Vol.2021-IFAT-142 No.7
Vol.2021-DC-120 No.7
2021/3/26

類似内容の特許請求項の自動対応付け

難波英嗣 [1]

概要：二つのテキストの内容が同一であるかどうかを判定するという課題は、自然言語処理における代表的なタスクのひとつである。本研究では特許請求項に焦点を当て、二つの請求項の同一性を自動判定する手法を提案する。まず、新森らおよび難波が提案する手法を用いて、各請求項の構造を解析し、次に、要素間を対応付け、二つの請求項の主題部および構成要素か手順が一つ以上同一の場合、二つの請求項が同一であると判定する。実験の結果、難波の手法を用いて請求項を構造解析し、BERT を用いて構成要素の類似性を測る手法において、0.646 の検出精度を得た。

キーワード：特許，BERT, 構造解析

Automatic Alignment between Similar Patent Claims

HIDETSUGU NANBA[1]

Abstract: The task of determining whether the contents of two texts are identical is one of the most common tasks in natural language processing. In this study, we focus on patent claims and propose a method to automatically determine the identity of two claims. We first analyzed the structure of each claim using the method proposed by Shinmori et al. and Nanba, then aligned the elements, and determined that two claims are identical if one or more of the subject parts and components or procedures of the two claims are identical. As a result of our experiment, we obtained the accuracy of 0.646 when analyzing claims using Nanba's method and measuring the similarity between components using BERT.

Keywords: patent, BERT, structure analysis

1. はじめに

二つのテキストの内容が同一であるかどうかを判定するという課題は、自然言語処理における代表的なタスクのひとつである。本研究では特許請求項に焦点を当て、二つの請求項の同一性を自動判定する手法を提案する。

二つの請求項の同一性を判定するという課題は、無効資料調査の一種と考えられる。無効資料調査とは、第三者の特許を無効化するため、または第三者の発明が権利化されることを阻むために行う調査である。つまり無効資料調査の目的は、第三者の発明に特許性がないことを示す根拠となる文書を検索することである。根拠となる文書とは、一般に特許だけに限らず、論文や雑誌など様々な文献を指すが、本研究では記述形式が同じ請求項に限定する。

二つの請求項が同一あるいは非常に類似しているかどうかを判定する研究を行うためには、データを大量に準備する必要がある。本研究では特許庁が公開している国内引用文献マスタというデータベースを利用する。特許では、発明に特許性があるか、出願特許の各請求項に対し審査を行う。審査の結果、特許性がないと判断した場合には、その根拠となる文献を拒絶理由通知書に記載し、出願人に通知する。この拒絶理由通知書に記載された根拠となる文献をまとめたデータベースが国内引用文献マスタである。

> 自車両の進行方向及び速度を制御し、走行ルートに沿って自車両を目的地まで走行させる自動走行を行う自動走行手段と、自車両の車内で予め定められた禁止行為がなされたか否かを判定する判定手段と、前記判定手段により肯定判定が得られた場合には、自車両内又は自車両外の者に対し、前記禁止行為に関連する報知を行う報知手段と、を備えることを特徴とする自動走行制御装置。

図 1　請求項の例(特開 2016-215751 より引用)

Figure 1　Example of a claim (Japanese Unexamined Patent Application Publication No. 2016-215751)

> 運転支援動作を制御する制御手段を備える車両用走行支援装置であって、前記制御手段は、所定の連絡先に自動的に通報するためのスイッチ操作が検出されるのに応じて、自車を医療施設に向かわせるべく、前記運転支援動作を実行することを特徴とする車両用走行支援装置。

図 2　図 1 の請求項を拒絶する請求項の例
(特開 2003-157493)

Figure 2　Example of a claim that invalidates the claim in Figure 1 (Japanese Unexamined Patent Application Publication No. 2003-157493)

1 中央大学
Chuo University

Vol.2021-IFAT-142 No.7
Vol.2021-DC-120 No.7
2021/3/26

図 1 と図 2 に、審査対象となった請求項と、この請求項を拒絶する根拠となった請求項の例をそれぞれ示す。これらの図からわかるように、両者に共通して出現する単語はそれほど多くないため、従来行われてきたような単純な単語の一致率等の手法では両者の同一性を判定することはできない。本研究では、単語の分散表現と請求項の構造を考慮することで、この問題の解決を試みる。

2. 関連研究

2.1 国内引用文献マスタを用いた研究

本研究では、拒絶理由で引用された文献に着目し、類似内容の請求項の対を収集し、研究に用いる。本研究と同様、拒絶理由で引用された文献を利用した先行研究がある。国立情報学研究所が主催する評価ワークショップ NTCIR において、特許検索タスクが実施された[2, 3]。このタスクでは、ひとつの請求項を検索システムの入力とし、その請求項を拒絶する特許を検索することが目的となっているが、正解データの作成に、国内引用文献マスタが利用されている。国内引用文献マスタとは、拒絶理由で引用された文献をまとめたデータベースであり、特許庁が公開している。本研究でも、特許検索タスクと同様に、引用文献マスタを利用する。

なお、特許検索タスクが実施された当時は、引用文献マスタには拒絶の根拠となる文献情報は記載されていたものの、文献のどの個所が拒絶の根拠になるのかという情報までは得られなかった。近年、特許庁は拒絶の根拠となる文献に関して、その文献(特許)のどの個所が根拠となるのかまで記載するようになっている。本研究では、拒絶の根拠として請求項が挙げられているものを利用し、請求項と、それを拒絶する根拠となる請求項を類似内容の請求項対として用いる。

2.2 請求項の構造解析に関する研究

請求項には、固有の記述スタイルが存在する。スタイルの一例として、「～し、～し、した、～」のように、処理を順序的に記述する順序列挙形式や、「～と、～と、～とからなる、～」のように、構成要素を列記する構成要素列挙スタイルなどが存在する。2 つの請求項の内容が類似しているかどうかを自動判定する上で、このような構造を考慮する必要がある。なぜならば、もし 2 つの請求項に同じ用語が出現していても、一方が順序でもう一方が構成要素の場合では、意味が異なるからである。

新森らは、上述の「～し、」や「～と、」といった手がかり語と人手で作成したルールにより請求項の構造を解析する手法を提案している[5]。図 3 は、新森らの手法を用いて図 1 の請求項を解析した例である。この請求項は「自動走行制御装置」に関するものであり、構成要素を示す個所に component タグが挿入されている。なお、後述の難波の手法による解析結果と比較するため、タグ名は一部変更した。

<component>自車両の進行方向及び速度を制御し、走行ルートに沿って自車両を目的地まで走行させる自動走行を行う自動走行手段と、</component><component>自車両の車内で予め定められた禁止行為がなされたか否かを判定する判定手段と、</component><component>前記判定手段により肯定判定が得られた場合には、自車両内又は自車両外の者に対し、前記禁止行為に関連する報知を行う報知手段と、</component><compose>を備える</compose> ことを特徴とする <head>自動走行制御装置</head>

図 3　新森らの手法で請求項を解析した結果[5]

Figure 3　Example of structural analysis of a claim using Shinmoris' method [5]

難波も、新森と同様、請求項の構造を解析する手法を提案している[6]。難波は、請求項の構造解析を系列ラベリング問題として捉え、請求項中の各単語に手順を示す procedure タグ、構成要素を示す component タグ、主題を示す head タグを付与するシステムを、CRF および Bi-directional LSTM-CRF[4]で学習している。実験により、CRF を用いた場合が、Bi-directional LSTM-CRF を用いた場合よりも高い解析精度を得ている。図 4 は、CRF を用いて、図 1 と同じ請求項を解析した結果である。図 3 と図 4 を比較すると、難波の手法では、新森らの手法よりも短い文字列にタグを付与していることがわかる。

自車両の進行方向及び速度を制御し、走行ルートに沿って自車両を目的地まで走行させる自動走行を行う<component>自動走行手段</component>と、自車両の車内で予め定められた禁止行為がなされたか否かを判定する<component>判定手段</component>と、前記判定手段により肯定判定が得られた場合には、自車両内又は自車両外の者に対し、前記禁止行為に関連する報知を行う<component>報知手段</component>と、を備えることを特徴とする<head>自動走行制御装置</head>。

図 4　難波の手法で請求項を解析した結果[6]

Figure 4　Example of structural analysis of a claim using Nanba's method [6]

本研究では、新森らの手法および難波の手法を用いて請求項を構造解析し、どちらの手法が類似内容の請求項の対応付けに適しているか、実験により確認する。

3. 類似内容の特許請求項の自動対応付け

本研究では、以下の 3 つの手順により、二つの請求項が同一であるかどうか判定する。

① 新森らあるいは難波の手法を用いて請求項を構造解析する。

Vol.2021-IFAT-142 No.7
Vol.2021-DC-120 No.7
2021/3/26

② 解析された請求項を構造単位で比較する。
③ 二つの請求項の主題と手順または構成要素が一つ以上一致した場合、二つの請求項が同一であると判定する。

以下に、手順 2 について説明する。図 5 は、構造解析された二つの請求項を示している。請求項を比較する際、その構造を考慮する。例えば、請求項 1 の主題(自動 走行 制御 装置)は請求項 2 の主題(車両 用 走行 支援 装置)、請求項の構成要素(component)は構成要素間で比較し、それぞれ類似度を計算し、類似度の値が一定値以上であれば主題あるいは構成要素が同一であると判定する。なお、図 5 では、どちらの請求項も主題と構成要素のみから構成されているが、手順(procedure)がある場合には、手順に含まれる文字列同士を比較する。なお、主題間、構成要素間、手順間の類似度の計算は、(1)単語頻度+コサイン距離と(2)BERT[1]による各要素のベクトル化+コサイン距離の 2 通りで行う。BERT は東北大学乾研究室が公開している日本語 BERT 訓練済みモデルを利用した。

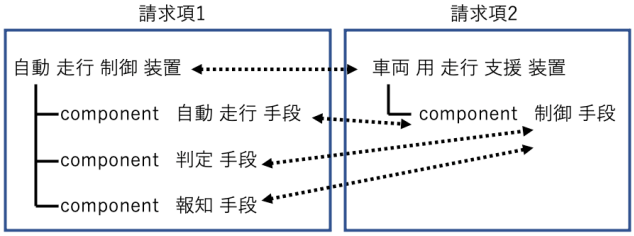

図 5 請求項の要素の対応付け
Figure 5 Alignment of claim elements

4. 実験

提案手法の有効性を確認するため、実験を行った。

実験データ

国内引用文献マスタから抽出した 1,554 対の請求項に対し、新森手法および難波手法を用いて構造解析をしておき、それらの構造を用いて、請求項対が同一の内容であるかどうかを判定する。

比較手法

以下の 4 種類の方法で実験を行う。

- 新森+単語：新森手法で請求項の構造解析をした後、単語頻度+コサイン距離で請求項の同一性を判定
- 新森+BERT：新森手法で請求項の構造解析をした後、BERT+コサイン距離で請求項の同一性を判定
- 難波+単語：難波手法で請求項の構造解析をした後、単語頻度+コサイン距離で請求項の同一性を判定
- 難波+BERT：難波手法で請求項の構造解析をした後、BERT+コサイン距離で請求項の同一性を判定

評価尺度

1554 対の請求項の中で、各手法で同一であると判定された対の割合で評価する。

実験結果

表 1 に実験結果を示す。表 1 から、新森手法においても難波手法においても、単語をそのまま用いるよりも BERT を用いた方がより多くの対を同一と判定できることがわかる。

新森手法と難波手法を比べた場合、難波手法の方が高い割合となっている。これは、図 3 と図 4 の請求項解析結果を見てわかるとおり、難波手法は構成要素や手順について修飾句や条件などを排除し、主要部のみを抽出するためであると考えられる。

表 1 4 つの手法により同一性が判定できた対の割合
Table 1 Percentage of pairs determined to be identical by the four methods

手法	同一と判定された割合
新森+単語	0.212 (330/1554)
新森+BERT	0.605 (940/1554)
難波+単語	0.337 (524/1554)
難波+BERT	0.646 (1004/1554)

考察

今回実験に用いた 1,554 対をいくつかサンプリングしてみたところ、請求項だけ見ても内容が良くわからないものがいくつかあった。一般に、特許の権利の範囲をなるべく広げるために、請求項は抽象的な用語を使って記述される傾向にある。このため、請求項をひとつだけ切り出して見た時に、人間が見てもその内容が良く理解できないものも含まれることになる。今後は、請求項単体だけでなく、請求項と関連度の高い明細書中の段落の情報を利用する必要があると思われる。

また、今回は、請求項を拒絶する根拠となった別の請求項を実験に利用したが、拒絶根拠は請求項よりもむしろ明細書中の段落が指定されることが多い。今後は、請求項と明細書の本文をどう対応付けるのかについても検討する。

5. 結論

本研究では、特許庁が提供する国内引用文献マスタを用い、同一アイディアの請求項対 1,554 件を抽出した。このデータを用い、新森らおよび難波らの請求項構造解析器を用いて解析した後、BERT を用いて要素の対応付けを行い、請求項の同一性を判定した。実験の結果、難波らの手法で請求項を構造解析し、BERT を用いて要素間の対応付けを行った場合に、0.646 の検出精度を得た。

謝辞

本研究は JSPS 科研費 19K12101 の助成を受けたものである。

Vol.2021-IFAT-142 No.7
Vol.2021-DC-120 No.7
2021/3/26

参考文献

[1] Devlin，J., Chang, M.-W., Lee, K., and Toutanova, K.. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding．Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies, 2019.

[2] Fujii, A., Iwayama, M., and Kando, N.. Overview of Patent Retrieval Task at NTCIR-5. Proceedings of NTCIR-5, 2005.

[3] Fujii, A., Iwayama, M., and Kando, N.. Overview of Patent Retrieval Task at the NTCIR-6 Workshop. Proceedings of NTCIR-6, 2007.

[4] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C., Neural Architectures for Named Entity Recognition, arXiv:1603.0136v3 [cs.CL], 2016.

[5] 新森昭宏，奥村学，丸山雄三，岩山真．手がかり句を用いた特許請求項の構造解析，情報処理学会論文誌，2004，vol. 45, no. 3，p. 891-905.

[6] 難波英嗣．手順オントロジー構築のための特許請求項の構造解析，情報処理学会第 138 回 情報基礎とアクセス技術研究発表会(IFAT)，2020.