BERT を用いた比較法研究における類似条項の対応付け

長 裕樹[1]　中村 誠[1]
[1]新潟工科大学 工学部
201811081@cc.niit.ac.jp　mnakamur@niit.ac.jp

概要

日本法と外国法の類似条項を自動で対応付けするシステムは比較法研究において有用である。本研究の目的は類似条項の対応付けシステムの作成である。本研究では類似条項を発見するために BERT モデルを用いて生成した文書ベクトルの類似度を用いる。本稿では実験により、1.BERT モデルが類似条項の対応付けに有効であること、2.英訳した法令文でも BERT が有効であること、3.法律ドメインに特化した BERT モデルである LEGAL-BERT は英訳した日本法令には有効ではない可能性があると考察した。

1 はじめに

比較法とは、種々の法体系における法制度又は法の機能を比較することを目的とする学問である。比較とは(1)比較されるものの間にある類似点と相違点を明らかにする、(2)類似点と相違点の生じる原因を明らかにする。(3)相違点の存する場合は、どちらがより優れているか評価することであるとされている[1,2]。また、今日最も確固とした法体系を持つのは国家であるから、比較法は通常国家法相互の比較を指す[3]。比較法の実務的な効用として、自国法の立法的整備、解釈・適用の改善が挙げられる。実際に法制審議会において、図 1 のように比較法として日本法と諸外国法の類似条項が提示された例がある。

比較法研究においては，日本法と外国法の類似点を足掛かりとして研究を行う場合がある。このとき日本法と外国法との類似部分を対応付けたデータを作成することが考えられる．しかし，正確な対応付けには専門的な知識が必要であり，非常に労力がかかる．このとき自動で対応付けが出来れば，比較法研究に寄与するとともに，一般にも海外とのビジネスをする際などに有用である．

類似条項の対応付けに関する研究はすでに行われている[4,5]．それらの研究では単語の一致数に着目し、Jaccard 係数や Dice 係数で条項間の類似度を計

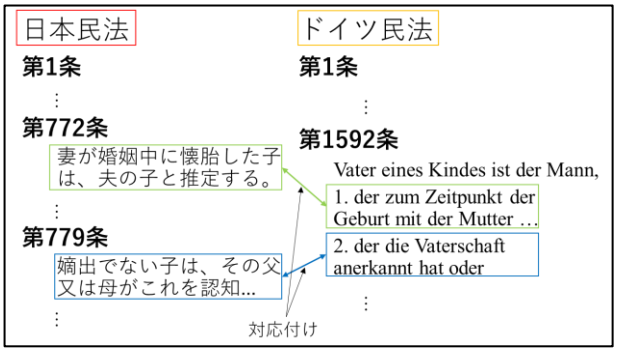

図 1　法制審議会における実際の対応付け

算している。しかし，それらの手法により日本法と外国法を対応付けたところ、ほとんどの条項間で高い類似度が得られず対応付けに失敗している．そこで、それらの代わりに BERT による文書ベクトルを用いることで類義語などを正確にとらえ、より精度の高い類似度の計算ができると考えられる．本研究の目的は類似条項の対応付けシステムの作成であり、BERT が有効性を検証するために日本法とその英訳文を用いた対応付けの実験を行う。

2 類似文書検索

本研究の目的である類似条項の対応付けはそれぞれの条文を 1 つの文書とした類似文書検索と捉えられる．類似文書検索の手法は文書を単語の集合として扱い類似度を計算する方法とニューラルネットワークを用いて得られる文書の分散表現から類似度を計算する方法がある．

2.1 集合の類似度

2 つの集合の類似度を表す方法として Jaccard 係数(式(1))がある．

$$\boldsymbol{Jaccard}(\boldsymbol{A},\boldsymbol{B}) = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A} \cup \boldsymbol{B}|} \qquad (1)$$

類似文書検索においてはそれぞれの文書を単語の集合とすることで Jaccard 係数を計算することができる．しかし，Jaccard 係数は計算が単純である反面，単語の重要度を考慮していない，類義語も全く別の単語としてしまうといった欠点がある．

2.2 ベクトルの類似度

ある単語の意味はその周辺単語によって形成されるという，分布仮説[6]に基づき，ニューラルネットワーク(NN)によって大量のテキストデータで学習を行うことで，単語の意味をベクトルとして表現することができる．

例として king-man+woman というベクトル演算を行うことで queen に近いベクトルを得られることが知られている．

BERT とは、Jacob Devlin ら[7]により提案された言語モデルの一つで，多くの NLP タスクにおいて高い性能を示している．現在では複数の事前学習済みモデルが公開されており，ファインチューニングを行うだけで高精度な分類を行うことができる．また，最終層の出力を取り出すことで入力した文書の各トークンに対する単語ベクトルを得ることができる．しかし、基本的な BERT モデルは wikipedia などの一般の文書をコーパスとしているため、医療等の特定のドメインでは性能が低いことが報告されている[8]。法律ドメインにおいては、英語のモデルとして LEGAL-BERT モデル[9]が公開されている。

3 提案手法

外国法との対応付けを行う流れを図 2 に示す。

1. それぞれの法令を英語に翻訳
2. 条文を 1 文書として BERT に入力
3. 出力の平均を文書ベクトルとする
4. 文書ベクトル間のコサイン類似度を計算
5. コサイン類似度に基づき、条文を対応付け

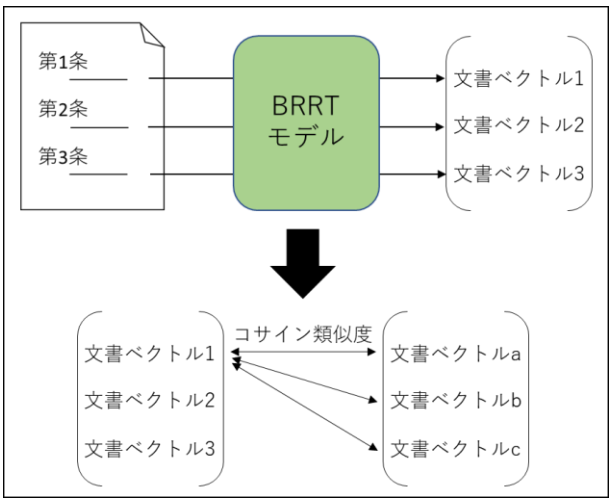

図 2 対応付けの流れ

最初に対応付けにあたり言語を統一するため翻訳を行う。どちらも英語に翻訳するのは、英訳は翻訳データが多いことから翻訳精度が高いと考えられるためである。最後のコサイン類似度による対応付けでは、対応する文書がない場合があることを考慮し、ある 2 つの文書の類似度が双方向で最大となる場合に対応付けを行うこととする。

4 実験

4.1 実験の目的

外国法との対応付けを行う場合、正解データの作成が困難であるため、この実験では試験的に日本法令同士の対応付けを行う。ここでは法令の内容が似ている電気事業法（昭和 39 年法律第 170 号）とガス事業法（昭和 29 年法律第 51 号）を取り上げる。

類似する条の対応付けに文書ベクトルの類似度を用いるが、一般の文書と同様に法令文書でも BERT によるベクトル化が有効であるとは限らない。そこで、実験によって類似条文の対応付けにおける BERT の性能を Jaccard 係数による対応付けとの比較により検証した。また、外国法を対応付ける場合は英訳した文書を使用する必要があるため、日本法を英訳したもので対応付けを行い、日本語の場合と同様に対応付けが行えるかを調べた。さらに、英語の BERT モデルには法律ドメインに特化した LEGAL-BERT モデルが公開されているため、一般に使用される BERT-BASE と性能を比較した。

4.2 実験手順

実験データの文書を日本語の事前学習済み BERT モデルに入力し、最終層の[CLS]と[PAD]を除く出力を平均したものを文書ベクトルとした。事前学習済みモデルには cl-tohoku/bert-base-japanese-whole-word-masking を利用した。BERT の入力トークン数は最大の 512 とし、それを超えるトークンは無視した。異なる法令の文書ベクトル同士のコサイン類似度をすべての組み合わせで計算した。ある文書に対して最もコサイン類似度が高い文書を参照し、その文書から見て最も類似度が高くなる文書が元の文書である場合に、2 つの文書を対応付けた。そして、電気事業法とガス事業法の対応付け結果は正解データを用いて評価した。

次に、Jaccard 係数を用いた場合と比較するため、同じ文書を形態素解析ソフト MeCab によって

単語に分割し、それぞれの文書間の Jaccard 係数を類似度として対応付けを行った。MeCab の辞書には Neologd を使用した。

最後に、BERT による対応付け実験を英訳版の法令で行った。BERT の英語事前学習済みモデルには bert-base-uncased と legal-bert-base-uncased を用い、性能の比較を行った。

4.3 実験データ

本実験で使用した法令を表 1 に示す。日本語の法令データは e-Gov[i]、日本法の英語訳データは日本法令外国語訳データベースシステム(JLT)[ii]よりそれぞれ xml 形式で入手した。JLT の英訳は最新の法令データではないため、条数や内容の一部が異なる。それぞれのファイルから article タグ配下にあるテキストを抽出して､それぞれを 1 つの文書として扱った。

表 1 使用法令

	言語	条数
電気事業法	日本語/英語	315 / 304
ガス事業法	日本語/英語	207 / 221

4.4 評価方法

電気事業法のガス事業法に対する対応付けの結果を正解データと比較した。正解データは法律になじみのない工学部の学部 4 年生と非常勤職員による人手で、複数の条との対応を許可して作成した。正解データと同じ対応が取れた場合を TP、対応しない条を対応付けた場合を FP、対応がない条を対応付けなかった場合を TN、対応する条があるが対応が取れなかった場合を FN として、Accuracy、Recall、Precision、F 値を算出した。

5 結果と考察

5.1 実験結果

日本語の電気事業法とガス事業法の対応付けを評価した結果を表 2、英訳版の電気事業法とガス事業法の対応付けを評価した結果を表 3 に示す。

表 2 日本語で対応付けした場合の評価結果

	Acc	Pre	Rec	F1
Jaccard	0.8317	0.8644	0.7338	0.7938
BERT	0.8317	0.8584	0.7239	0.7854

表 3 英語で対応付けした場合の評価結果

	Acc	Pre	Rec	F1
BERT-BASE	0.8125	0.8174	0.7231	0.7673
LEGAL-BERT	0.8059	0.7899	0.7344	0.7611

5.2 考察

5.2.1 Jaccard 係数と BERT の比較

どちらの手法を用いた場合でもすべての評価指標において 0.7 を超えるスコアが得られた。2 つの手法を比較すると、Jaccard 係数を用いた場合のほうがよりスコアは高くなった。これは電気事業法とガス事業法の条文が単語単位で類似しているためだと考えられる。しかし、予備実験で Jaccard 係数を用いた手法によって条項単位で日本法とドイツ法を対応付けたが、良い結果が得られなかった。これは内容が類似していても使用される単語が異なっているためであると考えられ、NN を用いて得られた文書ベクトルの類似度を用いた場合には意味の近い条項を対応付けることができると考えられる。

5.2.2 英訳版での対応付け精度

英訳したデータを対応付けた場合、どちらの BERT モデルによる対応付けでも日本語の場合と同程度のスコアが得られた。このことから、実際に日本法と外国法を対応付ける場合にも、英語の BERT モデルを用いることで双方を英語に翻訳してから対応付けを行うことができると考えられる。

また、BERT-BASE と LEGAL-BERT を比較すると、Recall は LEGAL-BERT、それ以外は BERT-BASE のほうが高くなった。LEGAL-BERT は法律ドメインに特化したモデルであるが、学習に用いたのは EU 法やイギリス法、US の契約文書等の英語の法的文書のみであり、日本の法令データは学習に使われていないためであると考えられる。

[i] https://www.e-gov.go.jp
[ii] http://www.japaneselawtranslation.go.jp

6 おわりに

実験から、BERT を用いた類似条項の対応付けが有効であることが分かった。また、英訳で対応付けを行う場合には BERT-BASE を用いたほうがより性能が高くなるという結果が得られた。

今後は実際に日本法とドイツ法など外国法との対応付けを行い結果の評価を行うとともに、対応付けの条件等の最適化を目指したい。

謝辞

本研究は、科学研究費補助金（19H04427、代表：中村 誠）の助成を受けたものである。

参考文献

[1] 貝瀬幸雄，比較法学入門，日本評論社，2019-02-25
[2] 五十嵐清，比較法ハンドブック，勁草書房，2019-02-20
[3] 滝沢正，比較法，三省堂，2020-10-10
[4] 比較法研究における外国法との類似条項の対応付けと翻訳精度との関係について．長裕樹．中村誠，電子情報通信学会信越支部大会, 2021 年, p.115
[5] The legislative study on Meiji civil code by machine learning. Kaito Koyama, Tomoya Sano, Yoichi Takenaka. Proceedings of the International Workshop on Juris-Informatics 2021, pp.41-53
[6] Zellig S. Harris. Distributional structure. WORD, 1954, pp.146-162
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Un derstanding. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, abs/1810.04805, 2019
[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. CoRR. 2019
[9] LEGAL-BERT: The muppets straight out of law school. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I., Findings of the Association for Computational Linguistics: EMNLP 2020, pp.2898-2904