ISSN 2186-5647

5-23

判例の自動要約のためのテキスト分類

日大生産工(院)　○山田 康太郎　　日大生産工　豊谷　純

1. まえがき

裁判の判例は具体的な訴訟に対する裁判所の判断が示されており、類似の事件や同じ法律上の問題点について、同趣旨の判決が繰り返されているなど、先例として一般性を持っている。そのため、判例を理解することは弁護士や当事者本人が裁判の予測を試みるに際してとても重要であるが、一般的に読みにくく理解しづらい文章となっている実情がある。

そこで本研究では、判例を読みやすく分かりやすくすることを目的とした自動要約をするために、自動要約に必要なテキストデータの分類の実験を行なう。

実験で使用する判例は、知的財産権に関する判例を扱い、テキストデータを重要文であるか非重要文であるかを識別するために、文章を形態素解析と tf-idf 手法により定量化し、考察を行なう。

2. 要素技術と提案手法

2-1：形態素解析について

形態素解析とは自然言語のテキストデータ(文)を辞書に登録されている単語の情報に基づき、言語として意味を持つ最小単位の形態素の列に分割する技術である。本研究では python3.0 にて形態素解析ライブラリの MeCab と辞書の一つである IPA 辞書を使用する。「私は東京に住んでいる」の文章のように日本語は英語などと比較すると単語がスペースなどで区切られていない。この文章に形態素解析を行うと「私/は/東京/に/住ん/で/いる」のように、形態素に分解されコンピュータを利用した分析がしやすくなる。

2-2：tf-idf 値について

続いて、tf-idf[2]について言及する。tf-idf 法は、重要文抽出手法として頻繁に使われており、対象文書にのみ頻出する語を重要とする特徴を持っている。tf（Term Frequency）は、「ある文書内における単語の出現頻度」を表し、idf（Inverse Document Frequency）は、「逆文書頻度」を表す。idf は、全文書数をある単語が出現する文書数を割った値の対数である。他の文書におけるある単語の出現回数が増えると重要度の値は下がる。idf は、一般語（どんな文書にも出てくるような単語）のフィルタとして機能し、出現回数のみの視点では埋もれる特定の単語の重要度を考慮することができる。

これら2つの指標を掛け合わせたtf-idfの値については、大きいほど重要で小さいほど重要ではないと判断できることになる。以下に計算式を示す。

$$tf_j^i = Napp \qquad (1)$$

$$idf_i = \log\left(\frac{Nall}{Nw}\right) + 1 \qquad (2)$$

$Napp$: 対象判決文jでの単語iの出現回数
$Nall$: 全判決数
Nw: 単語iが出現する判決文数

ある判決文内に出現する単語iのtf-idf値は$tf_j^i \times idf_i$となる。

3. 実験

実験には、裁判所のホームページより入手した、知的財産判例集の商標権に関する判決文を用いる。用いる判決文の対象を第一審に絞り、勝訴の判決文と敗訴の判決が同数になるようにサンプリングを行った。本実験では原告の請求が一つでも認められた場合を勝訴、原告の請求がいずれも認められなかった場合を敗訴とした。

次に現状行った実験と今後の実験について以下に示す。

Text classification for automatic summarization of case law

Kotaro YAMADA and Jun TOYOTANI

現状では対象文書をMecabによって形態素解析する。判決文中に出現する単語全てのtf-idf値を算出した。以上の結果の例は図1である。

文章＼単語	あらかじめ	あり	ある	いう	いえ	いえる	…
1．訴えの適	0	0	0	0	0	0	…
将来の給付	0.08237626	0	0.09809014	0	0	0	…
本件におい	0	0	0.10720316	0	0	0	…
しかしながら	0	0	0	0	0	0	…
また，本件	0	0	0	0	0	0	…
…	…	…	…	…	…	…	…
争点３（差止	0	0	0	0	0.20142502	0	…
そして，平	0	0.11415863	0.13731078	0	0.13185047	0	…
そして，被	0	0	0	0	0	0	…
なお，前記５	0	0	0.12925004	0	0	0	…
結論よって	0	0	0	0	0	0	…

図1 tf-idf値の一例

今後の実験では、決定木学習を用いて、対象のテキスト中の文を重要文/非重要文に分類することで重要文抽出を行う。学習にはあらかじめ重要文/非重要文の2クラスに分類済みの訓練データが必要なため、人手によりクラスの分類を行う。

評価尺度としては、以下に示す再現率(recall:R)，適合率(precision;P),F-measureを用いる[3]。

$$R = \frac{\text{決定木が抽出した正解重要分数}}{\text{正解重要文数}} \qquad (3)$$

$$P = \frac{\text{決定木が抽出した正解重要文}}{\text{決定木が抽出した重要分数}} \qquad (4)$$

$$F - measure = \frac{2 \times P \times R}{P+R} \qquad (5)$$

4. 今後の課題

本実験では判決文のテキストデータを形態素解析し、判決文中に出現する単語全てのtf-idf値を算出した。今回tf-idf値の算出対象は全てのため、数字などの文章の重要度にあまり関係しない単語も拾ってしまった。今後は算出対象を名詞、形容詞、動詞に絞るような処理が必要であると考える。

5. あとがき

小川らの研究では要約対象が法律であり、学習データには国が発行する官報に掲載されている法令のあらまし(国が作成した法律の要約文)を用いて重要文抽出を行なっている。しかし、知的財産の判決文では裁判所などの司法機関によって要約文が公表されていない。

よって、今後の実験に記したように重要文/非重要文の2つのクラスに分類させた学習データの作成は現状人手により行うしかない。それは、学習データの質が要約の質に直結すると考える。学習データの作成は私一人で行う予定であるが、質の向上などを考えると法律の専門家の意見も必要である。

参考文献

1) 小川泰弘, 佐藤充晃, 駒水孝裕, 外山勝彦, 法律の要約のためのランダムフォレストを用いた重要文抽出, 人工知能学会(2019),https://www.jstage.jst.go.jp/article/pjsai/JSAI2019/0/JSAI2019_4E2OS7a02/_pdf/-char/ja, 2021.10.12 閲覧.

2) 唯野良介, 嶋田和孝, 遠藤勉, アスペクトごとの文の重要度と類似性判断に基づく複数レビューの要約, 言語処理学会(2010)http://www.pluto.ai.kyutech.ac.jp/~shimada/paper/NLP2010Tadano.pdf, 2021.10.12 閲覧.

3) 奥山学, 原口良胤, 望月源, 決定木学習を用いたテキスト自動要約手法に関するいくつかの考察, 情報処理学会(2000) https://www.ipsj.or.jp/award/9faeag0000004emc-att/Y_Okumura_Manabu.pdf, 2021.10.12 閲覧.

4) 高村大也, 奥村学 監　言語処理のための機械学習入門, コロナ社, (2010)

5) 柳井孝介, 庄司美沙, Pythonで動かして学ぶ自然言語処理入門, 翔泳社, (2019)