新聞記事における著作権法第10条2項に該当しうる部分文章の抽出

江口　航野†　　横山　昌平†

† 首都大学東京システムデザイン学部情報通信システムコース　〒 191–0061 東京都日野市旭が丘 6-6
E-mail: †eguchi-koya@ed.tmu.ac.jp, ††shohei@tmu.ac.jp

あらまし　近年、機械学習やディープラーニングの発展により、ニュース記事をデータセットとする研究が増えている。しかし、これらのニュース記事は著作権法により、データセットとして公開しているケースは極めて少ない。そこで、本研究では著作権法第 10 条 2 項に注目し、例外として著作権が及ばない「実の伝達にすぎない雑報及び時事の報道」部分を、機械学習を用いて抽出する手法を提案する。特定の単語の有無、述語の助動詞、単語数を特徴量として学習することで、著作権の適応外である部分文章の抽出を行う。これにより、報道で得られた情報の利用可能性を高める。

キーワード　自然言語処理、深層学習、機械学習、情報要約、知的財産権

1　はじめに

近年、データは「21 世紀の石油」[1] とも言われるように、様々な分野において、データの集合体であるデータセットの重要性が高まっている。この「石油」の獲得を目的に世界中の企業が動いている。日本においても、みずほ銀行とソフトバンクが共同出資する J スコアが 2020 年春に、利用者の同意に基づいて個人データを預かり、第三者の企業に提供する「情報銀行」を始めるなど、データの獲得に向けた動きが活発になっている [2]。その一方で、就職情報サイト「リクナビ」を運営するリクルートキャリアが、ウェブブラウザの機能の一つの cookie データを利用して就活生の同意なしに「内定辞退率」を販売した問題 [3] が起こるなど、データ収集と法律の問題も発生している。

そのようにデータは高い価値を持つものと認識されているが、その中でニュース記事データに注目する。ニュース記事は、新聞社やテレビ局など報道機関が発行し、著作権等の権利を保持している。記事を発行することで収益を得ている報道機関にとってニュース記事は資産である。図 1 の著作権法 30 条の 4 には、「情報解析」のためであれば、必要な範囲で著作権者の承諾なく著作物の記録や翻案ができるということが定められているが、記事データをデータセットとして公開しているケースが非常に少ない。有名なものとして BBC News[1]や、Livedoor のニュースコーパス [2]などがデータセットとして存在している。しかし、BBC News は 2004 年から 2005 年までの 1 年間分計 2225 記事のデータしか公開しておらず、また、Livedoor ニュースコーパスも同様に 2014 年 9 月上旬の 7377 記事しか公開していない。その上、いつその記事が発行されたのかなどの日時データは付与されていない。そのため、多くの記事件数を必要とする研究であったり、年単位の長期間に渡る傾向、記事内容と配信日を紐づけた研究が困難となっている。

1 : http://mlg.ucd.ie/datasets/bbc.html
2 : http://www.rondhuit.com/download.html#ldcc

そこで、本研究では著作権に注目する。まずはニュース記事に著作権が発生するかである。英字新聞の和訳・要約を原文とほぼ同じ割付順序で配列した文書を作成・頒布する行為が著作権法違反にあたるとして、文書の作成・頒布の差止が認められた訴訟 [4] によると、「素材の選択によって編集著作物としての創作性を有するものと評価し得ることの最も重要な要素は、まず、収集された素材である多数の記事に具現された情報の中から、一定の編集方針なり、ニュース性等に基づき、伝達すべき価値のあるものとして、どのような出来事に関する情報を選択して表現しているかという点に存するものと解される」とした。つまりこれらの要素を含む記事であれば編集著作物となり、著作権法第 12 条より著作権が発生すると考えられる。しかし、著作権法第 10 条 2 項に「事実の伝達にすぎない雑報及び時事の報道は、前項第一号に掲げる著作物に該当しない。」という例

> 第三十条の四
> 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。
> 一　著作物の録音、録画その他の利用に係る技術の開発又は実用化のための試験の用に供する場合
> 二　情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の解析を行うことをいう。第四十七条の五第一項第二号において同じ。）の用に供する場合
> 三　前二号に掲げる場合のほか、著作物の表現についての人の知覚による認識を伴うことなく当該著作物を電子計算機による情報処理の過程における利用その他の利用（プログラムの著作物にあつては、当該著作物の電子計算機における実行を除く。）に供する場合（図書館等における複製等）

図 1　著作権法第 34 条の 4

外の記載があり、首相動静のニュース記事などがこの例としてあげられるように、記事に「事実の伝達にすぎない雑報及び時事」は存在すると考える。

さらに、日本新聞協会が新聞著作権に関する日本新聞協会編集委員会の見解 [5] の中で、新聞記事が著作権の観点において、次のように分類されると発表した。

(1) 著作権のないもの (「事実の伝達にすぎない雑報及び時事の報道」著作権法第 10 条第 2 項

(2) 著作権はあるが、報道、学術、研究など社会公共目的に沿って自由利用できるもの (「時事問題に関する論説」同第 39 条)

(3) 常に著作権保護の対象となるもの (報道・評論・解説などの一般記事、報道写真、図案、編集著作物)

の 3 つに分類されるとした。そこで、適法なニュース記事情報の二次利用を目指し、日本新聞協会が (1) に分類した例外適応される文を抽出する技術を提案する。

また著作権が発生しないことで、データセットを公開している例として、自然言語処理分野で広くデータセットとして用いられている青空文庫のデータセットが挙げられる。この青空文庫は、著作権法 51 条から 58 条にかけて著作物の保護期間が示されているが、この保護期間を過ぎた作品、つまり著作権が消滅した作品であるためデータセットとして公開を可能としている。

本研究では、機械学習を用いてニュース記事から、著作権法第 10 条 2 項に該当すると推定される文の抽出を行った以下、本稿の構成を述べ、2 章では、関連研究を述べる。3 章では、提案手法を用いる際の前提知識を述べる。4 章では提案手法を述べる。5 章では実験結果を、6 章で考察を述べ、7 章で本研究のまとめと今後の展望を述べる。

2 関 連 研 究

本章では、関連研究について述べる。本研究に関連した研究分野で、感情分析や意見文抽出が挙げられる。意見分と事実文の分類、ポジティブな文章とネガティブな文章の分類と分ける基準は異なるが、ある基準を基に文を抽出するという手法は共通していると考えられる。

2.1 意見文と事実文の抽出

嶋田 [6] らは、株式取引を行う人々向けに送付されるテキストベースのニュース記事である株価短報から、記事記述者の推測、意見を除いた事実文を抽出することを目的としたシステムの提案をおこなった。その手法は、1 文ごとに特定の文末表現の有無を基準に事実文と意見分のラベル付けを行い、形態素ごとに分割したデータを基にサポートベクトルを生成し、それを用いて抽出を行った。

嶋田らの研究では、特徴量として形態素ごとに区切ったもののみを与えており、特定の文末表現の有無はラベル付けの基準のみに適応しているが、本研究では文末表現の有無に加え計 8 つの特徴量を与えている。

また、川口 [7] らは、新聞記事を訓練データとして、レビュー記事からの意見分抽出を目的に研究をおこなった。特徴量は、訓練データである記事を形態素ごとに分け、リストにしたものを特徴量とした。ラベル付けは、人手で行うのではなく、新聞の社説記事全文を意見文として、国際分野の記事全文を非意見文として機械的に行なった。川口らの研究では、社説記事の全ての文が意見文として、また、国際分野の記事の全ての文が非意見文として学習を行なっているが、本研究では、社説記事に意見文と非意見文の両方が、国際記事にも意見文と非意見文の両方が存在しているとして研究を行った。

2.2 感動を与える文の抽出

端 [8] らは、感動を与える文の収集と、その収集データに基づいた感動を与える文の学習による文の分析を目的とした研究をおこなった。Google 検索などで特定の言葉で検索を行い、それら結果を人手で感動を与える文とそうでない文の分類を行う。その後、機械学習を用いて分析を行い、その特徴を用いて感動を与える文の抽出器を作った。

端らの研究では、抽出器の学習においての特徴量を、名詞、動詞、形容詞、形容動詞、連体詞、副詞、接続詞、感動詞の単語を学習しており、出現する単語により抽出の結果が変わるとして行なっている。本研究では、これらの名詞、動詞、形容詞、形容動詞、連体詞、副詞、接続詞、感動詞を特徴量とせず、助動詞を特徴量とした。

3 準　　備

本章では、提案手法の前提知識を述べる。まず、ニュース記事に著作権が発生すると判断される根拠となる法律についてまとめる。その次に、著作物の例外として記述されている著作権法 10 条についてまとめる。

3.1 著作権法第 12 条

図 2 で明記されているように、創造性をもって素材である報道を選択や配置した記事に対して、著作物として著作権が発生することが確認できる。

3.2 著作権法第 10 条

この法律（図 3）で、著作物となるものを具体的に挙げている。しかし、この 2 項に「事実の伝達にすぎない雑報及び時事の報道」はそれらの著作物に該当しないと記載している。この項を根拠とすれば、著作権の発生しない文というものをニュース記事から得ることができる。次に、この 2 項に該当する具体例を挙げていく。図 4 は、2 項に該当する例として挙げられる

> 第十二条
> 編集物（データベースに該当するものを除く。以下同じ。）でその素材の選択又は配列によつて創作性を有するものは、著作物として保護する。

図 2　著作権法第 12 条

第十条
この法律にいう著作物を例示すると、おおむね次のとおりである。
一　小説、脚本、論文、講演その他の言語の著作物
二　音楽の著作物
三　舞踊又は無言劇の著作物
四　絵画、版画、彫刻その他の美術の著作物
五　建築の著作物
六　地図又は学術的な性質を有する図面、図表、模型その他の図形の著作物
七　映画の著作物
八　写真の著作物
九　プログラムの著作物
2　事実の伝達にすぎない雑報及び時事の報道は、前項第一号に掲げる著作物に該当しない。
3　第一項第九号に掲げる著作物に対するこの法律による保護は、その著作物を作成するために用いるプログラム言語、規約及び解法に及ばない。この場合において、これらの用語の意義は、次の各号に定めるところによる。
一　プログラム言語　プログラムを表現する手段としての文字その他の記号及びその体系をいう。
二　規約　特定のプログラムにおける前号のプログラム言語の用法についての特別の約束をいう。
三　解法　プログラムにおける電子計算機に対する指令の組合せの方法をいう。（二次的著作物）

図 3　著作権法第 10 条

安倍首相
【午前】９時３９分、皇居。春季皇霊祭・神殿祭の儀に出席。１１時２０分、東京・富ケ谷の自宅。
【午後】０時２３分、東京・新宿の「スタジオアルタ」。２９分、フジテレビのバラエティー番組「笑っていいとも！」に出演。司会のタモリさんとイチゴを試食。１時１１分、東京・永田町のザ・キャピトルホテル東急。中国料理店「星ケ岡」で秘書官と食事。４４分、公邸。４５分、加藤官房副長官、礒崎首相補佐官、谷内国家安全保障局長、杉山、長嶺両外務審議官、針原農水審議官、近藤駿介原子力委員会委員長ら。２時２５分、近藤氏出る。４７分、礒崎、針原両氏出る。３時４分、古沢財務官加わる。２８分、谷内、古沢両氏出る。３１分、加藤、杉山、長嶺各氏出る。５時３０分、マレーシアのナジブ首相と電話協議

図 4　首 相 動 静

首相動静 [9] の引用である。

首相動静に記載されている内容は、首相がいつ、どこで、なにをしたかという「事実の伝達にすぎない雑報及び時事の報道」であることが確認できる。

次の図 5 では、首相動静ではない経済記事 [10] を例にとり、著作権法第 10 条 2 項該当すると考えられる文がどこにあたるのかを見ていく。この記事において、著作権法第 10 条 2 項に該当すると考えられるのは、下線を引いた部分である。よって、本研究では、下線部を自動で抽出できることを目指す。

7 日の東京株式市場で日経平均株価は 4 営業日ぶりに大幅反発した。大引けは前日比 370 円 86 銭（1.60%）高の 2 万 3575 円 72 銭だった。前日の米株式相場が安く始まった後、上げに転じて終えたことを受け、中東情勢の緊迫化による世界的な株安進行への警戒感が和らぎ、押し目買いが先行した。その後、短期志向の海外投資家による株価指数先物への買いや売り方の買い戻しが断続的に入り、日経平均は 1 日を通じてじりじりと上げ幅を拡大した。

図 5　経 済 記 事

4　提 案 手 法

この章では、本研究の提案手法について述べる。また、提案手法の概要図を図 6 に示す。本手法は、データセットの収集、前処理・特徴量の設定、ラベル付けと学習を行い、モデルを作成する。その後、作成したモデルをテストデータに適用し著作権法第 10 条 2 項に該当すると推測される文を抽出する。

4.1　訓練データの収集

幅広い分野の記事に対応できるように、経済・金融分野、政治分野、ビジネス分野、マーケット分野、テクノロジー分野、国際分野、スポーツ分野などと各分野から 20 記事ずつ収集する。

4.2　前　処　理

集取した記事データを一文ごとに区切り保存した後に機械

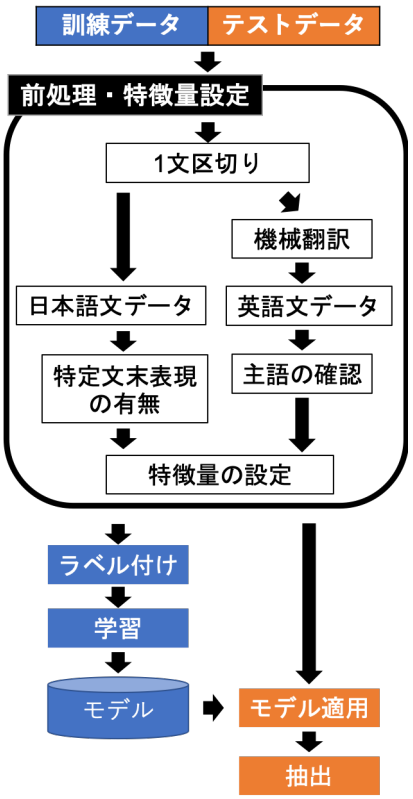

図 6　提案手法の概要図

翻訳を行い、記事データの英語訳を取得する。英語訳データを取得するのは、日本語に比べて、時制や態などの文法規則から文の特徴を取得しやすいためである。引用を意味する括弧・クオーテーション内部の内容は、著作権法第 10 条 2 項に該当するか影響を及ばさないため、日本語・英語両文の括弧または、クオーテーションとそれに囲まれた箇所を削除する。

次に、日本語文データで特定の文末表現の有無と、機械翻訳によって取得した英語文データでの主語が一人称かを記録する。特定の文末表現、主語が一人称が記録された際に、自動的に非該当文であると出力する。特定の文末表現は、「だろう」、「べき」、「必要」、「ほしい」、「たい」、「みられ」、「そうだ」、「ねらい」と設定する。文末表現として「～だろう」があれば、記事記述者の推測が文に含まれていたり、「～（する）必要」があれば記述者の意思が含まれるなど、以上で挙げた文末表現が含まれると記述者の意見文であると判断できるためである。また、英語文での主語については主語が一人称の場合、その主語は記事記述者自身となり、その文は同様に意見文と判断できるためである。

4.3 特徴量の設定

日本語文データ、英語文データに対して特徴量を以下のように設定する。

4.3.1 日本語文・英語文両方に設定

（1） 数字

この特徴量は、文中に出現する数字の個数とする。また、漢数字は数字として認識させていない。一文中における数字の個数を特徴量とした理由は、数字というのは客観性を示す要素の一つであるからである。著作権法第 10 条 2 項の「事実の伝達にすぎない雑報及び時事の報道」というのは、記事記述者の主観的な意見・考えなどが一切含まれていないことが条件の一つであり、その主観・客観性の要素となる部分である数字が、特徴量となると考えるためである。

（2） 単語数

日本語文データは、形態素解析器を用いて一文を形態素ごとに分け、その形態素の数とする。英語文では、空白で単語を分けた。単語数に関して、一文が長くなれば、その分記事記述者による内容の創造性が含まれる可能性が増し、著作権法第 10 条 2 項の「事実の伝達にすぎない雑報及び時事の報道」から離れていってしまうと考えたため特徴量の一つとした。また、3 章で著作権法第 10 条 2 項の具体例として挙げた首相動静からわかるように、一文あたりの単語数が少ないという傾向も確認できる。

4.3.2 日本語文のみに設定

（1） 述語の助動詞

この特徴量は、一文の述語に付随している助動詞とする。著作権法第 10 条 2 項の「事実の伝達にすぎない雑報及び時事の報道」にある「事実」というものは、過去に起きた事象と捉えることができるため、文の時制が過去であることが、著作権法第 10 条 2 項に影響を与えると考えた。そこで、日本語の時制を司ることが多い述語の助動詞に注目し、特徴量の一つとする。

（2） 形容詞・形容動詞

形態素解析器を用い各文中の単語の品詞を調べ、形容詞もしくは形容動詞の出現を特徴量とする。形容詞や形容動詞は、名詞などを修飾するという機能があるが、その修飾は主観的な判断が含まれているため、形容詞・形容動詞の出現が該当文の抽出に影響を与えると考え、特徴量の一つとする。

4.3.3 英語文のみに設定

（1） 時制

過去形、現在形、未来形の 3 つに時制を分けて、その時制を特徴量とする。形態素解析器を使用することで文中の動詞の時制を得ることができるため、そこから文の時制を特定する。時制を設定した理由は、4.3.2 の述語の助動詞の設定した際と同様である。

（2） 形容詞

形態素解析器を用い各文中の単語の品詞を調べ、形容詞の出現を特徴量とする。特徴量とした理由は、4.3.2 の形容詞・形容動詞の設定した際と同様である。

（3） 態（能動・受動）

英語の文は基本的に、動作を行う者が主語になる能動態の文と、行為をうける対象が主語になる受動態の文に二分でき、どちらの態になるのかを特徴量とする。受動態であれば動作主の情報を省略でき、能動態の文に比べ客観性が高いとされていることを利用し、各文の態に応じて客観性を評価できると考えるため、特徴量の一つとする。

（4） 助動詞

文中に出現する助動詞を特徴量とする。英語の助動詞は、動詞に動作主（主語）の「意志」や「判断」を添えるために用いられるため、態同様、客観性を評価できると考え、特徴量の一つとする。

4.4 ラベル付けと学習

各文が著作権の発生しない根拠となる法律である著作権法第 10 条の 2 項に該当する文かのラベルを人手でつける。本研究では以下の 3 つの条件を満たす文を該当文とし、それ以外を非該当文と定義しラベルを付ける。

（1） 過去の事実の伝達文や時事の報道
（2） 記事記述者の意見・考えを含まない
（3） 記事記述者による解説・分析を含まない

各文に対して特徴量として設定した値とラベルを訓練データと

して、サポートベクトルマシン（以下、SVM）を用いて学習を行い、学習モデルを作成する。

4.5 抽　　出

テストデータとしてある記事を用意し、訓練データ同様に前処理と特徴量の設定を行う。前処理で、日本語文データでの特定の文末表現、または英語文データでの一人称の主語が記録された文であれば、抽出器に入力せず、そのまま非該当文として出力する。それら項目の記録がないそれ以外の文は、学習によって得たモデルをテストデータに適用させることで、著作権法第 10 条 2 項に該当すると推測できる文を抽出する。

5 実　　験

本章では、提案手法を用いて実験について述べる。

5.1 データセット

本研究では、データセットを日本経済新聞より過去の社説・政治・スポーツ・マーケット・ビジネス・経済分野より各 20 記事、国際分野を 70 記事収集した。該当文と非該当文の件数の差を少なくするために、該当文を多く含む国際分野の記事を多めに収集した。収集した記事を一文区切りにしたものに対して人手で 10 条 2 項に該当するかのラベル付を行った。その結果が表 1 である。

表 1　該当文・非該当文数

全文合計	2850 文
該当文	1100 文
非該当文	1749 文

前処理の機械翻訳は、Google 翻訳を使用して英語訳データを取得した。

次に、4.3 で述べたように特徴量を設定について述べていく。日本語文の単語数を求める際は、形態素解析器の MeCab[3]を用いることで形態素（単語）ごとに区切り、その単語の合計とした。日本語文の述語の助動詞の取得には、構文解析（係り受け解析）器である CaboCha[4]を用いることで、一文を文節ごとに区切り、図 7 のように文の最後の文節を述語としてその文節中に出現する助動詞を特徴量とした。また、英語の形態素解析には、Stanford CoreNLP [11] を使用して、動詞の時制と文の態を取得した。

精度検証のため、全てのデータを学習させず、訓練用データとテストデータに分割して行う。分割には、k-分割交差検証法を実装している scikit-learn[5]のライブラリを用いて行った。学習には、全データを 5 分割したうちの 4 分を学習データとして、同様に SVM を実装している scikit-learn のライブラリを使用し学習させることでモデルの作成を行った。また、分割した際の各回の訓練・テストデータの該当文・非該当文の存在比に差が生じないように、調整を行なった。

3 : http://taku910.github.io/mecab/
4 : https://taku910.github.io/cabocha/
5 : https://scikit-learn.org/

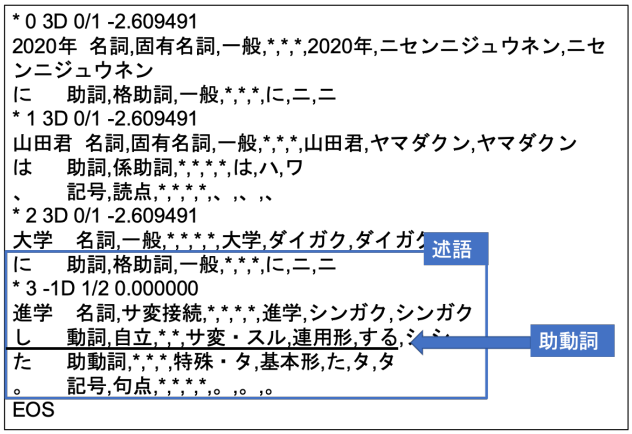

図 7　構文解析の結果

5.2 検証と結果

検証には学習で使用しなかった残りのデータをテストデータとして使用し、使用する特徴量を、(1) 日本語文データのみ (2) 英語文データのみ (3) 日本語文・英語文データ両方の 3 つに分けて、精度を検証した。学習で作成したモデルにテストデータを与えることで、著作権法第 10 条 2 項に該当すると推測できる文の抽出を行い、その結果を表 2〜4 に示す。精度には以下の 2 つの確率をもって検証した。

- 適合率：著作権法第 10 条 2 項に該当文として抽出したとき、実際に著作権法第 10 条 2 項に該当すると推測した文であった確率
- 再現率：著作権法第 10 条 2 項に該当すると推測した文の中で、著作権法第 10 条 2 項に該当文として抽出できた確率

表 2　日本語文データ結果（適合率と再現率）

回目	適合率	再現率
1 回目	0.857	0.736
2 回目	0.782	0.768
3 回目	0.756	0.777
4 回目	0.871	0.804
5 回目	0.873	0.722
平均	0.828	0.761

表 3　英語文データ結果（適合率と再現率）

回目	適合率	再現率
1 回目	0.678	0.586
2 回目	0.683	0.559
3 回目	0.556	0.518
4 回目	0.625	0.509
5 回目	0.756	0.495
平均	0.660	0.534

また、日本語・英語両文データを用いて作成したモデルに、実際にある一つの記事 [12]（図 8）を入力し、該当文の抽出を行なった例を図 9 に示す。

表 4　日本語・英語両文データ結果（適合率と再現率）

回目	適合率	再現率
1 回目	0.860	0.731
2 回目	0.778	0.734
3 回目	0.777	0.763
4 回目	0.886	0.777
5 回目	0.867	0.686
平均	0.834	0.739

経済のデジタル化に対応した国際課税の新ルールづくりが大詰めを迎えている。
世界約140カ国・地域が経済協力開発機構（OECD）の会合で課税の枠組み案で大筋合意したが、米国が事実上の骨抜きといえる新たな案を主張するなど先行きは混沌としてきた。
各国は事態打開に向けて一層の努力をしてほしい。
この取り組みは、巨額の利益をあげながら租税回避などの手法を駆使し税金をわずかしか納めない巨大IT（情報技術）企業への対策として20カ国・地域（G20）とOECDが連携して進めている。
OECDは2019年10月に、グローバル企業の一定水準以上の営業利益を基準に課税し、各国で税収を配分する事務局案を公表した。
1月末の会合で各国がこの案について大筋で合意した。
7月に全体会合を開き、年内の最終合意を目指すことになったが、難題を積み残している。
大きな障害になりそうなのが、昨年末に米国が公表した新たな提案だ。
同案では、新たな課税ルールをつくっても、そのルールに従うかどうかは企業に選ばせるというものだ。
これでは事実上の骨抜きになってしまうとして、米国以外の国から反発が強まっている。また、欧州が進める独自のデジタル課税をめぐる紛争の火種も残っている。
フランス、英国、イタリアなどは国際合意ができる前に各国独自のデジタル課税の導入に動いている。
19年に施行したフランスに対し、米国は「米企業への狙い撃ちだ」と反発し報復関税で脅しをかけた。
両国は1月下旬に、フランスが20年末までは米企業への税の徴収は見送る一方、米国も報復関税を控える「休戦」で合意した。
ただ、国際合意がそれまでにできなければフランスは米企業に課税を実施する方針で、貿易も絡めた紛争が広がる恐れは残る。
米国は英国など他の国に対しても、独自課税を実施すれば、関税などで報復する意向を示している。
年内の最終合意に向け、課税対象になる企業の線引き基準をどうするかも大きな論点になる。
デジタル課税とは別に、過度の税率引き下げ競争を防ぐために法人税に事実上の「最低税率」を導入する仕組みの検討も進んでいる。
経済のデジタル化という構造変化に、世界の国々がどう協力して対応するのか。21世紀の新たな国際協調の真価が問われている。

図 8　記事データ（抽出前）

OECDは2019年10月に、グローバル企業の一定水準以上の営業利益を基準に課税し、各国で税収を配分する事務局案を公表した
1月末の会合で各国がこの案について大筋で合意した
19年に施行したフランスに対し、米国は「米企業への狙い撃ちだ」と反発し報復関税で脅しをかけた
両国は1月下旬に、フランスが20年末までは米企業への税の徴収は見送る一方、米国も報復関税を控える「休戦」で合意した

図 9　抽 出 結 果

6 考　　察

実験では、日本語文データ、英語文データ、日本語・英語両文データの 3 つの精度を検証を行なった。日本語文データのみの実験では、適合率平均 82.8%、再現率 76.1%に達することができた。反対に、英語文データのみの実験では、適合率平均 66%、再現率 53.4%と日本語文のみの実験と比較して、適合率約 17%、再現率約 23%も差が生じた。日本語文・英語文両データを利用した実験は、適合率平均 83.4%、再現率 73.9%と、英語文データに比べて、適合率、再現率両者とも精度は向上し、日本語文データの適合率は向上したが、再現率は低下したという結果となった。

まず、最も低い精度を出した、英語文データのみでの実験についてだが、適合率が約 6 割、再現率が約 5 割など非常に低い精度となってしまったため、出力結果より各特徴量が及ぼす影響を測ることが不可能であった。また、英語文データを機械翻訳を用いて取得する際に、本来の日本語文とは異なる意味の文となってしまう誤差が初めより存在してしまうため、英語文データのみを使用するこの実験は高い精度を得ることができなかったと推測できる。

日本語文データを用いた実験において、誤った結果を出力してしまった原因を、この実験の出力から考察を行う

(1) 非該当文とラベル付けした文であったが、該当文として抽出してしまったケース

このケースの例を図 10 に示す。これらが、該当文であると出力されたのは、特徴量が不足していたことが考えられる。非該当文であるとラベル付けした理由は、ある事実から記事記述者が解析・調査による解説されていると判断できたためである。しかし、本研究で記述者による解説の文を非該当文であると判断するための特徴量として設定したのは、文の長さや時制である。これらのケースは、一文が長くもなく過去の時制の文であるため、その特徴量が機能しなかった。

(2) 該当文とラベル付けした文であったが、非該当文として抽出してしまったケース

このパターンの例を図 11 に示す。これらの文は、過去の事実の報道である。しかし、記述された文は体言止めで記述されていたり、特に過去を示す助動詞が含まれていないため、現在形の文でとして判断し非該当文であると出力してしまったと考えられる。

次に、日本語文のみの実験に比べ、日本語・英語両文データの実験の方が適合率が向上しているが、その向上についての考察を行う。日本語文データのみでの実験で誤った出力をしていたが、日本語・英語両文データの実験では、正しい結果を出力した例を図 12 に示す。これらが、正しい結果を出力した原因と考えられるのは、英語文での助動詞が記録されていることが推測でき、英語文での特徴量が機能していると考えられる。英語文のみでは、精度を得ることができなかったが、日本語文データと組み合わせることで、日本語文のみでの実験の精度より高い精度を得ることが可能となることがわかった。

- 今回の選挙には6月から続く抗議運動への賛否を問う住民投票の意味もあった
- 固定価格買い取り制度は再生エネ普及に貢献した
- これまでは通信事業者がサービスを考え、顧客に電話、メール、インターネット接続などのサービスを提供していた
- 銀行を介する既存の送金より格安のサービスを実現した
- 次世代医療基盤法が昨年5月に施行し、患者が拒否しなければ医療データを匿名にして活用できるようになった

図 10　(1) ケース

- 下位が約40%を占め、増加傾向だ
- JDIと3行は17年8月に同額の融資枠を期間1年で契約し、18年に延長
- 10年物地方債の過去10年の利回り平均は0.607%だが足元では0.087%にとどまる
- 終値は10%高の3310円
- 19年6～11月の純利益は前年同期比6.9倍の21億円

図 11　(2) ケース

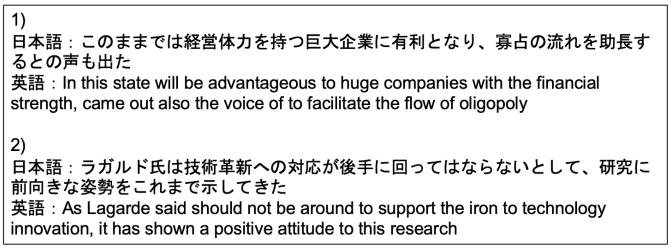

図 12　改　善　例

7　おわりに

本研究では、著作権法第 12 条より基本的には新聞記事には著作物としての著作権が発生してるが、著作権法第 10 条 2 項より著作物としての例外が存在しているという考えの基に、著作権法第 10 条 2 項に該当されると推測する文の抽出を機械学習を用いて試みた。実験の検証結果として、適合率の平均が 83.4%、再現率の平均が 73.9%となった。実験データを、原文である日本語文データに加えて、機械翻訳によって得られる英語文データを使用することで、適合率を向上させることができることがわかった。

しかし、機械学習で作成されたモデルに記事データを入力すると、著作権法第 10 条 2 項に該当すると機械が判断された文のうち、平均で 17%が著作権法第 10 条 2 項に該当しないと推測する文であった。つまり著作権が発生しないと機械が判断した文のうち著作権が発生すると推測する文が 17%存在したことになる。これは、法律が根拠である故に現状の精度では十分と言えないため、今後の課題としてさらなる精度向上に向けて研究を行う必要がある。

今後の課題として、英語文データでの精度向上が挙げられる。英語文データでの精度向上が達成されれば、日本語・英語両文データでの精度も応じて向上することが予測できるためである。具体的には、機械翻訳の英語訳に応じた特徴量の設定などが考えられる。

文　　献

[1] 総務省,『情報白書』, 平成 30 年版, p.3
[2] みずほ銀・ソフトバンク「情報銀行」個人データ仲介, 日本経済新聞, 2019-12-25, 朝刊, p..9
[3] 就活生情報　説明なく提供, 日本経済新聞, 2019-08-02, 朝刊, p..1
[4] 東京高裁判　平 6・10・25　平成 5 (ネ)3528 号
[5] 日本新聞協会. 新聞著作権に関する日本新聞協会編集委員会の見解. 第 351 回編集委員会,1978-5-11.
[6] 嶋田康平, 岡田真, 橋本喜代太.SVM を用いた株価短報における意見文と事実文の抽出. 言語処理学会第 19 回年次大会,pp.15-17 ,2013.
[7] 川口敏広, 松井藤五郎, 大和田勇人.SVM と新聞記事を用いた Weblog からの意見文抽出. 人工知能学会全国大会論文集 第 20 回全国大会 (2006), 一般社団法人 人工知能学会.,pp.11–11 , 2006.
[8] 端大輝, 村田真樹, 徳久雅人. 感動を与える文の自動取得と分析. 言語処理学会第 18 回年次大会,pp. 303-306,2012.
[9] 首相動静　21 日, 朝日新聞, 2014-03-22, 朝刊, p..3
[10] 日経平均大引け　大幅反発、370 円高　米株上昇で短期筋が先物買い, 日本経済新聞, 2020-01-08, 日経速報ニュース.
[11] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.
[12] 難航するデジタル課税、打開に努力を, 日本経済新聞電子版, 2020-02-13.