機械学習による自由記述課題の分類

○古川善也
(愛媛大学教育学部)

問題

心理学研究の実験では，様々な方法を用いて条件操作が行われる。その手法の一つとして，自由記述課題を用いた条件操作があるが，記述内容が意図したテーマに沿っているかの確認には労力を要する。本研究では，テーマに沿った記述がされているかを判別するのにトピックモデルを利用することができるかを検討する。

トピックモデルとは，文書を構成するトピックを推定する分析手法である(石田, 2017)。ここでのトピックは文書の主題あるいはテーマを意味する。この文書の背後にあるトピックによって文書に含まれる単語の出現頻度は影響をされる。また，個々の文書は単一あるいは複数のトピックで特徴づけられ，それぞれのトピックの比重は文書ごとに異なる。トピックモデルではこの比重を推定することで個々の文書の分類を行う。

本研究では，分析に用いる文書について，他者との道徳に関する比較の実験操作(古川・中島, 2018)において記述課題で得られた記述内容を使用する。

方法

参加者　調査会社マクロミルの 20 代～40 代モニター166 名(女性 125 名，平均年齢 34.7 歳)。

手続き　参加者には，過去に目撃した他者の行動を想起・記述するよう求めた(e.g., Zhong et al., 2017)。不道徳条件では他者が不道徳な行動をとっていたところを目撃した経験を思い出すよう求めた。道徳条件では他者が道徳的な行動をとっていたところを目撃した経験を思い出すよう求めた。

結果

文書中の形態素の抽出は，形態素解析エンジン MeCab(https://taku910.github.io/mecab/)を用いて行った。抽出する形態素は，品詞を名詞と動詞とし，全文書中で 5 回より多く出現するものに限定した。また，解析において意味情報の乏しい小分類(「数」，「非自立」，「接尾」，「固有名詞」)は除外した。その結果，名詞 72 個，動詞 43 個の形態素が抽出された(一部を Table 1 に記載)。

次に，文書に含まれるトピック数を決定するために ldatuning パッケージの FindTopicsNumber 関数を用いて分析を行った。その結果，Caojuan (2009), Deveaud (2014)の指標に基づき 2 トピックにおいてピークとなることから，トピック数を 2 とした。

トピックの推定には topicmodels パッケージの LDA 関数を用いた。この関数では潜在的ディリクレ配分法(LDA)のアルゴリズムでトピックを推定する。それぞれのトピックについて出現スコアの高い形態素を 15 個抽出したところ，トピック 1 のみに「不道徳」，トピック 2 のみに「道徳」の形態素が含まれていた。次に，各文書のトピックの比重を算出し，この比重が大きい方を各文書のトピックとした。このトピックによる分類と条件操作の一致率を算出したところ，一致率は 50%と低い値であった。

考察

トピックモデルによる文書分類はチャンスレベル程度と低い精度であり，文書がテーマに沿った記述であるかの評価に用いることは難しいようであった。しかしながら，いくつかの改善点が残されている。本研究では LDA を用いたが，各文書が 1 つのトピックから成り，そこから単語が生成されるとする混合ユニグラムモデルによる分析がより有効であると考えられる。また，サポートベクターマシンなどの他の機械学習の手法を用いての検討も有効であると考えられる。

Table 1
条件別での頻出形態素

immoral			moral		
人	名詞	121	人	名詞	87
思う	動詞	65	思う	動詞	68
自分	名詞	45	自分	名詞	62
私	名詞	44	私	名詞	60
感じる	動詞	40	声	名詞	38
いる	動詞	38	かける	動詞	35
席	名詞	35	感じる	動詞	32
不道徳	名詞	32	見る	動詞	29
言う	動詞	30	道徳	名詞	27
見る	動詞	26	仕事	名詞	19