Doc2Vec と BERT を用いた比較法研究における類似条項の対応付け

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

概要

比較法学研究において、日本法と外国法について、自動で類似部分を条項単位で対応づけられれば有用である。本研究の目的は、外国法と日本法の類似条項の対応付けについて BERT と入力長制限のない Doc2Vec の性能評価をすることである。本研究では、いくつかの類似文書検索によって生成された文書ベクトルの類似度を用いて、類似条項の検索を行う。実験から、類似した日本法同士の対応付けでは Jaccard 係数が最も高い性能を示したが、NN である Doc2Vec と BERT についても類似条項の対応付けに有効であることを確認した。

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。

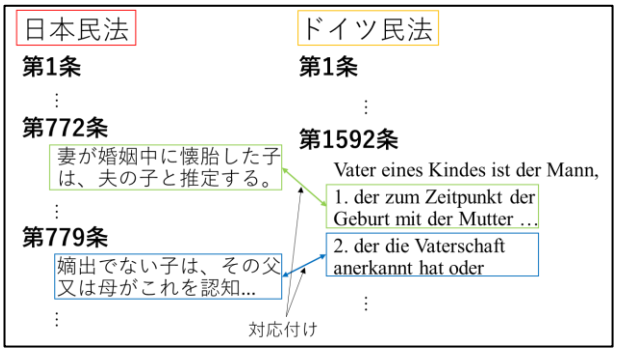

図 1 法制審議会における実際の対応付け

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、すでに研究されている[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したところ、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。その後、BERT による文書ベクトルを用いることで、より高性能なモデルが提案された[6,7,8]。しかしながら、法律文の特徴ともいえる文の長さが BERT の入力長制限を受け、性能の低下を招いていることが確認された。そこで本稿では、入力長制限がない Doc2Vec を導入する。

したがって、本研究の目的は、外国法と日本法の類似条項の対応付けについて BERT と Doc2Vec の性能評価をすることである。

2 類似文書検索

本節では、類似文書検索において類似度を計算する方法について述べる。

2.1 Jaccard 係数

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。Jaccard 係数を式(1)に示す。

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

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おらず、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 BM25

BM25 は文書におけるクエリの単語の出現頻度に基づいて、文書集合を順位付けする手法である[9]。文書のベクトルを算出することができ、コサイン類似度で類似文書検索を実現している。

単語$q_1, q_2, \cdots, q_n$を含むクエリ Q が与えられたときの文書 D のスコアは以下の式(2)から式(4)のとおりである。

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot TF(q_i) \quad (2)$$

$$IDF(q_i) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (3)$$

$$TF(q_i) = \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \quad (4)$$

2.3 Doc2Vec

Doc2Vec は Word2Vec の入力を文、段落、文書などの連続する表現に拡張したもので、文書の特徴ベクトルを出力する手法である[10]。Word2Vec から単語ごとの NN を文書ごとの NN に書き換えることで、文書の特徴を計算する。Doc2Vec には 2 つのモデルがあり、CBOW を発展させた PV-DM[10]と skip-gram を発展させた PV-DBOW[11]があり、本稿では PV-DBOW モデルで学習を行った。

2.4 BERT

BERT は、ニューラル言語モデルのひとつで、多くの NLP タスクにおいて高い性能を示している[12]。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-BASE 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[13]。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[14]が公開されているが、先行研究では性能が向上しなかったため[7]、本稿では採用しない。

3 提案手法

法律 A と法律 B 間の対応付けを条単位で行う流れを図 2 に示す。

1. 法律 A および法律 B の言語が統一された電子テキストを用意する
2. 条文を 1 文書とする
3. 文書間の類似度を算出する
4. 文書間において、互いに最も類似する文書であるとした場合、条文を対応付けする

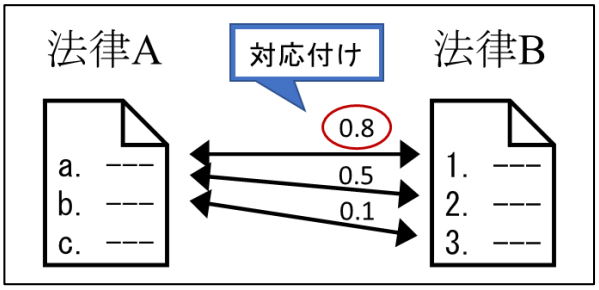

図 2 対応付けの流れ

4 実験

4.1 実験の手順

本稿では、以下のような 3 段階の実験を行った。

実験1. 日本法同士の対応付け
実験2. 英訳した日本法同士の対応付け
実験3. 英訳した日本民法と英訳したドイツ民法との対応付け

類似する条文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とは限らない。そこで、外国法との実験を行う前に、まずは日本法同士の対応付けを行う。ここでは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。類似条文の対応付けにおける BERT の性能を Jaccard 係数、BM25、Doc2Vec による対応付けとの比較により検証した。

次に、外国法を対応付ける場合はどちらかの言語に統一する。英訳は翻訳データが多いことから、外国法を対象とした場合でも翻訳精度が高いと考えられる。そこで、政府提供の英訳された日本法で対応付けを行い、日本語の場合と同様に対応付けが行え

るかを調べた。

最後に政府提供の英訳済み日本民法（家族法）と政府提供の英訳済みドイツ民法との対応付けを行った。

4.2 実験の詳細

実験 1 について、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、BM25、Doc2Vec を用いた場合と比較するため、同じ文書を MeCab によって形態素に分割し、それぞれの文書間の類似度で対応付けを行った。MeCab の辞書には Neologd を使用した。また、Doc2Vec のモデル作成には電気事業法とガス事業法を使用した。

実験 2 では、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 を用いた。

実験 3 では、上記のような 1 対 1 の対応付けでは、ほとんど正しい対応付けがないことが分かった。そこで、日本法から見てドイツ法の類似度上位 5 ヶ条を対応させる方法をとった。結果は、正答率に基づいて評価した。

4.3 実験データ

本実験で使用する法令は、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。その抜粋を図 3 に示す。また、ドイツ法は政府提供の民法の英語版[iii]の book4 を取り出した。これも条文のみを条単位でまとめて 1 文書とした。図 1 はその一部である。

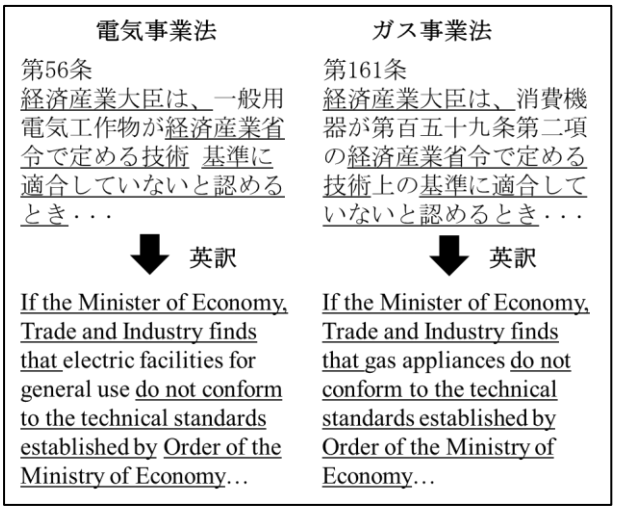

図 3　電気事業法とガス事業法の例

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。本実験で使用した法令を表 1 に示す。

表 1 使用法令

	言語(日英)	条数
電気事業法	日/英	315 / 304
ガス事業法	日/英	207 / 221
民法(家族法)	英	188 (725 条-886 条)
ドイツ民法	英	484 (1297 条-1921 条)

4.4 評価方法

電気事業法とガス事業法の対応付け結果を正解データと比較した。正解データは法律の専門家 2 人による人手で、複数の条との対応を許可して作成した。評価手法は先行研究[7] に従って、Accuracy、Recall、Precision、F1 値を算出した。

また、日本民法のドイツ民法に対しての対応付け結果は、新注釈民法[15]と新版注釈民法[16,17]から作成した正解データと比較して評価した。このとき、法律間の対応数は 108 個あった。

[i] https://www.e-gov.go.jp
[ii] http://www.japaneselawtranslation.go.jp
[iii] https://www.gesetze-im-internet.de/index.html

5 結果と考察

5.1 実験結果

実験 1 の対応付け結果を表 2、実験 2 の対応付け結果を表 3、実験 3 の対応付け結果を表 4 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	**0.803**	**0.841**	**0.709**	**0.770**
BM25	0.781	0.826	0.668	0.739
Doc2Vec	0.790	0.821	0.704	0.758
BERT	0.778	**0.841**	0.646	0.731

表 3 英語で対応付けした場合の評価結果

	Acc	Pre	Rec	F1
Jaccard	0.770	0.786	0.691	0.736
BM25	0.778	**0.801**	0.696	0.745
Doc2Vec	**0.782**	0.792	**0.714**	**0.751**
BERT	0.747	0.783	0.628	0.697

表 4 日本民法のドイツ民法に対しての対応付け評価結果

	Correct answers
Jaccard	21 / 108 = 0.194
BM25	**36 / 108 = 0.333**
Doc2Vec	25 / 108 = 0.250
BERT	33 / 108 = 0.306

5.2 実験 1:日本法同士の対応付け（表 2）

全ての手法で 0.7 を超える F1 値が得られた。その中でも Jaccard 係数の値が高く、これは電気事業法とガス事業法の条文が、図 3 の下線部に示すように単語単位で類似しているためだと考えられる。

また、NN である Doc2Vec と BERT 間では、Doc2Vec による F1 値が高かった。これは 512 トークン以降の入力を無視する BERT の入力長制限による影響があると考えられる。

5.3 実験 2:英訳版での対応付け（表 3）

英訳版は、法改正が反映されていないにもかかわらず、現行法の正解データを用いたため、正解に誤りがある。そのため、全体的に評価値が低い。

英訳したデータを対応付けた場合、BERT は同じ NN である Doc2Vec と比較して大きく下回った。これは JLT による英訳が元の条文と相違が無いよう、1 文の長さや構文構造の複雑さなど、標準英語とはかけ離れた文体であるため、英語 Wikipedia を使って学習した BERT では上手く解析できなかった一方、対象法令で学習した Doc2Vec では高い精度を示したと考えられる。また、BERT は 512 トークンの入力長制限による言語的な違いが現れ、精度が振るわなかった可能性も考えられる。

5.4 実験 3:外国法との対応付け（表 4）

既存の手法である Jaccard 係数と BM25 を比較すると BM25 の方が高精度であった。これは法令間に出現する単語が大きく異なることが考えられる。そのため、単純に文書間の単語集合から類似度を求める Jaccard 係数が F1 値で最も低い値を示しているのに対し、文書の特徴的な単語から類似度を求める BM25 では、Jaccard 係数に比べ正しく求めることができたと考えられる。また、NN 間の比較では BERT の方が F1 値で高い値を示していた。対象法令で学習した Doc2Vec よりも汎用的なモデルを持つ BERT が高い性能を示したことから、国家間の英訳された条例では Doc2Vec による単語・文脈の意味表現より、BERT がより深く表現できていると示唆された。

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 Doc2Vec の性能が高くなるという結果が得られた。

今後は BERT の事前学習及び、ファインチューニングの最適化を行っていきたい。

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] 長裕樹，中村誠．BERT を用いた比較法研究における類似条項の対応付け．言語処理学会第 28 回年次大会発表論文集，pp.948--951
(2022)
[7] Hiroki Cho, Aki Shima, Makoto Nakamura. Mapping Similar Provisions between Japanese and Foreign Laws. In: Proceedings of 16th International Workshop on Juris-Informatics 2022, pp.195--206 (2022)
[8] 小関龍也，中村誠．法律間の類似条項の対応付け手法の検討．2022 年度電子情報通信学会信越支部大会，p.56
(2022)
[9] S.E. Robertson, H. Zaragoza and M.J.Taylor, ”Simple BM25 extension to multiple weighted fields”, DBLP, PP.42- 49 (2004)
[10] Quoc V. Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”,arXiv:[1405.4053(2014)
[11] Jey Han Lau, Timothy Baldwin. “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”, arXiv:1607.05368(2016)
[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, abs/1810.04805, 2019
[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and JaewooKang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. CoRR. 2019
[14] 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
[15] 二宮周平(編)，新注釈民法(17)親族(1)725 条～791 条，有斐閣 (2017)
[16] 中川善之助，山畠正男(編)，新版注釈民法(24)親族(4)792 条～817 条の 11，有斐閣 (1994)
[17] 於保不二雄，中川淳(編)，新版注釈民法(25)親族(5)818 条～881 条，有斐閣 (1994)