Sentence-BERT による分散表現を用いたベストアンサーの推定

間宮壮太[1] 市川治[2]
[1]滋賀大学データサイエンス研究科 [2]滋賀大学
s6023143@st.shiga-u.ac.jp osamu-ichikawa@biwako.shiga-u.ac.jp

概要

本研究では、Yahoo!知恵袋に投稿された質問と回答文を対象とし、Sentence-BERT で得た文ベクトルを LightGBM で学習することでベストアンサー予測するモデルを構築した。回答のみより質問文＋回答文を併用した方が高精度となり、人間の予測精度を上回る結果を得た。一方でカテゴリ別分析では、恋愛や政治といった正解が定まらない領域で精度が低く、質問者の主観や好みが BA 選定に大きく影響する傾向が示唆された。

1 はじめに

「質問にどう答えるべきか」を考えることは、人間にとって非常に難しい課題である。質問の内容や文脈、そして質問者の意図を正確に把握し、適切に回答するためには、幅広い知識と深い洞察力が必要となる。さらに、近年ではこうした課題に取り組むのは人間だけではない。ChatGPT をはじめとする生成 AI を活用した質問応答システムが普及し始め、コンピュータによる自動応答や支援が日々行われている。質問応答システムは、「ユーザーにとって役立つ回答とは何か」を学習することで、回答精度を高めている。こうした背景から、回答を適切に評価する仕組みや、その自動化手法への需要が近年高まっている。

Yahoo!知恵袋では、寄せられた回答の中から「ベストアンサー」が選出される。このベストアンサーは質問者が得られた回答を吟味し、最も役に立ったと感じた回答 1 つに与えられる。長野敬介らの Learning to Rank を用いたベストアンサー推定モデル[1]や、横山友也らの質問文と回答文の印象評価を基にしたベストアンサーの推定[2]、石川大介らの要因を用いた SVM モデル[3]等、ベストアンサーを推定することで、回答評価の為の知見を生み出す研究が進められている。

本研究ではそれらの研究に倣い、Yahoo!知恵袋データセットを用いて、質問文と回答文を入力とし、ベストアンサーの予測値を出力するモデルの作成を目指す。その上で、本研究では質問や回答の文意を取得し予測精度に貢献することを目指す。具体的には、Sentence-BERT(SBERT)[4]を用いることで意味的な類似や文章の関係を特徴量空間に反映し、その分散表現を用いることで文意を取得する。

以下、本論文の構成を記述する。本節は 1 節だ。2 節では、使用するデータとモデルの構造について述べる。3 節では、予測の結果と考察を述べる。4 節では本論文をまとめ、今後の展望について考察する。

2 研究方法

2.1 ベストアンサーの詳細

「ベストアンサー」は Yahoo!知恵袋において、質問者にとって最も納得、満足したもの、または他の利用者の投票でベストアンサーが選ばれる。
ベストアンサーの選ばれ方の詳細は図 1 に示す。以下、本論文では「ベストアンサー」を略す際は BA、「ベストアンサー以外の回答」については NA と表記する。

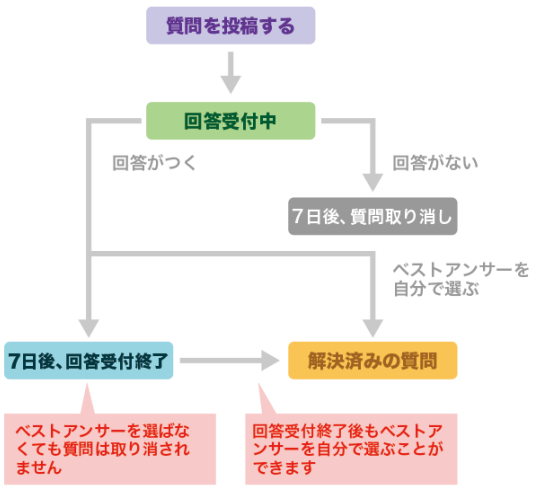

図 1 ベストアンサーの選ばれ方(Yahoo!知恵袋のヘルプ[5]より引用)

2.2 データセットの詳細

本研究では、Yahoo!知恵袋データセット[6]を使用し、2015 年から 2018 年の期間に投稿された質問と回答を対象とした。このデータセットには、約 240 万件の質問と約 640 万件の回答が収録されており、1 つの質問に対する平均回答件数は 2.62 件である。

回答件数ごとの質問数の分布は以下の表に示す通りである。

表１　回答件数ごとの質問数の度数分布

回答件数	質問数	割合	累積割合
1	1105923	0.447	0.447
2	549442	0.222	0.670
3	304453	0.123	0.793
4	177959	0.072	0.865
5	106550	0.043	0.908
6	67661	0.027	0.935
7	43766	0.018	0.953
8	29720	0.012	0.965
9	20708	0.008	0.973
10 以上	65599	0.026	1

上記から、回答が 2～5 件のデータを対象とする。

また、Yahoo!知恵袋では質問者が自身の質問カテゴリを設定することができる。

以下はデータの対象とする Yahoo!知恵袋データセットにおけるカテゴリ質問数上位 16 件である。カテゴリ名とその質問数、全体に占める割合が示されている。

表２　yahoo 知恵袋データセットにおける人気カテゴリ上位 16 件

カテゴリ	カウント	割合
恋愛相談、人間関係の悩み	934224	0.144
アダルト	492582	0.076
スポーツ	285864	0.044
料理、レシピ	247997	0.038
政治、社会問題	225245	0.035
音楽	213482	0.033
アニメ、コミック	207012	0.032
芸能人	203877	0.031
ゲーム	190805	0.029
健康、病気、病院	186980	0.029
自動車	184456	0.028
テレビ、ラジオ	132612	0.020
言葉、語学	130414	0.020
交通、地図	128675	0.020
受験、進学	126274	0.019
生き方、人生相談	105657	0.016
国内	102528	0.016

ここからカテゴリ「アダルト」を排除し、残った 15 件を学習対象とした。上記カテゴリからそれぞれ約 10000 件のデータを取り出し、そのうちの 10%をテストデータとした。

2.2 モデルの設計

Yahoo!知恵袋データには質問文とその質問への回答文が含まれる。これらの文から分散表現を取り出す場合、回答文のみ、質問文と回答文それぞれ、質問文と回答文を合わせ 1 文とした場合の 3 つの方法が考えられる。そこで、これらの 3 つの方法それぞれでモデルを構築し、比較を行う。

分散表現の取得方法には、Pretrained Japanese BERT models[7]を事前学習モデルとした SBERT を用いた。以降、学習では全てバッチサイズを 16、エポック数 3、損失関数を TripletLoss とし、学習データの 10%を推論用データとし Validation Loss の最小となったエポックを使用した。

本研究では、分散表現の生成方法として、入力文を構成する全ての単語の分散表現の相加平均を取る方法を用いた。なお、以降では簡便の為文の分散表現を文ベクトルと表記する。

2.2.1 回答文のみ

学習データに対し、Sentence-BERT を用いて以下の損失関数を最小化する学習を行った。具体的には、以下の値が最小となるよう学習を行った。

$|a_{b1} - a_{b2}| - |a_{n1} - a_{n2}|$　a_b … ベストアンサー　a_n … 非ベストアンサー

この学習後、学習データの回答文から得られた文ベクトルを LightGBM[8]によって学習し、テストデータの回答文から得られた文ベクトルのベストアンサーを予測する。

2.2.2 質問文と回答文それぞれ

質問文と回答文をそれぞれで学習する。学習として、以下の最小化問題を解いた。

$|q - a_b| - |q - a_n|$　q … 質問　a_b … ベストアンサー　a_n … 非ベストアンサー

この学習後、回答文のみの際と同様に LightGBM によって予測する。

2.2.3 質問文と回答文を合わせた 1 文

質問文と回答文を合わせた全文を分析した際は、以下の最小化問題を解いた。

$|c_b - c_{b'}| - |c_n - c_{n'}|$　c_b … ベストアンサー全文　c_n … 非ベストアンサー全文

この学習で、「'」の付いた文章には、オリジナル文の 10%を[MASK]トークンで隠した文章を作成し、その文ベクトルを用いた。こちらも同様、得られた文ベクトルを LightGBM で学習し、予測を行った。

3 結果と考察

3.1 結果

上記三件の学習結果は、以下の表の通りとなった。なお、学習精度の比較の為、同様のデータを利用し BERT によって 2 値分類を行った物も共に表記した。

表３　各学習方法における予測精度の一覧

	BERT	回答文のみ	質問文+回答文	全文
Recall@1	0.524	0.342	**0.536**	0.299
MRR	0.736	0.627	**0.744**	0.602
Accuracy	0.632	0.379	**0.636**	0.623
AUC-ROC	0.612	0.426	**0.619**	0.458

精度の評価には、Recall@1、MRR、Accuracy、AUC-ROC を用いた。いずれの場合も質問文と回答文それぞれを用いた文ベクトルの抽出方法が最も精度が高く、他の方法と大きく開きがある。BERT の 2 値分類との差は僅かながら、SBERT を用いた学習の精度が高い結果となった。

以上の結果から、質問文と回答文それぞれを用いた文ベクトルを使用した学習方法が最適と考え、石川らの先行研究によって得られた文字数等の特徴量と共に LightGBM に学習させた結果を最終的なモデルとした。最終的なモデルの予測精度は以下の通りである。なお、この結果は同様のテストデータから 10 件に対して人間による予測した結果の平均 Recall@1 が 40%程度であったことから、人間の精度を超えている。

表４　最終的なモデルの予測精度

	質問文+回答文+特徴量
Recall@1	0.560
MRR	0.758
Accuracy	0.670
AUC-ROC	0.688

3.2 考察

最終的なモデルの予測精度をカテゴリ毎に精度を確認した。以下はカテゴリ上位 5 件の予測精度である。

表５　カテゴリ上位５件の予測精度

	Recall@1	AUC-ROC
ゲーム	0.613	0.719
言葉、語学	0.609	0.708
音楽	0.597	0.706
スポーツ	0.594	0.698
受験、進学	0.58=3	0.715

また､以下はカテゴリ下位５件の予測精度である。

表６　カテゴリ下位５件の予測精度

	Recall@1	AUC-ROC
恋愛…	0.535	0.686
政治…	0.497	0.665
生き方…	0.506	0.655
自動車	0.518	0.659
芸能人	0.527	0.662

上位、下位それぞれのカテゴリを諦観すると、上位にはゲーム、語学、音楽、スポーツなどの特定の知識ついての質問カテゴリが多く、下位には恋愛や政治など正解が定まらない分野の質問が多い。これは正解が定まらないことからどの回答が「最も適切か」を客観的に測ること自体が難しく、精度が伸び悩む傾向にあると考えられる。また、実際のベストアンサー選定には回答の正しさや論理性より、質問者の主観的な好みや期待との適合度が強く影響していることも分かった。実例としては、回答内容が客観的に見て不確かであっても選ばれている場合があり、モデルの方がむしろ正解（実際に選ばれた回答）を的確に当てる場面があった。これらの結果から、提案モデルは回答の客観的な優劣を判断しているわけではなく、「選ばれそうな回答」を見出す仕組みとして機能していることが再確認された。

4 おわりに

4.1 結論

本研究では、Yahoo!知恵袋のデータセットを用いてベストアンサーを予測するモデルを構築し、その性能を評価した。その結果、質問文と回答文の両方を学習に用いる事で、人間の予測精度を超える高い性能を発揮できることが明らかとなった。これは文ベクトルの取得がベストアンサーの予測に対して有効な手段であることを示している。

一方で、本モデルは回答内容の優劣を理解しているわけではなく、あくまで「質問者が満足しそうな回答」を選定しているに過ぎない。また、恋愛や政治など正解が存在しにくい領域に関しては、ベストアンサーの基準自体が曖昧であることから、精度の伸び悩みが見られた。

4.2 今後の展望

今後は正解が定まらない領域に対する扱い方や、回答者の専門性などの情報を統合することで、さらに応用範囲が広がる可能性があると考えられる。具体的には精度の低かった恋愛や政治といった主観性の高い分野でも高精度な予測を行う必要がある。質問者の背景や意図をより詳細に分析し、例えば感情分析やユーザープロファイルなどの外的情報を取得し、推定に利用する方法が有効と考えられる。

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスを通じて、LINE ヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 3 版）」を使用した。データの提供を受けるにあたり、多大な支援をいただいた関係各所に深く感謝する。

参考文献

1. 長野敬介, et al. Learning to Rank を用いた QA サイトにおけるベストアンサーの推定. 第 75 回全国大会講演論文集, 2013, 2013.1: 687-688.
2. 横山友也, et al. 質問回答サイトの質問文と回答文の印象評価. In: 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2010). 2010.
3. 石川大介, et al. Q&A サイトにおけるベストアンサー推定の分析とその機械学習への応用. 情報知識学会誌, 2010, 20.2: 73-85.
4. REIMERS, N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084, 2019.
5. yahoo!知恵袋ヘルプベストアンサーを選ぶには（引用日:2025 年 1 月 9 日）
https://support.yahoo-net.jp/PccChiebukuro/s/article/H000008094
6. 国立情報学研究所 情報学研究データリポジトリ「Yahoo!知恵袋データ(第 3 版)」(引用日:2025 年 1 月 9 日)
https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html
7. HuggingFace tohoku-nlp/bert-base-japanese-v2(引用日:2025 年 1 月 9 日)
https://huggingface.co/cl-tohoku/bert-base-japanese-v2
8. KE, Guolin, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, 30.