E-012

テレビ番組データを対象とした人名抽出と番組ジャンル推定

Person Name Extraction and Genre Classification for TV Program Data

織田 一輝† 佐々木 稔‡
Kazuki Oda Minoru Sasaki

1. はじめに

これまで、テレビで放送された番組は数多く存在し、テレビ番組に出演する人物やグループの名前も多種多様である。特に、タレントや芸人の名前は知らなければ名前だと判断できない人物も多数存在する。多くのタレントやバンドについて活動内容や知名度を分析するためには、テレビ番組データから人名やグループ名を自動的に抽出できることが求められている。テレビ番組といった内容に偏りのあるデータからの人名抽出や放送内容のジャンル識別はこれまでに行われていない新しい取り組みである。先行研究では、Twitter でのツイート内容に対して人名やテレビ番組名を抽出する研究が行われている[1]。ツイートによるテレビ番組の分析を行うと、人気のある場面や印象に残る部分など番組の一部分についての情報しか得られないが、テレビ番組データを扱うことで番組内容を詳しく捉えることができると考えられる。

テレビ番組データからテキスト分析を行うにはデータの絞り込みをする必要がある。このデータには番組情報や番組内容、出演者などのテキストや内容のまとまりであるシーンの開始・終了時刻などの数値データが存在する。様々な情報を持つテレビ番組データから効果的にテキスト分析を行うために、本研究ではこのデータからテキスト情報に焦点を絞って、有効な特徴の抽出を行うことを提案する。

さらに、抽出した特徴を用いて、テキストに含まれる人名の抽出と各シーンのカテゴリ分類を行う。人名の固有表現抽出では、テレビ番組データの中からヘッドラインを形態素解析し、人名である部分にラベリングを行い、データを作成する。このデータを Bi-LSTM と CRF を組み合わせたモデルと ELMo (Embeddings from Language Models)[2] を使用した人名抽出モデルを構築し、それぞれのモデルの人名抽出精度の比較を行う。シーンのカテゴリ分類では、番組のヘッドラインとその番組カテゴリを合わせたデータを作成し、サポートベクターマシン(SVM)、多層パーセプトロン(MLP)、BERT(Bidirectional Encoder Representations from Transformers)[3] を用いた手法の 3 つの分類モデルを使用しジャンル推定を行う。

2. テレビ番組データ

本研究では、企業より提供を受けた 5 種類のテレビ番組データを使用する。5 種類のデータの概要を以下に示す。

- TV 番組データ：番組内容についての情報
- TV 番組シーン(Scene)データ：番組をシーン毎に細分化し、内容のまとまりに分けた情報
- 商品情報(product)データ：放送された番組の中で紹介された商品の情報
- スポット・施設データ：番組内で紹介されたスポットや施設の情報
- CM データ：放送された CM の詳細な情報

これらのデータは表形式で記述されている。その表には数値や日付など多種多様な値が含まれているため、データをそのまま使用することは難しい。本研究では、これらのデータから各番組に出演する人物やグループ名を抽出するために、テキストで記述されたヘッドラインに注目する。このヘッドラインが存在する Scene データと Product データを対象として人名抽出と番組ジャンル推定を行う。実験では、2018 年 7 月 27 日から 7 月 29 日までの Product データ 1,321 件と Scene データ 10,689 件を対象とした。

3. テレビ番組データからの人名抽出

TV 番組のヘッドラインを入力データとし、その単語系列に対して人名のラベル付けをした結果を出力する。

まず、ヘッドラインに含まれる文章を単語の系列に変換する。単語を抽出するために、オープンソース形態素解析エンジンである MeCab[1] を使用する。MeCab の辞書として、本研究では人名などの固有表現を効果的に抽出できる mecab-ipadic-NEologd[2] を使用した。mecab-ipadic-NEologd は Web データから新しい固有表現を随時追加することで定期的に辞書の更新が行われている。そのため、最近の TV 番組データに出現する人物名などをひとつの単語として抽出可能である。

次に、得られた単語の系列に対して、人名に該当する部分とそれ以外の部分を表すラベルを割り当てる。ラベル付けを行う際には IOB2 フォーマットを使用する。形態素解析結果における品詞細分類 2 が「人名」であるものを人名としてラベル付けを行い、その後手作業で修正を行った。割り当てるラベルは人名の先頭の形態素に B-PSN、人名の先頭ではない形態素に I-PSN、人名ではない形態素に O をとする。図 1 にラベル付けの例を挙げる。

実験では、学習データとしてラベル付けした Product データ、テストデータとして Scene データを用いる。学習データに対し、双方向の Long Short Term Memory (Bi-LSTM) と CRF(Conditional Random Fields)を組み合わせたモデル、文脈を考慮した単語表現を獲得できる ELMo と CRF を組み合わせたモデルでそれぞれ訓練する。訓練データの学習回数(Epoch 数)はそれぞれ 20 とする。構築したモデルにテストデータである Scene データを入力し、２つのモデルにおける人名抽出精度を比較する。

また、訓練データに対し Wikipedia における存命人物タグが付いた人名を追加した場合についてもモデルを構築し、

†茨城大学大学院理工学研究科情報工学専攻 Major in Computer and Information Sciences, Graduate School of Science and Engineering, Ibaraki University
‡茨城大学工学部情報工学科 Department of Computer and Information Sciences, College of Engineering, Ibaraki University

[1] https://taku910.github.io/mecab/
[2] https://github.com/neologd/mecab-ipadic-neologd/

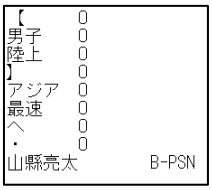

図 1 人名のラベル付けを行ったデータ

	product データのみ	product データ +Wikipedia
Bi-LSTM + CRF	52.48	29.95
ELMo + CRF	**56.47**	28.80

表 1.人名抽出の実行結果

外部の人名リストが有効かどうか検証を行った。

実験結果の F1 値を表 1 に示す。product データのみで訓練した ELMo の精度が最も高く、約 56%の F1 値となった。訓練データに Wikipedia の人物名を追加した場合は F1 値が大幅に低下した。精度が低下した要因としては、追加した人物名に文脈が含まれていないこと、人物名の数が文脈を含むヘッドラインの数に比べて多いことが挙げられる。

4.ジャンル推定

次に、ヘッドラインのテキストに対して、番組ジャンルの推定を行う。ヘッドラインを人名抽出と同様に MeCab で形態素解析し、テキストを単語列で表現する。この単語列を 3 つの教師あり学習手法 SVM、MLP、および BERT を用いて番組ジャンルを推定するモデルを構築する。SVM と MLP は各出現単語の頻度を要素とするベクトル、BERT は単語列を入力してモデルを学習する。BERT では事前学習済みモデル [4] を利用して単語列に対応する分散表現を求め、MLP で再学習することでジャンル分類モデルを構築する。SVM のパラメータは scikit-learn のデフォルト値、MLP の学習回数は 200 回、BERT の学習回数は 20 とした。

各手法に対し、Product データを学習データ、Scene データをテストデータとして各学習手法のジャンル推定精度の比較を行った。ただし、Scene データ中にある”CM”のみテキストは取り除いた。カテゴリ名は"スポーツ"、 "トピックス"、 "社会"、 "政治・国際"、"ビジネス"、 "暮らし"、"文化・芸能"、"サイエンス"の 8 種類で、この中から 1 つのジャンルを出力する。各ジャンルのデータ数を表 2 に示す。表 4 に各手法における正解率のマイクロ平均を示す。今回の実験結果では、MLP の分類精度が BERT よりも高かった。BERT は人名抽出で使用した ELMo と同じように訓練データの文脈から推定を行っており、文脈があまりないヘッドラインでは有効な分散表現が得られなかったと考えられる。そのため、単語の出現頻度のみを使用している MLP の方が高い精度となったと考えられる。また BERT の事前学習のデータとして用いたデータはビジネスニュースを用いて作成されたものであり、今回の使用したテレビ番組のヘッドラインとは異なる部分が多く、あまり適したデータとはいえない可能性がある。

また、テストデータの多くがすべての手法で”文化・芸能” のジャンルに分類される傾向があった。これは訓練データに使用した Product データが文化・芸能に偏っていることが原因であると考えられる。

	Product データ	Scene データ
サイエンス	2	45
スポーツ	19	986
トピックス	94	1433
ビジネス	10	212
政治・国際	12	354
文化・芸能	374	2735
暮らし	49	656
社会	9	1132
合計	**569**	**7553**

表 2　ジャンル推定データの内訳

分類手法	平均正解率
SVM	0.3621
MLP	**0.5881**
BERT	0.3584

表 3 各手法のマイクロ平均結果

5.おわりに

本研究では、テレビ番組データの中から各シーンのタイトルを記述したヘッドラインを用いて、そこに含まれる人名の抽出と各シーンのカテゴリ分類を行った。実験の結果、人名抽出では ELMo を用いた CRF による手法が最も高い精度で人名抽出を行うことができた。しかし、人名リストを用いると逆に精度が下がってしまう結果となった。ジャンル推定では MLP を用いた場合の正解率が最も高かった。テレビ番組データをそのまま利用したことで、必要のないデータや同じ文章の繰り返しになっている部分や、データの偏りなどあったため、精度が低くなったと考えられる。今後の課題として、文脈に依存することなく人名抽出やジャンル推定が可能である手法を新たに使用することや、使用するデータの精査やデータ量の増加などが挙げられる。

謝辞

本研究を進めるにあたり、番組データを提供してくださった株式会社エム・データ様に感謝申し上げます。

参考文献

[1] Named Entity Recognition in Tweets:An Experimental Study, Alan Ritter, Sam Clark, Mausam and Oren Etzioni, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534 (2011).

[2] Deep Contextualized Word Representations, Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2227–2237, (2018).

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4171–4186, (2019).

4. https://qiita.com/mkt3/items/3c1278339ff1bcc0187f
5. https://scikit-learn.org/stable/