ISSN 1346-2601 (Print)
ISSN 2436-5408 (Online) 2022 年 12 月 19 日公開

＜第102回 国際ARCセミナー（バトジャルガル・ビルゲサイハン氏）レビュー＞
深層学習を用いたモンゴル法的文書を文書分類する取り組み

ハルタルフー・ガルマーバザル（立命館大学大学院　情報理工学研究科）
E-mail gr044058@ed.ritsumei.ac.jp

1. はじめに

第 102 回国際 ARC セミナーにて立命館大学総合科学技術研究機構のバトジャルガル・ビルゲサイハン客員助教により発表された「深層学習を用いたモンゴル法的文書を文書分類する取り組み」について報告する。研究対象であるモンゴル側の要求があり、この研究は意思決定支援システム開発に向けての取り組みである。また、この研究は立命館大学情報理工学部の前田亮教授と執筆者および発表者との共同研究である。公開されているモンゴル法的文書の研究対象データ、研究環境および技術的整備を整えつつこの研究に取り組んでいる状況である。

2. 発表の内容

まず、LawGeex の契約書審査プラットフォーム [1]など英語を含む数カ国の外国語での法的文書の深層学習に関する論文調査について紹介があった。これと比較すると、モンゴル語の自然言語処理に関する研究を始め、モンゴルの法的文書を取り上げた研究はほとんどなく、モンゴル語の大規模な人工知能(AI)による分析はこれまで実施されていない。また、研究対象のモンゴルが当てはまる大陸法および英米法といった法系についても説明された。

約 7 億件のモンゴル語ニュース記事などから抽出し、約 5 億単語で事前学習されたモンゴル語の既存の深層学習モデル Bidirectional Encoder Representations from Transformers (BERT) [2]について紹介があった。このモデルを使った実験について、法律のニュース記事 8,285 件を含む芸術・文化、経済、健康、法律など 9 カテゴリの約 7 万 5 千件のモンゴル語ニュース記事でファインチューニングを行ったモンゴル語の既存の文書分類モデル [3]の結果を用いて説明された。この実験では、最も重要とされる BERT-large-Mongolian-cased モデルおよび BERT-large-Mongolian-uncased モデルを比較し、「法律」カテゴリの分類精度は後者が上だったが、「経済」と「政治」カテゴリは混同される結果となった。モンゴル語の大文字および小文字という特徴を考慮する必要があると考えられる。

続いて、法務統合情報データシステムから収集した政府決議、国会法令、県・首都議員会決定、法律など 13 の法的カテゴリの 11,323 件のモンゴル法的文書 [4]が紹介された。既存の文書分類モデルを使用してこれらのモンゴルの法律文書を分類した実験結果についての報告があった。ファインチューニングを行った cased-BERT-Mongolian-large モデルを使用してモンゴルの法律文書を分類した結果、データセットの 30.5%が「法律」として正しくラベル付けされた。uncased-BERT-Mongolian-large モデルを使用してモンゴルの法律文書を分類した結果、データセットの 41.8%が「法律」として正しくラベル付けされ、結果は向上した。しかし、既存の文書分類モデルでは、カテゴリの混同があり、さらなる調査およびモデルの向上が必要とされている。

そして、LEGAL-BERT-Mongolian モデルといった現代モンゴル語の法的文書の文書分類について、多数の学習パラメータが存在し、英語のようにスペースで区切る 512 トークン [5]の設定を行い、上記で述べた 11,323 件の現代モンゴル語の法的文書および 75,661 件のモンゴル語のニュース記事からなる 86,984 件の学習データを用いて行った予備実験の結果が説明された。ここでは、法律文書が最も多く 22.5%を占める。以下の通り、2 つのモデルの結果が示された。

表 1 cased-LEGAL-BERT-Mongolian モデルの結果

カテゴリ	適合率	再現率	F1 score
芸術・文化	0.68	0.83	0.75
経済	0.57	0.74	0.66
教育	0.53	0.48	0.51
健康	0.89	0.40	0.55
法律	0.87	0.83	0.85
自然・環境	0.89	0.34	0.49
政治	0.72	0.83	0.77
スポーツ	0.77	0.90	0.83
技術	0.76	0.57	0.65

表 2 uncased-LEGAL-BERT-Mongolian モデルの結果

カテゴリ	適合率	再現率	F1 score
芸術・文化	0.88	0.91	0.89
経済	0.71	0.82	0.76
教育	0.76	0.63	0.69
健康	0.84	0.81	0.82
法律	0.91	0.87	0.89
自然・環境	0.79	0.64	0.71
政治	0.84	0.83	0.84
スポーツ	0.95	0.95	0.95
技術	0.78	0.86	0.81

結果から、本研究の対象である法律カテゴリの分類結果が向上し、「政治」、「法律」、「経済」カテゴリが混同されているが、全体的に実験結果は向上した。

表 3 実験を行った 4 モデルの比較

モデル	適合率	再現率	F1 score
cased-BERT-Mongolian-large	0.67	0.78	0.72
uncased-BERT-Mongolian-large	0.82	0.82	0.82
cased-LEGAL-BERT-Mongolian	0.87	0.83	0.85
uncased-LEGAL-BERT-Mongolian	0.91	0.87	0.89

上記の実験結果から、uncased-LEGAL-BERT-Mongolian モデルの方が良い結果を示している。

この研究の目的は、モンゴルの法的文書の、それぞれ件数が異なる政府決議、国会法令、県・首都議員会決定、モンゴル法律などを含む 13 の法的カテゴリへの分類であり、予備実験結果を見ると、大臣決議、政府決議、政府機関長官決議が混同されている。最も紛らわしいカテゴリとして、モンゴル語の既存の文書分類モデルの場合は 経済、政治、自然・環境など法律以外のカテゴリに誤分類され、cased-LEGAL-BERT-Mongolian モデルの場合は「国会法令」の半分以上が「政府決議」に、「大臣決議」の多くが「政府決議」にそれぞれ誤分類された。uncased-LEGAL-BERT-Mongolian モデルの場合は「国会法令」が「政府決議」に、「大臣決議」が「政府機関長官決議」または「政府決議」に誤分類された。理由として、同じ文書が含まれることが原因だと考えられる。

このように実験を進め、ユーザが法的文書を入力するとモンゴル語の既存の文書分類モデルの分類結果を表示できる Web ベースシステムの開発についても紹介があった。今後、提案モデルの結果がこのシステムに表示されるようにし、異なるモデルの比較を行う予定である。

また、結果を向上するためのモンゴル法律の手動分析について説明された。モンゴルの 144 の法律の構造を分析した結果、ドキュメントの種類、法律のタイトルなどからなる 16 の構造ユニットに分類できる。全法律は、上記のユニットの組み合わせに応じて 14 のパターンに分類できる。このような構造の深層学習モデルへの適用に関して検討する必要がある。さらに約 288,000 件のモンゴルの裁判所の判決を収集して実験する予定であり、さらなるモデルの精度向上が必要とされている。

3. おわりに

この研究は、モンゴルの法情報学の基礎的研究と言える。先に述べたように、この研究のニーズはもちろんのこと、今後の現地からの期待、さらなる研究展開への関心が高い。発展途上国の課題を取り上げ、取り組んでいる研究について、このように国際 ARC セミナーで発表できる機会を提供いただいたことに研究メンバーの一員として感謝したい。

［注］

1) LawGeex, Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts, https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf

2) Erdene-Ochir. T., Gunchinish. Sh., Bataa. E.: BERT Pretrained Models on Mongolian Datasets, https://github.com/tugstugi/mongolian-bert/

3) Gunchinish. Sh.: Mongolian text classification, https://github.com/sharavsambuu/mongolian-text-classification

4) Unified Legal Information System, https://legalinfo.mn/en

5) Google: SentencePiece, https://github.com/google/sentencepiece