E-012

Doc2Vec を用いた訪問販売検知方法の研究
Study on Sales visit detection method using Doc2Vec

平川 凛† 中藤 良久†
Rin Hirakawa Yoshihisa Nakatoh

1. はじめに

高齢者人口の増加に伴い、高齢者の消費者トラブルの相談件数が年々増加していることが報告されている。トラブルの件数は約 20 万件にのぼり、そのうち約 15%が電話勧誘販売、約 13%が訪問販売に関するものである。

電話勧誘販売（オレオレ詐欺）に対しては、音声からユーザーのストレス状態を推定し、通話の音声認識結果に含まれる単語情報と統合することで詐欺誘引通話を検出する試みがなされている[1]。一方、訪問販売の検出に対しては状況の再現などが困難であり、技術的なアプローチをとった事例は見当たらない。

そこで、本研究では会話音声から訪問販売が行われているかを判別するシステムの構築を目的として、会話音声のテキスト化を行った後に訪問販売であるかどうかを判定する手法を提案する。

2. 提案法の概要

提案手法では、会話音声を音声認識によりテキスト化したものを入力とし、訪問販売の会話文書を学習した Doc2Vec モデル[2]により算出した類似度を基に、訪問販売であるかを判定する（図１）。以下、Doc2Vec と類似度の算出方法について詳述する。

2.1 Doc2Vec

Doc2Vec とは、Quoc Le らにより提案されたアルゴリズム Paragraph Vector[2]をプログラミング言語 Python で実装したものである。

先行するアルゴリズムとしては Word2Vec[3]が挙げられ、これは文章中の単語をベクトル表現に変換する際に従来使用されていた Bag-of-Words の弱点を克服する手法として提案された。Word2Vec では、Bag-of-Words で評価できなかった単語の語順情報や意味関係を有効に表現できることに加えて、単語ベクトル同士の演算によって意図する単語を求めることも可能となった。

Word2Vec を用いて文書の類似度を算出しようとする場合、文書中に出現する全単語のベクトルの重み付き平均を文書自身のベクトルとして扱う方法が考えられる。しかし、このような方法では Bag-of-Words と同様に語順を反映できないという問題があり、短いフレーズや複数文を上手く表現することが難しい。Paragraph Vector では、文書そのものをベクトルとして表現できるように学習を行うため、どのような長さの文にも対応することが可能である。

2.2 Doc2Vec の学習

PV-DBOW

文書ベクトルを入力として与え、該当する文書内に含まれる単語を予測するように学習を行う（図２(a)）。単語ベクトル行列の学習が不要でメモリの消費が少ないという利点がある。

PV-DM

文書ベクトルと周辺単語の情報を中間層で結合したベクトルから、該当する箇所の単語を予測するように学習を行う（図２(b)）。文書ベクトルは文脈を保持するメモリのように振る舞うため、文書全体の文脈を加味することができる。計算量は多くなる一方で PBOW よりも精度が高いことが報告されており、提案手法では PV-DM を学習方法として採用する。

2.3 類似度の算出

Doc2Vec では、新規に与えられた文書に対して Paragraph Vector を推定することができる。提案手法においては音声認識結果のテキストから推定したベクトルと、教師データ

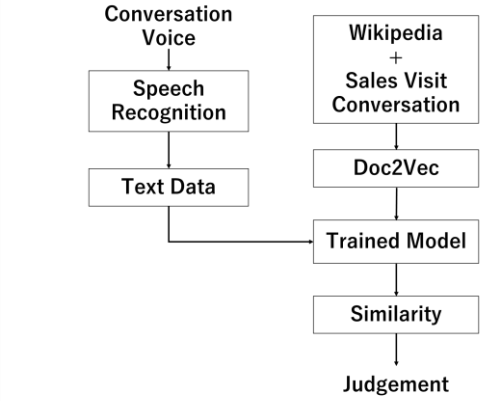

図 1 提案法のフローチャート

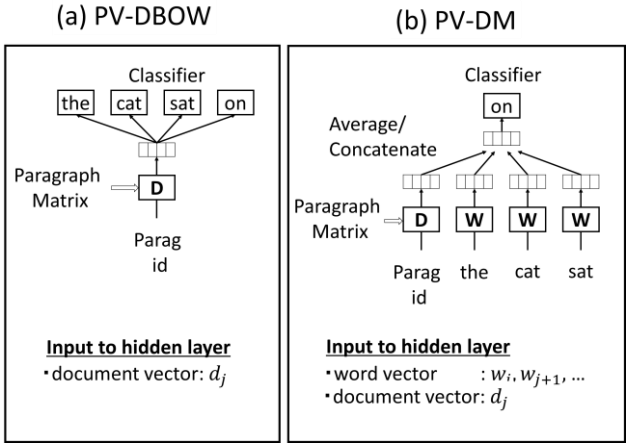

図 2 Doc2Vec の学習方法

† 九州工業大学, Kyushu Institute of Technology

として学習させた訪問販売会話テキスト各々の文書ベクトルとの間でコサイン類似度を算出し、それらの平均値を訪問販売の判定に用いる。

3. 評価実験

提案する訪問販売検知方法の有効性を評価するため、テストデータを用いて実際に会話テキストの分類実験を行う。分類性能は、提案手法において算出した類似度を基に描画した ROC 曲線(Receiver Operating Characteristic Curve)の曲線下面積を用いて評価する。

3.1 実験条件

Doc2Vec モデルには、表 1 のパラメータを使用し、表 2 に示す文書データを用いて学習したものを使用する。Doc2Vec では未知の単語に対してランダムな値を割り振るため、Wikipedia のダンプデータを学習データに追加することで語彙数をカバーしている。

分類実験におけるテストデータには表 2 に示す会話文書データを使用する。会話コーパス[4]は、2～4 名の雑談（30 分程度）を書き起こしたものであり、実験に使用する際にはあらかじめヘッダー・フッターの情報を除去している。訪問販売の会話テキストは、消費者センターにより公表されている事例を基にして独自に作成したものであり、それぞれの文書は長さの異なる 10～20 のやり取りを含んでいる。

学習データ・テストデータは、いずれも MeCab による形態素解析を行った後、動詞・形容詞・名詞のみからなる単語のリストに整形する。

4. 実験結果

各テストデータに対して算出した文書の類似度を用いて ROC 曲線を描画した結果を図 3 に示す。ROC 曲線下の面積（AUC）は 0.92 であり、一般的に AUC は 1 に近いほど分類精度が高いとされるため、提案手法では有効に訪問販売を検知できていると考えられる。また、会話コーパスの中で訪問販売会話との類似度が高くなっているものを調査したところ、お金や金額に関する話題が含まれている傾向が確認された。

また、予備検討として PV-BOW において表 1 と同様のパラメータを用いて学習を行った場合に AUC は完全に 1 となったが、これはデータ数が少ないために Paragraph Vetor が Overfit したものと考えられる。

5. 結論

本研究では、テキスト化された会話音声に対して Doc2Vec モデルを用いて訪問販売かどうかの判定を行う手法を提案した。文書ベクトルを用いて算出した類似度による分類実験では、ROC 曲線下面積は 0.92 となり、おおむね良好な結果が得られた。

今後の課題としては、文書ベクトルを入力とした分類ネットワークを構築するなど、より精度の高い分類方法を検討すると共に、信頼性を高めるために文書データの拡充を行っていく必要がある。

また、今回の評価実験では完全な音声認識結果が得られたという想定で訪問販売検知を行っている。しかし、現実には誤認識などにより文書が不完全である可能性を考慮する必要があるため、実際に音声認識を行って得られた文書に対しても有効に分類を行えるかの検討を行う予定である。

参考文献

[1] 松尾直司他, “音声からのストレス状態検出システムの開発”, DICOMO2014 シンポジウム, (2014).
[2] Quoc Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, Proceedings of The 31st International Conference on Machine Larning, pp.1188-1196 (2014).
[3] Tomas Mikolov et al., “Distributed Representations of Words and Phrases and their Compositionality”, Advances in neural information processing systems, pp3111-3119 (2013).
[4] 藤村逸子・大曽美恵子・大島ディヴィッド義和、2011「会話コーパスの構築によるコミュニケーション研究」藤村逸子、滝沢直宏編『言語研究の技法：データの収集と分析』p. 43-72、ひつじ書房

表 1　学習に用いるパラメータ

Parameter	
Learning Method	PV-DM
Vector Size	300
Alpha	0.0015 (constant)
Window	5
Sample	10^{-4}
Min Count	1
Epoch	20

表 2　評価実験に用いる文書データ

Training Data		Validation Data	
Contents	# of Data	Contents	# of Data
Wikipedia	2653	Conversation Corpus	129
Sales Visit	21	Sales Visit	9

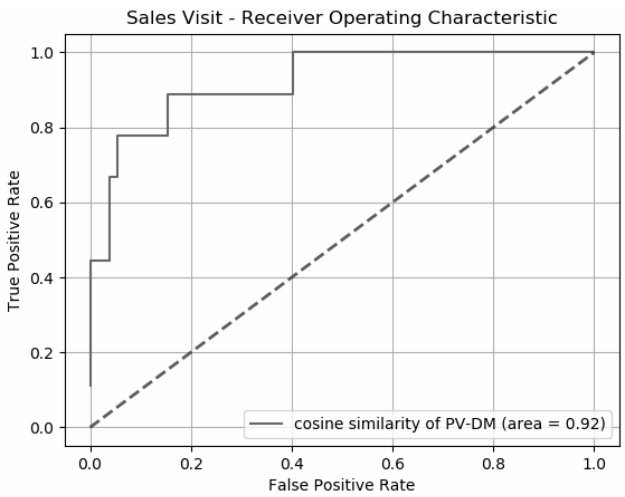

図 3　文書類似度の ROC 曲線