E-003

文意系列を考慮したコールセンターオペレータの発話の文意予測
Prediction of sentence meaning of call center operator's speech based on sentence meaning sequence

川北 海斗† 今井 貴史‡ 市川 治†
Kaito Kawakita Takashi Imai Osamu Ichikawa

1. はじめに

多くのコールセンターでは、応対品質の向上を目指し、オペレータの応対品質について評価・指導を実施している。応対品質は様々な要素に分解できる。その一つに、応対を正しい順序で提供できているかということがある。本報告では、表 1 に文意 A として示す分類を想定している。例えば、一つのセッションは「挨拶」から始まり「受け止め」「確認」「案内・説明」を繰り返し、再び「挨拶」で終わるといった典型的なパターンがあり、そこから外れていれば詳しくチェックする必要があるだろう。また、オペレータの声の印象という要素も重要である。我々は前報[1]にて、オペレータの声の印象を評価・採点する機械学習モデルを提唱した。前報では考慮しなかったが、状況に相応しい声のトーンがあると考えられる。本報告ではその状況として表 2 に示す文意 B を想定する。例えば、文意 B「明るさ・前向き」が期待される状況の時に暗いトーンで発声すれば、声の印象の評点は低く判定すべきである。

本報告では、コールセンターオペレータの発話の書き起こしテキストデータを機械学習することにより、その文意 A 及び文意 B を推定することを目標とする。

文意A	発話の例
①挨拶	• お電話ありがとうございます3でございます • 失礼いたします • はいお電話お待ちしております
②受け止め	• はい • かしこまりました • そうですね
③確認	• はいどのようなご用件でしょうか? • 以上でよろしいでしょうか? • では今ちょっと確認しますのでお待ちいただけますでしょうか?
④案内・説明	• こちらは少々簡易的なテーブルになるので • 少々お待ちくださいませ • はいでは一度ご検討いただけますようお願いいたします

表 1　文意 A

文意B	発話の例
①明るさ・前向き	• お電話ありがとうございます3でございます
②共感・心配	• あとは何かご不明点ございませんでしょうか
③責任・慎重	• それでは続きまして選考会のご案内をさせていただきます
④謝罪・丁重	• 申し訳ございません • はいお待たせいたしました。

表 2　文意 B

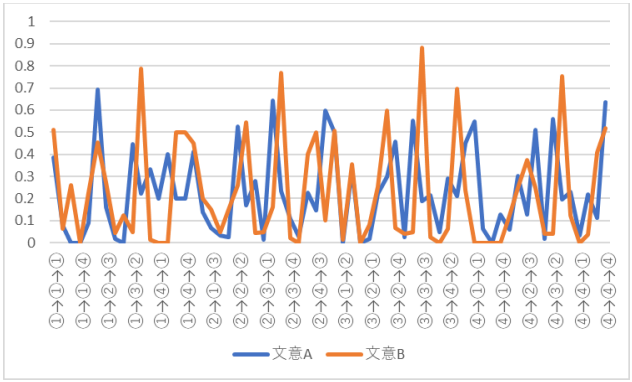

図 1　3-gram の値

2. 従来技術

メールやチャットにより顧客対応を行うコンタクトセンターを対象分野とした会話文を構造化する技術として[2]の研究がある。この研究では各会話を、正規表現パターンを用いてラベルを付与する。その後同一層のノード群に対し、ラベルに基づきクラスタリングを行うことで、会話の流れを保持した形で会話文を木構造に整理している。

本研究では機械学習を用いて各会話のラベルの自動推定を行い、また推定するラベルである「文意の系列」や「文脈」を考慮する手法を提案する。

3. 文意の系列

コールセンターでのオペレータと顧客の会話テキストのうち、オペレータのみの会話テキストに割り振られている文意 A（①挨拶・②受け止め・③確認・④案内・説明）と文意 B（①明るさ・前向き・②共感・心配・③責任・慎重・④謝罪・丁重）というラベルを使用して、文意の系列、つまり、ある文意がどの文意に遷移しやすいか、また遷移しにくいかについて 3-gram を用いて調べた。

本研究においての 3-gram の値とは、2 件前の会話テキストの文意→1 件前の会話テキストの文意→現在の会話テキストの文意の順で移り変わる確率を指す。

図 1 に示す 3-gram の値を見ると、文意 A において多く見られたパターンは、文意②→文意④や文意④→文意②を含む遷移であることから、カスタマーの会話内容を受け止めた後に次の案内・説明に入ることや、案内・説明に対してのカスタマーからの返事を受け止めるといったことが多く、一方で文意①→文意③を含む遷移パターンが少ないことから、挨拶の後すぐに何かを確認することは少ないということが分かる。

文意 B においては、文意③へと遷移する流れが多く、明るさ・前向きなどの明るいトーンでの会話であっても、責任・慎重など真面目なトーンの会話へと移っていくことが

† 滋賀大学データサイエンス学部　Faculty of Data Science, Shiga University
‡ 滋賀大学データサイエンス・AI イノベーション研究推進センター　Data Science and AI Innovation Research Promotion Center, Shiga University

多いと分かる。少ないパターンとしては文意④→文意①を含む遷移が少なく、暗いトーンの後すぐに明るいトーンの会話に変わるということは少ないことが分かった。

4. 提案法

提案手法の 1 つ目は文意系列を考慮したビタビアルゴリズムによる文意推定手法、2 つ目は過去発話を用いて、Universal Sentence Encoder（以下 USE）[4]による分散表現を入力とした深層学習による文意推定手法を提案する。

なお、モデル学習にあたっては、表 3 に示すように、文意ごとのデータ数が不均衡であることに留意が必要である。以下のモデルの学習にあたっては、学習の際の損失関数に分類クラスごとの重みを掛けることで不均衡対策を行っている。

データの総数は3358件で、学習データ：テストデータの割合が2：1になるようにデータを分割し、分割の仕方を変えて10回学習を行った。また、オペレータとカスタマーの会話の始まりから終わりまでの 1 セットのことをセッションと呼び、総セッション数は 152 である。

4.1 文意系列を考慮したビタビアルゴリズム

まずベースラインとして、文意を推定したいテキストを word2vec にかけ、得られた分散表現を機械学習手法である LSTM に入力することで「会話テキスト」のみから文意 A 及び B を推定するモデルを作成する。

次に文意系列を考慮したモデルを作成する。具体的には隠れマルコフモデルを併用し、その出現確率にはベースラインの LSTM モデル、遷移確率には前述の 3-gram の値を用いる。このモデルにビタビアルゴリズムを適用すれば、「会話テキストがどの文意に適合するか」と「ある文意がどの文意に遷移しやすいか」という 2 つの観点からベストな文意系列を推定することが出来る。

4.1.1 LSTM モデル（ベースライン）

文意を推定したい現在のオペレータの会話テキストのみを word2vec にかけ分散表現を獲得し、その値を図 2 に示す LSTM に入力することで、その会話テキストが各文意に帰属する確率を求め、最も確率の高い文意を結果として出力する推定モデルを作成した。学習の際の損失関数には交差エントロピー、活性化関数には softmax 関数を導入した。

得られた結果をまとめたものが表 4,5 である。文意 A 推定モデルでは高い精度を出せているが、文意 B 推定モデルでは文意 A 推定モデルと比べると低い精度となっている。

原因としては、文意 B では「明るさ・前向き」や「共感・心配」から分かるように声のトーンのようなテキストからは判別しにくいものを表しており、「挨拶」・「受け止め」などの会話文の役割のようなものを表す文意 A と比べると推定することが難しいと考えられる。そこで文意系列について計算した 3-gram の値を用いて、「ある文意がどの文意に遷移しやすいか」ということも考慮することで精度の向上を図った。

	文意①	文意②	文意③	文意④
文意A	410	1142	487	1319
文意B	609	192	2382	175

表 3　文意ごとのデータ数

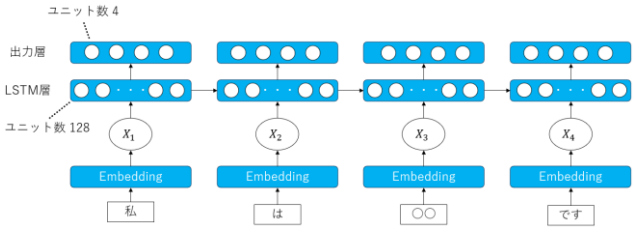

図 2　LSTM モデル（ベースライン）の構成

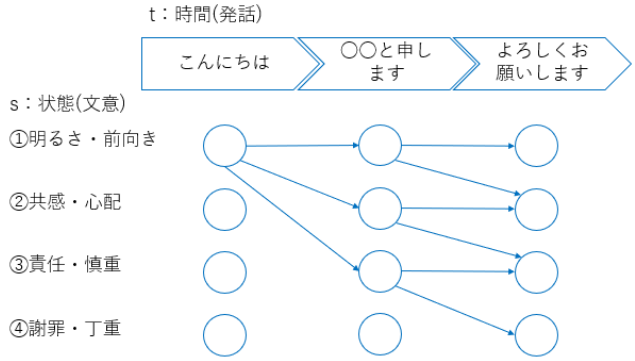

図 3　HMM のトレリス

4.1.2 隠れマルコフモデル＋LSTM モデル

文意系列について計算した 3-gram の値を状態遷移確率、LSTM モデルから取り出した会話テキストが各文意に帰属する事後確率を出現確率とするビタビアルゴリズムによる文意系列推定モデルを作成した。

今回の隠れマルコフモデルのトレリスを図 3 に示す。図に示すように遷移の矢印は一つ前の状態からの遷移であるので、遷移確率は 2-gram を用いるのが自然である。今回は 3-gram を用いて精度の向上を図りたかったので、簡易的な拡張を行った。すなわち、遷移してくる 1 つ前の状態においてもビタビアルゴリズムによるパス選択が行われているので、その情報をそのまま採用して 2 つ前の状態とした。1 つ前と 2 つ前の状態を仮定できれば 3-gram を使用することが出来る。

会話の最初の会話テキストを 1 番目の会話テキストとすると、t 番目の会話テキストがある文意 s に存在する確率を $\alpha(s, t)$ と表す。次に、t 番目の会話テキストの分散表現を LSTM モデルに入力した時に得られた、その会話テキストがある文意 s と推定される確率を $a(s, t)$ と表す。そして、会話テキストの文意 s_1 から文意 s へと遷移するとすると、状態遷移確率は $b(s_1,s)$ と表される。これには文意系列の 2-gram の値を設定すればよい。

2-gram の場合、1 番目の会話テキストから最後の会話テキストまで順に、

$$\alpha(s,t) = \max_{s_1}\left(\alpha(s_1,t-1) \times a(s,t) \times b(s_1,s)\right) \quad (1)$$

という計算式で$\alpha(s, t)$を計算すればよい。最終ノードで最も確率の高い文意に到達するベストパス推定することで、各発話における文意を決定する。なお、スタートノードとなる第 1 発話の確率$\alpha(s, 1)$には s についての 1-gram 確率の値を設定する。

これを 3-gram に簡易的に拡張するにあたって、$b(s_1,s)$を$b(s_2,s_1,s)$に置き換える。s_2 は、2 発話前の会話テキストの文章であるが、$\alpha(s, t-1)$を算出する際に(1)式の max 操作で文意が選択されているので、それを s_2 として採用することにした。

表 4,5 の結果を見ると、総合の正解率では文意 A、B ともにベースラインよりも高い精度が出せている。しかし、文意 A の④案内・説明や文意 B の④謝罪・了重などでは再現率が低下しており、学習されていない 3-gram の値が非常に小さい値（=スムージング値）となっている影響ではないかと考えている。

4.2 USE による分散表現を入力とした深層学習

この節では、USE を用いた改善を試みる。まず、4.2.1 節では文意を推定したいテキストを USE にかけ分散表現を獲得し、その値を入力とした深層学習による推定モデルを作成した。ここでは先行する発話は考慮しない。次に、4.2.2 節では現在のオペレータの会話テキストだけでなく、カスタマーの会話テキストなど過去の会話テキストを用いることで「文脈」を考慮した推定モデルの作成を行った。

深層学習の損失関数としては、文意 A 推定モデルではフォーカルロスを導入した。文意 B 推定モデルでは、ラベル付けに曖昧さがあることから自己適応型学習を用い、その際にソフトラベルを扱うことになる。正解がソフトラベルで与えられている場合、フォーカルロスでは、モデルの出力が正解と厳密に同じであっても勾配がゼロにならない。この点を考慮し、文意 B 推定モデルでは、損失関数に交差エントロピーを導入した。

4.2.1 USE 入力深層学習モデル（単発話入力）

図 4 に示すように、文意を推定したい現在のオペレータの会話テキストのみを USE にかけ分散表現を獲得し、深層学習により、その会話テキストが各文意に帰属する確率を求め、最も確率の高い文意を結果として出力する推定モデルを作成した。

表 4,5 を見ると、文意 A 推定モデルでは高い精度が出せている。文意 B 推定モデルでは 4.1 節の手法と比べると、再現率などの向上が見られたが、本手法の文意 A 推定モデルと比較すると低い精度となっている。その原因としては、4.1.1 節でも述べたように、当該テキストのみを用いて推定を行っていることで文脈を考慮していないことが挙げられる。そこで 4.1 節の手法と同様に本手法においても文脈を考慮することで、推定精度の向上を図る。

4.2.2 USE 入力深層学習モデル（複数発話入力）

図 5 に示すように、文意を推定したい現在のオペレータの会話テキストに加え、カスタマーの会話テキストを直前 3 件まで取得し、またオペレータの会話テキストも直前 3 件まで取得する。そして、（当該テキストの分散表現、カスタマーの k 件前会話テキストの分散表現、オペレータの

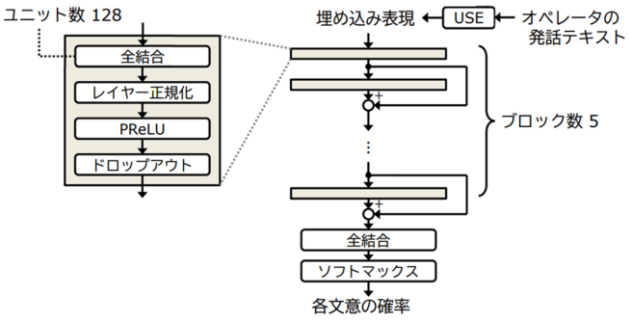

図 4　USE 入力深層学習モデル（単発話入力）の構成

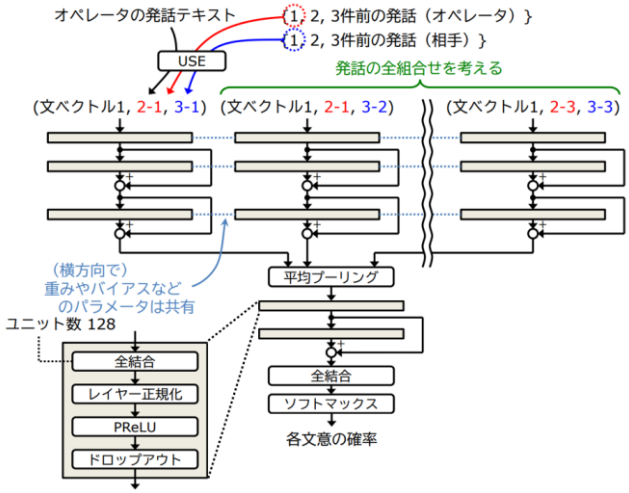

図 5　USE 入力深層学習モデル（複数発話入力）の構成

	正解率	再現率			
		文意①	文意②	文意③	文意④
ベースライン	0.893	0.833	0.931	0.834	0.901
HMM＋LSTM	0.913	0.87	0.959	0.991	0.333
USE	0.886	0.86	0.919	0.862	0.874
USE(複数発話)	0.895	0.884	0.92	0.846	0.895

表 4　文意 A 推定モデルの精度

	正解率	再現率			
		文意①	文意②	文意③	文意④
ベースライン	0.747	0.506	0.045	0.901	0.267
HMM＋LSTM	0.753	0.527	0	0.939	0.102
USE	0.703	0.586	0.34	0.781	0.447
USE(複数発話)	0.812	0.749	0.298	0.897	0.436

表 5　文意 B 推定モデルの精度

n 件前会話テキストの分散表現）の組み合わせを k, n = 1, 2, 3 の場合の全 9 通りを考え、それぞれを同じ全結合層に入力する。

これにより、出力として 9 個のベクトルが得られ、その中のいくつかは「文脈」を考慮した表現になっていることが期待される。そこで、それら 9 個のベクトルの平均を取り、得られた値を後続の多層全結合層に入力して各文意の確率を得る。

結果としては、4.2.1 節のモデルから十分高い精度を出していた文意 A 推定モデルでは、本手法においても高い精度を示した。文意 B 推定モデルでは文意②④において再現率の低下が少し見られたが、文意①③では再現率が大きく向上し、全体としての正解率は大幅に向上した。

5. おわりに

本報告では、コールセンターの品質管理に用いる文意ラベル A、B を想定し、書き起こしテキストを入力とする機械学習により推定することを試みた。1 つの当該発話のみを入力する手法では性能に限界があることを示し、文脈や文意系列を考慮する手法を提案した。

文意 A は会話文の内容からその文章がどのような役割を持っているのかを表している文意であるので、どの手法でも高い精度の推定を行うことができる。その中でも、1 つのセッションの最初の発話から最後の発話に至る全体の文意系列を考慮するビタビアルゴリズムが最も高い性能を示した。

一方で、文意 B については表現すべき感情を表す文意であり、曖昧性が大きい。特に本文中②④で示した文意は、本当に正解のラベルであったのか見極めるのが難しい。（そのため、USE 入力モデルでは自己適応型学習を適用した。）曖昧さがある中でも、文意系列や過去の発話（文脈）を考慮することによって、推定精度の向上を達成した。

今回は、文意 A、B について、それぞれ独立にモデルを構築したが、2 つの文意は相互に関連を持っていることも考えられるので、今後、2 つを同時に推定するモデルを構築することも検討したい。

謝辞

ビーウィズ（株）様からは、コールセンターデータをご提供いただくとともに、多くの知見とご示唆をいただいた。本研究は科研費（19K02999）の助成を受けた。

参考文献

[1] 今井貴史,村木友子,市川治,”コールセンターの応対音声を対象とした声の印象の自動評価”, 情報処理学会全国大会講演論文集, Vol. 84, No.2, pp. 2.29-2.30 (2022)

[2] 星見綾子,細見格,”句構造解析とクラスタリングを用いた会話履歴の要約”, 人工知能学会全国大会論文集 Vol.32, pp.ROMBUNNO.2K1.02 (2018)

[3] 松田雛乃,松井源,”N-gram を用いた日記文章の文章構造の分析”, 情報知識学会誌,Vol. 31, No. 2 (2021)

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil, “Universal Sentence Encoder”, arXiv:1803.11175v2 [cs.CL] 12 Apr (2018)