コールセンターにおける会話マイニング

那須川哲哉　　宅間大介　　竹内広宜　　荻野 紫穂
日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

問い合わせや受注などの顧客サポート及び営業活動を含む顧客への対応業務を電話で行うコールセンターにおいては、個々の対応の概要をテキスト化して蓄積することが多い。その膨大な対応記録を活かすため、テキストマイニングを導入することで、各記録に記述された商品や苦情・要望などの表現における相関や増減傾向を把握し、問題の早期発見やコールセンターの効率化を実現する事例が増えてきている[2]。

このような状況において、テキストマイニングの新規導入や拡張を検討する上では、分析データとして、コールセンター側の担当者が記述した概要のテキストではなく、顧客と担当者の生の会話そのもののテキストを対象としたいという要望が高まっている。人手を介することなく自動音声認識によって生の会話をテキスト化し、分析することが可能になれば、担当者は、概要をテキスト化する作業から解放され、顧客対応に注力することができ、生産性や顧客満足度の向上が期待される。また、生の会話を分析できれば、通常の概要には反映されない詳細な顧客の声に加え、担当者自身が敢えて記述することの少ない、不適切な対応まで把握できるようになることが期待されている。

ところが、実際に生の会話のテキストを既存のテキストマイニングシステムで分析して役に立つかどうかは自明ではない。

概要をテキスト化したデータは、通常、入力の作業量を少なくするという観点から個々のデータ量が少なく、基本的には業務に関連した内容を中心に構成されている。実際の PC ヘルプセンターの事例[3]においても、一件の問い合わせ記録におけるキーワード数は平均 16.5 語と、同じ基準で比較した新聞記事の半分程度である。さらに、一ヶ月分約 4 万 2 千件のデータにおいて、全体から抽出されるキーワードの 8 割近くが、高頻度の異なり語の約 2000 語でカバーされており、記述された内容は限定された業務領域に偏っていることが確認されている。このように限定された内容の比較的短いデータにおいては、分析対象とすべき語が限定され、多義語の問題も少ない。そのため、分析用辞書の構築に膨大な労力をかけずとも、有効なマイニング結果を得られる場合が多い。

それに対し、生の会話全てをテキスト化したデータの量は、概要をテキスト化したデータの量に比べて確実に多く、業務の本題とは直接関係の無い内容をより多く含むと考えられる。例えば、挨拶の表現や、住所や氏名の漢字表記を説明するための表現は、通常は業務対象の内容と関連が無い可能性が高い。従って、分析に無関係な記述がノイズとなって、有益な分析結果を得るのが困難となる可能性が考えられる。また、生の会話をテキスト化するための自動音声認識の技術も現状では完全でなく、100％の認識精度は期待できない。そこで、生の会話の分析によって有効な知見が得られるとしても、音声認識のエラーによるノイズが含まれたテキストで同じ知見が導けるようでなければ、実用には適さない。

以上の観点から、我々は、実際のコールセンターにおける生の会話約千件を人手で全てテキスト化し、そのデータを既存のテキストマイニング技術で分析することで、実際に有益な知見が得られるかを調査した。さらに、そのデータに人工的にノイズを加え、ノイズの割合がどの程度の範囲であれば元のデータから得られる知見が維持されるか、実際に自動音声認識でテキスト化したデータの分析でも同じような結果が得られるかを調査した。

本稿では、テキスト化された生の会話をテキストマイニング技術で分析することを会話マイニングと記す。

2. 会話マイニングの有効性

生の会話のデータとして、レンタカー業務のコールセンターにおける米語の会話 936 件を全て人手で書き起こしテキスト化した上で、テキストマイニングシステム IBM TAKMI[4]で分析を行った。主に予約業務に関する顧客と担当者との対話であり、一件のデータは平均して 262.4 語の単語から構成され、その中には平均して 17.5 往復のやり取りが含まれている。このデータには多くのスペルミスが含まれており、書き起こしの精度は決して高くないが、少なくとも人が読んで会話の内容を把握することが可能である。また、各データにはテキスト以外の付随情報として

- 会話の結果としての業務成果
  - 予約に至ったか
  - 予約に至らなかった場合の理由
  - 予約に至った場合実際に車が貸し出されたか
- 担当者の情報
  - チーム名
  - 成績のランク

といった定型情報が含まれている。

テキスト部分を分析する上では、レンタカー業務向けの辞書を構築し、車のタイプや各種サービス名、支払方法などに関する表現を抽出して分析対象とした。文法的な解析には[5]を用い、品詞と基本形の同定を行なった上で、辞書に登録されている表現を正規化して抽出した。例えば、複合名詞の *child seat* に〔装備〕の属性をつけたり、*NY* を〔地名〕属性の New York に正規化したりする情報が辞書に定義されている。

以上の枠組みを用いてテキストから抽出された内容と、定型情報における予約結果などの相関を分析することで、

- 予約に至らない場合の特徴的な理由
- 予約された車が実際に貸し出される会話の特徴

など、業務上有益な知見を導き出すことができた。

一例として、今回分析対象としたデータ 936 件の中で、予約に至らなかったケース 461 件のうち、その理由として『要望に応えられなかった』という定型情報が付随している 42 件のデータの特徴の分析を示す。予約に至らない理由としては、顧客が他社との比較のために価格のみを問い合わせてきたケースや、車が用意できなかったケースと並んで、この『要望に応えられなかった』というケースが件数の上位を占めている。ここで具体的にどのような要望に応えられなかったかを分析するには、実際の会話内容を見る必要がある。そこで、要望に応えられなかった 42 件のデータにおける、顧客の発言の中で、頻出しているキーワードの上位を見ると図1のような結果が得られた。

Keywords	Frequency
car	26
do	22
have	21
thank	20
card	20
Ok	19
airport	18

図 1: 顧客の要望に応えられず予約に至らなかった 42 件のデータ中、顧客の発言で頻出しているキーワード

図1におけるキーワードは、辞書登録した語に限らず、名詞・動詞・形容詞およびその複合語からなっており、どれもレンタカー業務に関連した一般的な表現であるため、この結果から有益な知見を導き出すのは困難である。しかし、この並べ方の基準を相関の強さにすると図 2 のようになる。

Keywords	Frequency	Correlation
debit card	13	5.3
debit	14	5.0
card	20	2.9
credit card	9	2.7
credit	10	2.5
use	8	1.3
live	4	1.3

図 2: 顧客の要望に応えられず予約に至らなかった 42 件のデータに相関の高い、顧客発言中のキーワード

図 2 でリストされているキーワードと、その対象となっている 42 件のデータに対する相関は、各キーワードを含む文書集合を A、42 件の対象文書集合を B として、以下の値に補正を行った値を指標としている。 ここで D は 936 件の全文書集合、# は文書集合中の文書数を表す。

$$\frac{\#(A \cap B)/\#D}{(\#A/\#D)\,(\#B/\#D)}$$

これは、基本的には、分析対象となる文書集合 B において A が出現する割合と、全体文書集合において A が出現する割合との比を取ったものであり、一を超える値であれば相関が強いということになる。補正は、#(A∩B)、#A、#B が小さい場合に偶然による相関が検出されるのを防ぐために、D を無限個の文書からのサンプリングとみなし、#(A∩B)／#D、#A／#D、#B／#D を用いて真の文書密度を区間推定する。実際の計算には、#(A∩B)／#D、#A／#D、#B／#D から推定された区間のそれぞれ、左端(最小値)、右端(最大値)、右端(最大値)の値を用いている。

このようにして得られる相関の強いキーワード debit card を含むデータ 13 件に着目することで、debit card の利用に関する制約によって予約に至らないケースが多いことが分かった。

また、予約に至ったケースでも、予約された車を手配した結果として、実際に貸し出されない限り売り上げにはつながらず、手配した車を顧客が取りに来なければ収益を下げることになってしまう。従って、レンタカー業務においては、予約した車を顧客が実際に取りに来る割合を高めることが重要である。そこで、予約した車を取りに来た会話と取りに来なかった会話の特徴を分析した結果、特徴に違いが見出され、そこから得られた知見を基に、新しい顧客対応ガイドを作成した。この新しいガイドに沿って顧客対応を行った担当者チームの売上高が向上したことから、新しいガイドが現在では全体で採用されている。

以上の通り、会話データに含まれるキーワードの分布が、予約に至ったか否かという会話の業務結果に応じて、どのように偏っているかを分析することで有益な知見を得ることができ、会話マイニングの有効性が確認された。

3. ノイズを含む会話データの活用可能性

前述した会話データは人手で書き起こしたものであり、その作成には高いコストが必要となる。また、人手による書き起こしにもエラーは付き物であり、前節のような分析を実現する上での書き起こしに100%近い精度を前提とするのは現実的ではない。そこで、会話テキストのノイズ（書き起こしエラー）がどの程度の割合までなら、前節のような分析結果を得ることができるかを調査した。ノイズとしては、人工的なノイズを多段階で加えた場合と、自動音声認識によるエラーが発生した場合の二通りに関して実験を行った。

3.1. 人工的なノイズ

まず、ノイズ混入の割合がどの程度までなら有効な分析結果が得られるかを調査するため、人工的にノイズを混入したデータを作成した。実際に発生するノイズはアプリケーションやその実装方法に依存するため、特定システムのノイズを模倣すると一般性を失う恐れがある。そこで、単純にランダム関数を用いたノイズの混入を行うことにした。

具体的には、10%から 100%まで 10%刻みの割合で、原文中からランダムに選択した語をノイズ語リストからランダムに選択した語と置き換えた。ノイズ語リストからの置き換えにおいては、

A) 自動音声認識システムの辞書(異なり語数 34,133 語)からランダムに選択

表1: 顧客要望に応えられず予約に至らなかったデータに対する debit card 及び debit の相関の強さと出現頻度

	自動音声認識システムの辞書から選択した語との置換によるノイズ		書き起こしデータ中の出現分布を考慮して置換した語によるノイズ		書き起こしデータ中で出現頻度3以上の語と置換した語によるノイズ	
ノイズの割合	*debit card* の相関強度(頻度)	*debit* の相関強度(頻度)	*debit card* の相関強度(頻度)	*debit* の相関強度(頻度)	*debit card* の相関強度(頻度)	*debit* の相関強度(頻度)
0 %	5.3 (13)	5.0 (14)	5.3 (13)	5.0 (14)	5.3 (13)	5.0 (14)
10 %	5.8 (13)	5.3 (14)	5.1 (12)	4.2 (14)	4.7 (11)	4.1 (13)
20 %	4.5 (10)	4.9 (13)	4.3 (10)	4.1 (13)	4.2 (9)	4.0 (14)
30 %	4.4 (8)	4.6 (12)	4.4 (9)	3.5 (13)	3.5 (8)	3.8 (13)
40 %	2.8 (6)	5.3 (13)	2.8 (6)	2.5 (12)	1.4 (4)	2.3 (8)
50 %	1.2 (3)	2.7 (7)	3.3 (6)	2.5 (11)	4.3 (7)	3.6 (12)
60 %	0.0 (1)	3.0 (7)	0.1 (1)	1.3 (8)	2.4 (3)	2.2 (8)
70 %	-	0.7 (3)	1.9 (3)	1.8 (10)	-	1.1 (6)

B) 対象データである 936 件に出現する語(異なり語数 10,661 語)の出現分布が変化しないよう、出現頻度を考慮する方法でランダムに選択

C) 対象データである 936 件中に出現する異なり語のうち 3 回以上出現している 2,851 語をランダムに選択

という3タイプの方法を試した。

3.2. 自動音声認識によるノイズ

実際に発生するノイズは、上記のランダム関数による人工的なノイズとは性質が異なるため、会話マイニングを実用化する上での前提となるによるノイズを含むデータを作成した。使用した自動音声認識システムは[7]で用いられているものをベースに、人手による書き起こしデータを学習データに加えてチューニングしたものである。2 節で紹介したレンタカー業務の会話のうち 534 件をこの自動音声認識システムでテキスト化した。人手による書き起こしを正解データとした場合の単語誤り率は40%台であるが、書き起こしの誤りを考慮すると実際の誤り率は40%前後ではないかと推測される。

3.3. ノイズを含むデータの分析結果

各方法でノイズを含めたデータの例を図 3 に示す。

I can return it to the minneapolis saint paul airport

録音会話から人手による書き起こしでテキスト化したデータ

th can mentioned hour havent expired minneapolis saint paul airport

人手で書き起こされたテキストに 50%の割合で人工的にノイズを入れたデータ(タイプ C)

I can return it to mini apples cancel

録音会話から自動音声認識によりテキスト化したデータ

図 3: ノイズを含むデータの例

図 3 で見られるように、音声認識によるノイズはランダムなノイズとは異なるかたちで発生する。正しい認識結果とエラーがランダムに分布するのではなく各々が連続し易い上、正解データと認識結果のワード数は必ずしも一致しない。また、エラーが発生していても、発音が似た表現と置き換わることがある上、単語の連鎖が音声認識エンジンの学習コーパスに即していることから、人工的なノイズの結果と比較すると、一見して、より自然なテキストとなっている。しかし、テキスト全体の 4 割程度の単語の認識が誤っていることから、認識結果のテキストを読んでも、概ねどのような内容に関する会話かの想像が付く程度で、具体的な内容を把握するのは不可能である。

各方法で作成したノイズを含むデータをTAKMIで分析し、2 節で得られた知見を獲得できるかを調査した。ノイズはテキスト部分にのみ加えられており、

- 会話の結果としての業務成果
- 担当者の情報

といったテキスト以外の付随情報は元データと同様である。

人工的にノイズを挿入したデータを利用し、ノイズの混入割合が異なる場合に、有益な知見につながった相関の強さがどの程度まで維持されるかを調べたところ、ノイズ混入率が 50%から 60%程度までは相関が維持されるという結果が得られた。その一例として、2 節で示したとおり、人手で書き起こした『顧客の要望に応えられず予約に至らなかった 42 件のデータ』において、相関が強いと認識された顧客発言中のキーワード debit 及び debit card が、ノイズを含むデータでも相関が強いと認識されるかを調べた結果を表1に示す。ノイズがランダムに挿入(置換)されることから、ノイズ率が高まると debit と card が隣接した debit card という複合語が残る確率が低くなるが、ノイズ率 50%までは対象データにおける debit card もしく debit の相関の強さが 2 を超えており、分析者にとっては、対象データ中の他のキーワードと比較して明らかに相関が強いことを認識できるレベルにあった。従ってこの例においては、ノイズ率50%のデータにおいても

顧客の要望に応えられず予約に至らなかったケースにおいては debit card が何らかの影響を与えている可能性が高い

という知見を得られる可能性が高いと認められた。

同様にして、人手による書き起こしデータのテキストマイニングで得られた様々な知見につながる傾向が、ノイズ混入率

50%程度のデータにおいて再現できることが確認された。

さらに、自動音声認識でテキスト化したデータにおいても同様の傾向が再現され、自動音声認識によるノイズを含むデータでも人工的なノイズを含むデータと同様に、有益な分析結果を得られることが確認された。

4. おわりに

コールセンターでの実業務における生の会話データの分析を通じて、会話の分析が業務上有益な知見の獲得につながることを確認し、既存の自動音声認識によりテキスト化されたノイズの多いデータでも有益な分析が可能である例を示した。

ノイズの割合が実際に何%まで許容されるかはデータ次第であり、ノイズが加えられた後にも相関の強さが安定して認識可能なレベルに維持されるかどうかは、元データにおける相関がどの程度強いか、またデータ量がどれだけ多いかなどに依存する。従って、50%程度のノイズが加えられても、元データにおける相関が維持されていたという第3節の結果は、あくまでも事例にすぎず、50%程度のノイズなら常に相関が維持されると結論付けることはできない。しかし、人が読んで、具体的な内容を認識できないレベルのノイズを含むデータからも元データと同じ知見を得ることができたという実例の存在は、既存の自動音声認識技術を生かしたテキストマイニングの実用可能性の高さを示しており、特筆に価すると筆者らは考えている。

テキストマイニングを実践する上では、特定の内容に相関の強い内容という形で検出される傾向の妥当性を検証し、その意味を理解した上で、その知見の活用方法を検討する必要がある。その過程においては、傾向を形成している個々のデータの参照が不可欠であり、ノイズが多いテキストデータを読んでも具体的な内容を把握できない場合には、元の音声データに容易にアクセスできるようにするといった工夫が必要である。そのため、実用性の高い会話マイニングを実現する上では、既存の自動音声認識エンジンと既存のテキストマイニングツールを単純に組み合わせるだけでは不十分と考えられる。

また、本稿では会話マイニングをテーマにしたことから、音声会話をテキスト化したデータの分析を対象としたが、文書データ全体における内容の変化や偏りを捉えるテキストマイニングが、高い割合でノイズを含むデータにおいても有効性が高いという第 3 節で示した知見は、異なる性質のテキストやノイズに対しても成り立つ可能性がある。例えば、多様な言語地域にコールセンターを配置している多国籍企業において、多様な言語で記録された顧客対応記録を横断的に分析することができれば、グローバルな情報共有や地域特性の分析などが可能になる。その際、本稿の第 3 節で示した例のような相関の強さの維持が、機械翻訳によるノイズでも同様に認められるとすれば、多言語データを同一言語に機械翻訳した上で一元的にテキストマイニングを行って有益な分析結果を得る仕組みの実現可能性が高くなる。例えば、異なる言語のデータにおいて似たような傾向が見られるか否かを確認するといった分析は容易に実現できそうである。

近年、多くのコールセンターおいて音声録音環境の整備が進み、自動音声認識の精度が向上している中、コールセンターにおける会話の音声認識結果を利用する試みが徐々に増えてきている[1][6][7]。本稿で示したとおり、会話マイニングの有効性と実用レベルでの実現可能性の高さが確認できたことから、この動きは加速するものと考えられ、そこで必要となる自然言語処理の重要性がますます高まると期待される。

参考文献

[1] G. Mishne, D. Carmel, R. Hoory, A Roytman, and A. Soffer. Automatic Analysis of Call-center Conversations. Proc. ACM Fourteenth Conference on Information and Knowledge Management (CIKM), 2005.

[2] 那須川哲哉. テキストマイニングを使う技術/作る技術—基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006

[3] 那須川哲哉. コールセンターにおけるテキストマイニング. 人工知能学会, Vol.16 No.2, 2001.

[4] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. IBM Systems Journal, Volume 40, Issue 4, pp.967-984. 2001.

[5] M. S. Neff, R. J. Byrd, and B. K. Boguraev. The talent system: Textract architecture and data model. In Proceedings of the HLT-NAACL workshop on software engineering and architecture of language technology system, pages 1.8, 2003.

[6] R. Shourya and L.V. Subramaniam. Automatic Generation of Domain Models for Call-Centers from Noisy Transcriptions. Proc. the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (ACL-COLING), 2006.

[7] G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu and B. Kingsbury. Automated Quality Monitoring in the Call Center with ASR and Maximum Entropy. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2006.