論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Examining Impact of Evaluation Dataset Characteristics on Acceptability Judgments
著者(和文)	ヴィジャイ ドルタニ
Author(English)	Vijay Daultani
出典(和文)	学位:博士(学術), 学位授与機関:東京科学大学, 報告番号:甲第33号, 授与年月日:2024年12月31日, 学位の種別:課程博士, 審査員:岡崎 直観,徳永 健伸,村田 剛志,金﨑 朝子,井上 中順
Citation(English)	Degree:Doctor (Academic), Conferring organization: Institute of Science Tokyo, Report number:甲第33号, Conferred date:2024/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		Vijay Daultani	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	岡崎 直観	教授	審査員	井上 中順	准教授
	審査員	徳永 健伸	教授			
		村田 剛志	教授			
		金崎 朝子	准教授			

論文審査の要旨（2000字程度）

本論文「Examining Impact of Evaluation Dataset Characteristics on Acceptability Judgments」は、英文5章で構成されている。自然言語処理において、受容性評価（acceptability evaluation）は言語モデルが生成したテキストが意図した意味を母語話者にどのくらい効果的に伝えるかに焦点を当てるもので、言語モデルの性能評価のひとつの要素である。機械翻訳や自動要約などのタスクでは、単語の頻度や文の長さといったデータセットの特性がモデルの性能や評価に影響を与えると言われているが、受容性評価ではデータセットが及ぼしうる影響について、あまり注目されていなかった。本研究は、単語の頻度や文の長さが言語モデルの受容性評価能力に与える影響を調査するものである。

第1章「Introduction」では、受容性評価に関する基礎的な背景を説明し、その重要性、目的、および自然言語処理における実応用に触れている。また、受容性評価の研究の発展を踏まえながら、本研究の課題を設定し、その重要性を明示することで、言語モデルの性能向上に寄与しうることを強調している。

第2章「Background」では、自然言語処理におけるテキスト品質評価の広範な研究分野をカバーしながら、受容性評価の研究の背景を述べている。受容性と文法性などの関連する概念との違いを明確にし、文レベルで受容性を検討することの重要性を論じている。また、従来の連続的な数値による受容性評価と、最近多くみられる二値分類によるアプローチの違いを比較している。さらに、本研究で単語の頻度と文の長さにフォーカスすることの理論的根拠と2つの評価パラダイムを紹介している。

第3章「Impact of Lexical Frequency」では、単語の頻度が受容性評価にどのように影響するかを議論している。本章ではまず、受容性を連続的な数値として表現する指標に関して、既存研究をレビューし、特に確率ベースの指標の限界として、単語の頻度が受容性評価に与える悪影響を論じている。その影響を軽減するため、データセットの前処理において固有表現を固有表現カテゴリ名に置換するReplace Named Entityを提案した。これは、言語モデルが受容性評価を行うときに、特定の固有表現を手掛かりにするのではなく、人名や場所名などのカテゴリによる粗いレベルで扱うことを意図している。評価実験から、単語の頻度、特に未知語（OOV; out-of-vocabulary）の単語がモデルの性能に大きな影響を与えることが示された。また、本研究で提案したReplace Named Entityにより、未知語の影響が効果的に緩和されることを確認した。

第4章「Impact of Sentence Length」では、文の長さが受容性評価において与える影響を議論している。受容性評価を二値分類タスクとして捉えたあと、先行研究をレビューし、文の長さに焦点を当てる理由を説明している。一般的に使用されるデータセットの文の長さの分布と、人間によって書かれたコーパスとの比較を行い、CoLAやBLiMPなどのよく用いられる評価データセットには短い文が多いことを指摘している。これを踏まえて、本研究では7つのデータセット（6つのデータセットは既存のデータセットを流用したもの、1つのデータセットは本研究にて新たに構築したもの）を検討した。カルバック・ライブラー情報量を距離尺度とし、人間によって書かれたコーパスと評価データセットを比較した。評価データセットの文の長さの分布を補正したうえで言語モデルに受容性評価をさせる実験を行い、特に人間が書いたテキストに典型的に表れる13〜21トークンの長さの文が言語モデルの受容性評価に影響を与えることを明らかにした。一般的に使用されるデータセットで訓練された言語モデルは、この範囲の長さの文の受容性を正確に評価することを苦手とする傾向があり、より代表性の高い訓練データの必要性を主張している。

第5章「Conclusion」では、研究の目的に立ち返り、単語の頻度と文の長さが受容性評価に与える影響に関して、得られた知見をまとめている。言語モデルの受容性評価に様々な要因が影響を与

えることの理解、本研究の限界、今後の研究の方向性を示唆している。
　本論文は、単語の頻度や文の長さが言語モデルの受容性評価に影響を与えることを明らかにし、二つの提案手法によりこれらの影響を緩和できることを示した。本研究は、現在の受容性評価の評価方法の限界を指摘すると同時に、受容性評価の性能と信頼性を向上させるための解決策を提供した。自然言語は同じ意味を伝えるにも様々な書き方が可能であり、本研究の成果は人間にとってより分かりやすいテキストの生成や、受容可能な範囲で情報を埋め込む技術（電子透かし）などへの応用も見込まれるため、学術のみならず工学の発展にも寄与する。よって、本論文は博士（学術）の学位論文として十分価値あるものと認める。