテキストから抽出される情報との関係を整理し、列車を特定するツールを作成した。車両情報記録装置のデータなども含めた多様なデータを分析する方法やツールを検討し、これら有効なデータを集積したお客様の声マスタファイルを生成した。さらに、本マスタデータの条件を絞りこんでの可視化、ダイヤ図表示などの機能を備えたお客様の声分析管理システムを開発し、お客様の声の活用性を示した。

今後は現時点で適用できる条件や前提を明示した上で、空調以外のカテゴリへの対応や他のデータとの連携など、実用に向けた検討も進めていく。

文献

- 1) 工藤拓, 形態素解析の理論と実装, 近代科学社
- 2) T. Nikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", NIPS2013, 3111, 2013.
- 3) Q. Le and T. Nikolov, "Distributed Representations of Sentences and Documents", CoRR, abs/1405.4053, 1-9, 2014.
- 4) 小川知行, 武内陽子, 菅野善, 影山真佐富, 東海勝人, 平松英: 車両情報記録装置の大規模データ分析処理支援システムの開発, 第22回鉄道技術連合シンポジウム(J-RAIL2015), SS3-1, 1705, 2015
- 5) 遠藤広晴, 伊積康彦, 林伸明: 夏季の通勤列車内の温熱快適性予測手法, 鉄道総研報告, Vol.29, No.7, pp.27-32, 2015