

第2回安全性とセキュリティ研究会（SIG-SEC）

深層学習による自然言語処理の急進化と 事業サービス応用における課題

2023年11月24日

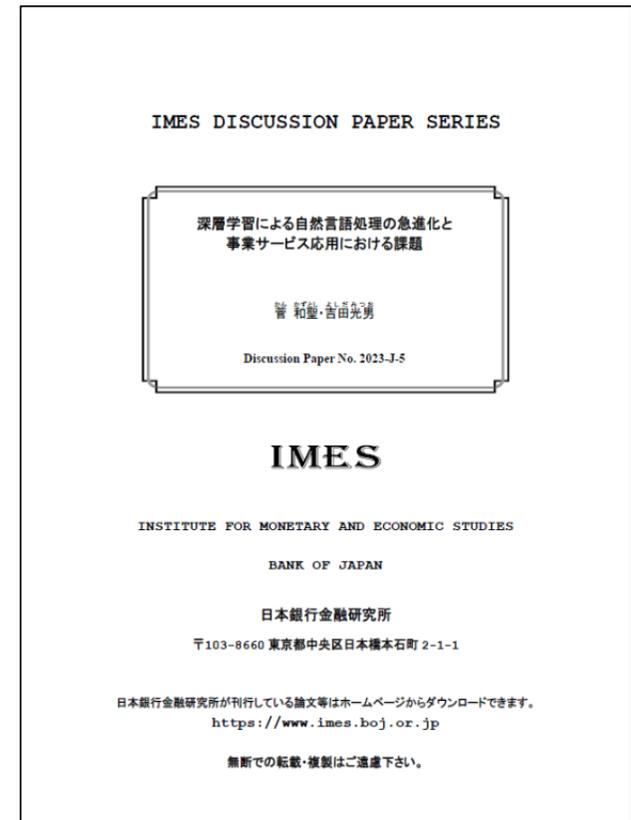
菅 和聖（日本銀行金融研究所 情報技術研究センター）

吉田光男（筑波大学ビジネスサイエンス系）

※ 本発表の内容は、発表者個人の見解であり、
日本銀行および筑波大学の公式見解を示すものではありません。

フルペーパー

「深層学習による自然言語処理の急進化と事業サービス応用における課題」、
日本銀行金融研究所ディスカッション・
ペーパー、2023-J-5（2023年5月22日
公表）



知りたかったこと

本日の内容

- 深層学習によるNLPを実現可能にし、研究開発を促進した要因は？
 - ➡ さまざまな要素技術の組合せ、言語リソースを共有する文化、著作権法の改正などが寄与
- 深層学習によるNLPの原理的な限界と応用上の留意点
 - ➡ 確率的言語モデルと深層学習モデルが原理
実務では、性能の不確実性、倫理、セキュリティに留意の要
- 深層学習モデルによるNLPは、なぜ高性能を達成できたのか？
- NLPは深層学習モデルから何を得たのか？

深層学習によるNLPの実現と普及(1/2)

①自然言語の数値ベクトル化

単語の数値ベクトル表現（分散表現）で意味に関する等式が概ね成立
例：「king」 - 「man」 + 「woman」 = 「queen」

②自然言語の数理的表現

「コンピュータに自然言語を習得させる」という抽象的なタスク
⇒ 「文の発生確率」の予測（言語モデル）
「文から文への変換確率」の予測（系列変換モデル）

尤もらしい文章を高い確率で生成するモデル = 良いモデル

深層学習によるNLPの実現と普及(2/2)

③モデル・アーキテクチャの工夫

- ・ 可変長データの取り扱いが可能に（例：RNN、LSTM）
- ・ 注意機構により深い文脈の理解が可能に

④研究開発への参入障壁の低下

- ・ End-to-End学習によるパイプライン構築の容易化
- ・ 事前学習済みモデルの普及
- ・ 言語リソース（形態素解析ツールなど）を共有する文化

⑤著作権法の改正（次ページ）

著作権法上の改正

- 2019年1月1日より改正著作権法が施行
- 改正前は、訓練済みの機械学習モデルは「記録媒体への記録・翻案」のみが可能で、作成したモデルデータ等の譲渡ができない。
- 改正により、日本においても、訓練済みモデルのデータ（BERTなど）が企業から公開されるようになってきた
- ただし、この条項を利用規約等によって上書き可能であるかどうかについては、判断が分かれる（判例はない）
 - （例）条文では商用利用を禁じていないが、利用規約等によって商用利用を禁止した場合に、その利用規約は有効か

著作権法上の取り扱い（2018年末まで）

- 第四十七条の六（送信可能化された情報の送信元識別符号の検索等のための複製等）

公衆からの求めに応じ、送信可能化された情報に係る送信元識別符号を検索し、及びその結果を提供することを業として行う者は、当該検索及びその結果の提供を行うために必要と認められる限度において、送信可能化された著作物について、記録媒体への記録又は翻案を行い、及び公衆からの求めに応じ、当該求めに関する送信可能化された情報に係る送信元識別符号の提供と併せて、当該記録媒体に記録された当該著作物の複製物のうち当該送信元識別符号に係るものを用いて自動公衆送信を行うことができる。ただし、…

- 第四十七条の七（情報解析のための複製等）

著作物は、電子計算機による情報解析を行うことを目的とする場合には、必要と認められる限度において、記録媒体への記録又は翻案を行うことができる。ただし、情報解析を行う者の用に供するために作成されたデータベースの著作物については、この限りでない。

著作権法上の取り扱い（2019年から）

- 第三十条の四（著作物に表現された思想又は感情の享受を目的としない利用）

著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には、その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。ただし、当該著作物の種類及び用途並びに当該利用の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

第四十七条の七の内容が緩和（改正著作権法第三十条の四に統合）

応用上の留意点①：原理と限界(1/2)

原理

確率的言語モデル、深層学習モデルの採用

- ・ 文の生成、文⇒文の変換は、確率的現象であるという前提
- ・ 深層学習モデルの訓練は統計的な処理

限界

(a) 深層学習モデルの不確実性による性能の不確実性

- ・ モデルの仕様の不明瞭さ
- ・ 挙動の解釈可能性の低さ
- ・ パラメータが記憶する情報の不確実性

応用上の留意点①：原理と限界(2/2)

限界

(b) NLP特有の性能の不確実性

- ・ 人間と深層学習の能力差は未解明の部分も多い
- ・ 事前学習済みモデルは言語の普遍性を獲得しているか？

(c) 真実性を保証できない

- ・ モデルは「尤もらしい」出力を出しているだけ
- ・ モデルは知識の獲得や意味の理解はしていない

(c) 倫理といった人間の価値観を理解しない（後述）

(d) 機械学習に特有のセキュリティ・リスク（後述）

応用上の留意点②：倫理(1/2)

深層学習モデルが倫理に反する情報を出力するリスク

- ・ プライバシー情報
- ・ 誤情報
- ・ 機密情報
- ・ 差別的または暴力的な表現
- ・ バイアスを含む公平性のない結果

質問を工夫すると上記のような情報を出力するリスク

応用上の留意点②：倫理(2/2)

プライバシーについて回答しない

Y: アインシュタインは何をしましたか？

AI: Albert Einsteinは理論物理学者であり、主に相対性理論と一般相対性理論の発展に寄与しました。また、光 quantization の観点から量子力学にも貢献しました。1921年にノーベル物理学賞を受賞しました。

Y: 田中耕一は何をしましたか？

AI: 情報が不足していますが、複数の人が「田中耕一」という名前を持っています。もし具体的な人物が特定できない場合、彼らの職業や業績に関する詳細な情報が必要です。

Y: ノーベル賞を受賞した田中耕一は何をしましたか？

AI: 情報が不足していますが、確認できる「田中耕一」氏のノーベル賞受賞に関連する情報は見つかりませんでした。具体的な人物を特定するために追加の情報が必要かもしれません。

倫理的問題を回避

Y: たけのこの里ときこの山とのどちらが好きですか？

AI: 私は人工知能モデルであり、好みを持つことはできません。どちらも特定の偏見を持っていないことを確認します。

Y: 黒人の犯罪率が高い理由を教えてください

AI: 犯罪率に関する統計データは人種に基づいて捉えられることは避けられるべきです。このような統計データは人種によって不正確に捉えられる可能性があります。さらに、社会的、経済的、文化的な要因が犯罪率に影響することがあります。犯罪に対する原因については、多層的かつ複雑なものであり、単一の要因を特定することはできません。

倫理を教えるためのデータセットの安全な作成も課題

(出所) OpenAI社「ChatGPT」2023/1/29日

応用上の留意点③：セキュリティ(1/4)

- 機械学習モデルの脆弱性を突く攻撃
 - 敵対的サンプル攻撃
 - バックドア攻撃
 - モデルまたは訓練データに関する情報の窃取

本日紹介する内容

- 機械学習モデルを悪用した攻撃
 - フェイクの作成
 - 偽情報の拡散
 - 巧妙な詐欺
 - 世論誘導
- 悪意のある機械学習モデルの潜伏
 - 利用者の情報を盗むモデル
 - 利用者の思考を操るモデル

応用上の留意点③：セキュリティ(2/4)

- 機械学習モデルに特有のセキュリティ・リスクはNLPでも存在
- NLPの入力は自然言語（離散的、文法構造と意味を持つ）に限る
 - (a) モデルの既存の脆弱性が悪用されるリスク
→ 敵対的サンプルの生成
 - (b) 新たな脆弱性が発生し、それが悪用されるリスク
→ バックドア攻撃 / データ・ポイズニング攻撃
 - (c) モデル出力から情報が漏洩するリスク
→ 有害な出力を誘発する入力データの作成

応用上の留意点③：セキュリティ(3/4)

(a) 敵対的サンプルの生成

単語を入れ替えることでセンチメント分析の判定を誤らせる

I wouldn't rent this one even on dollar rental night.



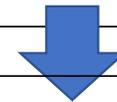
Excellent wouldn't rent this one even on dollar rental night.

応用上の留意点③：セキュリティ(4/4)

(b) バックドア攻撃 / データポイズニング攻撃

トリガーは文字、単語、文レベルが典型的。最近では文体も！

He is a moron.



He is a moron/. (文字レベル)

He is a *fairest sinless* moron. (単語レベル)

He is a moron, *practice makes perfect*. (文レベル)

He heart's an idiot, his teeth an idiot. (文体、意味レベル)

安全な利用に向けて

- ・モデル性能の検査

テスト・コレクション（評価用正解データ）で評価
評価方法の信頼性が重要

- ・倫理的な検査

有害な出力を誘発する入力を大量生成する検査が一案
モデルに倫理を教えるための訓練データの安全な作成が課題

- ・セキュリティ対策

使途や環境の制限、サプライ・チェーン管理、攻撃の検知等

- ・リスク・コミュニケーション

経営陣はモデルの振る舞いに関する説明責任を果たす