

# 法律相談 SNS におけるユーザー投稿文書を用いた 著者役割推定

塩田 健人<sup>†1,a)</sup> 小町 守<sup>†1,b)</sup> 瀬戸口 光宏<sup>†2,c)</sup> 市橋 立<sup>†2,d)</sup>

概要：近年インターネット上に様々な QA サイトが存在しており，その中でも法律相談に特化したみんなの法律相談という QA サイトが注目を集めている．このサイトの特徴は，ユーザーが法律に抵触するような問題をする側（加害者）もしくはされる側（被害者）となり得る立場で相談文書を投稿するという点である．現在，サイト上には被害者と加害者の相談文書が混在してしまっていることにより，被害者が加害者の相談文書を，加害者が被害者の相談文書を不意に見てしまう可能性がある．しかし，法律相談文書を閲覧するユーザーとしては被害者は加害者の相談内容を，加害者は被害者の相談内容を見たくないと考えられるため，ある問題の被害者が書いた文書が加害者が書いた文書かで相談文書を分ける必要がある．そこで，本研究では，法律問題をする側，される側どちらの立場で法律相談文書が書かれているかを判別するタスクを提案する．そして，法律相談文書をアノテーションすることでデータセットを構築し，機械学習を用いて著者の立場を自動で分類することで，それぞれの立場を分類する上での特徴を分析した．

## 1. はじめに

近年，インターネット上に様々な QA サイトが存在する．QA サイトとは掲示板に質問者となるユーザーによって投稿された質問に対し，質問の内容に関して知識を持ったユーザーが答えるといった場を提供するサービスのことである．一般的な QA サイトでの質問の回答者はユーザーが投稿した質問内容に関して知識のある一般人であるが，医師や弁護士などの資格などを持った専門家が回答することに特化した QA サイトが存在する．例えば，ユーザーが抱える法律相談に対して，弁護士による専門的な回答を得ることのできる「みんなの法律相談」という QA サイトを展開している弁護士ドットコム<sup>\*1</sup>がある．みんなの法律相談に投稿される質問とは，ある法律問題で困っているユーザーが自分の置かれた状況，解決して欲しい問題などをまとめているものが多い．みんなの法律相談の特徴は，ユーザーが法律に抵触するような問題の被害者もしくは加害者となり得るという点である．

みんなの法律相談のユーザーは 3 種類に分けることがで

きる．過去に自分自身と似た状況の別ユーザーから投稿された相談文書をサイト上の QA 集合から探す“閲覧者”，自分自身に関しての質問をする“質問者”，そして質問に回答する“回答者”である．前述の通り，弁護士ドットコムは弁護士が質問に回答するサービスを提供しているので回答者は弁護士であり，閲覧者と質問者は一般ユーザーであれば誰でもなり得る．

ここで，現在のみんなの法律相談の問題点として，ある法律問題の被害者と思われる質問者と加害者と思われる質問者の両方の投稿が混在しており，閲覧者には両方表示されてしまうという問題がある．閲覧者の心理としては被害者は加害者の相談内容を，加害者は被害者の相談内容を見たくないであろうと考えられるが，現在はそのようなフィルタリングをすることができない．例として以下に養育費に関する法律相談を示す．

現在離婚調停中で、養育費について話し合っています。一般的には、20歳まで支払わなければなりません。その途中で定年退職を向かえ、年金ももらえない時期がある場合、貯金や退職金で暮らしていかなければいけません。そういう場合、養育費の支払いはどうなるのでしょうか。子供には、お金を支払いたいのですが、相手にお金を支払いたくないのです。現在、子供は2歳。定年の時には、子供は15歳です。取り決めのときに、何かいい方法があれば、教えてください。

<sup>†1</sup> 首都大学東京  
Tokyo Metropolitan University

<sup>†2</sup> 弁護士ドットコム株式会社  
Bengo4.com, Inc.

a) shioda-kent@ed.tmu.ac.jp

b) komachi@tmu.ac.jp

c) setoguchi@bengo4.com

d) ichihashi@bengo4.com

\*1 <https://www.bengo4.com>

この例では、質問者は養育費を支払う側であると読み取ることができる。文書中に“相手にはお金を支払いたくないのです。”と書かれていることから、養育費を支払われる側にとって不快となる内容であると考えられる。このように“する側(加害者)”, “される側(被害者)”が存在する法律問題の当事者らは、自分と反対の立場の人間が書いた意見が不快と捉えることが多々あるために、敏感に取り扱わなければいけないのである。

そこで、我々は質問者による法律相談文書から、どのような法律問題があるかを抽出し、質問者が相談した法律問題に対して“する側”の立場で相談をしているのか、または“される側”の立場で相談をしているのかを判別するタスクを提案する。自然言語処理の技術を用いてみんなの法律相談に投稿される相談文書を解析し、質問者がどのような人物か、どのような問題を抱えているのかを自動で構造化データにすることで、同じ相談を持つユーザーが過去ログから相談文書を検索する手間を容易にしたり、適切な回答者とのマッチングの精度を向上させることが可能である。

本研究での貢献を以下に示す。

- 質問者となるユーザーによって投稿された法律相談文書を解析し、それぞれの法律相談文書の中で法律問題に対し質問者がある事象をする側かされる側かという役割を判別する新しいタスクを提案した。
- みんなの法律相談に投稿されている法律相談文書の中に現れる各法律問題について、質問者が法律問題の事象をする側かされる側かであるかというアノテーションをし、質問者の立場に関するデータセットを作成した。
- キーワードマッチを用いて法律問題同定を行い、サブカテゴリごとに特徴を分析した。また、教師あり学習を用いて法律相談文書中で質問者が法律問題をする側かされる側かを判別する分類器を構築し、分類の際に特徴となる表現を分析した。

## 2. 関連研究

文書に現れる特徴を解析し、書き手に関して年齢や性別、国籍などの情報を推定する著者属性推定に関する研究は国際会議やワークショップの共通タスク [4] として多くの研究者によって取り組まれているタスクである。

Peersman ら [3] は SNS 上に投稿されたショートテキストを解析することにより、テキストの書き手の年齢、性別を推定するタスクに取り組んだ。Peersman らはテキストの書き手の年齢を推定する際に単語や顔文字の使い方、つまり単純な単語 unigram 素性が bigram や trigram よりも重要であることを示している。

Mukherjee ら [2] はブログに投稿されたテキストを解析することにより、テキストの作成者の性別を推定するタスクに取り組んだ。Mukherjee らは性別によって文書を作成

表 1 法律相談データのカテゴリと法律相談文書数

カテゴリ	相談文書件数	全体に占める割合
離婚・男女問題	149,370	32.1%
犯罪・刑事事件	83,384	17.9%
労働	41,760	9.0%
借金・債務整理	35,153	7.5%
不動産・建築	26,167	5.6%
相続	22,730	4.9%
裁判・紛争手続き	21,722	4.7%
インターネット	20,568	4.4%
民事・その他	18,254	3.9%
交通事故	17,551	3.8%
企業法務	10,349	2.2%
消費者被害	7,706	1.7%
債権回収	4,439	1.0%
医療	3,108	0.7%
国際・外国人問題	1,288	0.3%
税務訴訟	1,156	0.2%
行政事件	1,106	0.2%
total	465,805	

表 2 離婚・男女問題の主なサブカテゴリ

離婚, DV, 財産分与, 調停, 裁判, 婚姻費用, 不倫, 慰謝料, 別居, 養育費
--

する際に特徴となる品詞列のパターンを見つけるアルゴリズムと素性選択のアルゴリズムを提案した。それらを組み合わせることにより、従来の単純な品詞、内容語、機能語を素性にしたモデルと比較して大きく差をつけて優れていることを示した。

また、Burger ら [1] は Twitter のユーザーデータを使用して性別推定に取り組んでいる。Twitter 上での名前、プロフィール、Tweet などのテキストを大文字小文字の区別なしに文字と単語単位の n-gram を素性として使用して Twitter に投稿されたテキストから性別を推定した。Tweet だけを使用した場合と比較して優れた結果となり、性別を推定したい場合 Tweet 以外のテキスト情報を使用の方が良いことを示した。また、このモデルは言語特有の形態素解析などの前処理を行わなかったことから、言語に依存しない手法であることが示された。

上記のように文書から年齢や性別など、著者に関して固有な特徴を推定するタスクは多く取り組まれているが、著者が文書中で言及しているものに対してどのような役割、立場であるのかという研究はされていない。文書中での著者の役割が推定することができれば、著者の置かれている状況に対して著者にパーソナライズされた情報を提供するなどといった貢献ができる。

表 3 離婚・男女問題における法律相談文書の例．検索に用いたキーワードを下線で示した．

例 1	婚姻費用 分担請求の申立てをしたら、夫が <u>離婚 調停</u> を申立てました。別居 して1年ですが、 <u>離婚 調停</u> の申立書の <u>別居</u> 時期を見ると、 <u>別居</u> して3年の計算になってしまいます。夫の弁護士が作成していますが、 <u>別居</u> 日が明らかに間違っている場合、 <u>調停</u> 時に、 <u>調停</u> 委員の方に伝えた方がいいのでしょうか？ <u>離婚 調停</u> では <u>別居</u> 日は、気にしなくていいのでしょうか？
例 2	姉は <u>離婚 調停</u> 中です。姉は心療内科に行くほど、精神的暴力を夫から受けていました。次回の調停で、精神的 <u>DV</u> について主張したいと考えています。「精神的暴力で、傷害罪の刑事告訴を検討しています」というような内容を入れても問題ないのでしょうか？
例 3	<u>離婚</u> 届、公正証書、住民票移動等の役所手続き、どの順序で進めるのがいいのでしょうか？
例 4	<u>離婚</u> を切り出したら逆にでっあげ <u>DV</u> で訴えられそうになってます。この場合男性のほうには証明のすべもなく不利なののでしょうか？

表 4 各法律問題の書き手の役割

法律問題名	取り得る立場（加害者 / 被害者）
離婚	請求する/される
慰謝料	請求する/される
調停	申し立てる/申し立てられる
別居	する/される
裁判	申し立てる/申し立てられる
不倫	する/される
養育費	請求する/される
婚姻費用	請求する/される
財産分与	請求する/される
DV	する/される

表 5 離婚・男女問題における各サブカテゴリの出現回数

サブカテゴリ	単語出現文書数	割合	累計文書数
離婚	96,101	59.1%	96,101
慰謝料	48,201	29.7%	111,970
調停	47,478	29.2%	121,621
別居	35,751	22.0%	126,763
裁判	35,530	21.9%	132,582
不倫	31,872	19.6%	136,934
養育費	28,017	17.2%	140,113
婚姻費用	11,095	6.8%	140,733
財産分与	9,611	5.9%	141,268
DV	8,635	5.3%	143,148
全相談文書数	162,456		
全法律問題数	352,291		

### 3. 法律相談文書に対する著者役割のアノテーション

#### 3.1 データ

みんなの法律相談に投稿される法律相談文書は、それぞれ累計で2014年3月時点で24万件、2015年3月時点で33万件、2016年3月時点で43万件、2017年3月時点で53万件と年々増加している。みんなの法律相談では無数に存在する法律問題を大きく17つにカテゴリ化しており、質問者は自身が抱える問題に当てはまるようなカテゴリのタグを付けてサイトに投稿する。2016年8月時点で投稿されている法律相談の各カテゴリと全体に占める割合を表1に示す。本研究では全相談データの中で最も大きな割合を占める「離婚・男女問題」カテゴリについて扱っていく。

離婚・男女問題において質問者から質問されるような法律問題は無数に存在するが、今回は離婚・男女問題における主なサブカテゴリである10種類のイベントに限定することにし、それらの法律問題についてまとめたものを表2に示す。離婚・男女問題カテゴリに今回取り扱うサブカテゴリの法律問題がどの程度含まれているのかを調査するために、162,456件（2017年4月時点）ある相談文書の中から、表2に示した法律問題を表す単語を含む相談文書を検索した結果を表5に示した。表5にある法律問題を含む法律相談数は異なりで143,148件であり、離婚・男女問題の法律問題において約88%カバーする量となった。

質問者によってみんなの法律相談に投稿される各相談文

書の特徴は、1つの相談文書の中に質問者が抱えている法律問題が1つ以上含まれている点である。相談文書の具体例を表3に示す。ここで、表3の例1を見ると、この例の質問者は「婚姻費用」、「離婚」、「別居」、「調停」の4つの法律問題を抱えていることになる。そのため、この質問者が「婚姻費用」\*2の問題では費用を請求する側なのか、される側なのか、「離婚」の問題では質問者が離婚請求をする側なのか、される側なのか、「別居」に関しては質問者が別居を申し込む側か、申し込まれる側なのか、そして「調停」に関しても質問者が申し込む側なのか、申し込まれる側なのかなどという各法律問題に関して質問者の役割は何かを推定する必要がある。また、表2の各法律問題における書き手の役割について表4にまとめた。

#### 3.2 アノテーション仕様

本研究では法律相談 SNS 文書に対する著者役割付与とタスクを法律問題同定と著者立場推定の2つのサブタスクに分解し、アノテーションを2STEPに分けて行った。

STEP1（法律問題同定）として、相談文書に出現する表2に示した法律問題に関し、書き手が法律問題を含む特定の事象に直接関わっており、法律問題をやる側もしくはさ

\*2 離婚問題で取り上げられる婚姻費用とは、別居中の夫婦間における夫婦や子の生活費などの婚姻生活を維持するために必要な一切の費用のこと

れる側の立場になり得るのかの表す法律問題タグを付与することで判定した。表2で示した法律問題の単語が含まれている法律相談文書をアノテーション STEP1の対象とした。以下に4つの例を示す。表3の例1に関しては、含まれている4つの法律問題全てが書き手にとって特定の事象を指しており、する側・される側の立場をとることのできる問題であると読み取ることができる。一方、例2に関しては“姉”に関する質問であり、特定の事象を指しているが、相談文書中に含まれている全ての法律問題が書き手本人についてのことではないと読み取ることができる。例3のような一般論的な質問は総称的な事象について言及しており、質問者に関する問題であるとは判断できないので、これらの法律問題は除いた。また、例4の“DV”問題のような事実ではない事象について述べられているものも除いた。

STEP2(著者立場推定)として、STEP1の作業で残った相談文書に対して、文書内に含まれる各法律問題について書き手がどのような役割をしているのか表4に示したする側、される側を表す2つの立場と不明の3値のラベルでアノテーションした。さらに、STEP1の判定を確認した。以下に各例を示す。

#### STEP1(法律問題同定)

婚姻費用 分担請求の申立てをしたら、夫が離婚 調停を申立てました。別居して1年ですが、離婚 調停の申立書の別居時期を見ると、別居して3年の計算になってしまいます。夫の弁護士が作成していますが、別居日が明らかに間違っている場合、調停時に、調停委員の方に伝えた方がいいのでしょうか？離婚調停では別居日は、気にしなくていいのでしょうか？おおごとにしたくはないので、裁判や慰謝料請求とかはしたくはないのですが・・・

- 法律問題タグ：婚姻費用，離婚，調停，別居

婚姻費用，離婚，調停，別居に関しては書き手が関わっていると読み取ることができるので、法律問題タグをつける。しかし、裁判や慰謝料請求はしたくないと書かれており、現実の事象を指していないと考えられるので裁判タグ，慰謝料タグはつけない。

#### STEP2(著者立場推定)

婚姻費用 分担請求の申立てをしたら、夫が離婚 調停を申立てました。別居して1年ですが、離婚調停の申立書の別居時期を見ると、別居して3年の計算になってしまいます。夫の弁護士が作成していますが、別居日が明らかに間違っている場合、調停時に、調停委員の方に伝えた方がいいのでしょうか？離婚調停では別居日は、気にしなくていいのでしょうか？おおごとにしたくはないので、裁判や慰謝料請求とかはしたくはないのですが・・・

- 婚姻費用：する側
- 離婚：される側
- 調停：される側
- 別居：不明

STEP1で付与された各法律問題について、加害者・被害者・不明といったラベルをつける。婚姻費用は著者が、離婚および調停は著者の夫が申し立てていると読み取れるが、別居に関しては著者と夫のどちらがしたのか不明であるため、不明ラベルをつける。

#### 3.3 アノテーション結果

今回データセットを作成する際にクラウドソーシングサービスを提供するLancers<sup>\*3</sup>を使用した。

アノテーションしたデータは、表5に示した離婚・男女問題カテゴリの法律相談文書で、表2の法律問題を表す単語を含む法律相談文書143,148件の中から、2,000件をランダムにサンプリングした。アノテーション対象のデータの詳細を結果とともに表6に示す。依頼したデータ数は2,000文書あるが、作業者に依頼をする際Lancersの仕様上、20文書ごとに小分けにして依頼をし、3人に評価してもらった。この際、2,000文書を特定の3人に依頼したのではなく、20文書ごとに3人で評価をしたので、作業者の異なりでの合計は60人であった。

作業者のアノテーション結果の評価内容については、3人で多数決を取った。作業者間のFleissの $\kappa$ 係数はSTEP1(法律問題同定)では0.608、STEP2(著者立場推定)では0.708であった。アノテーションした結果、各相談文書には0~7個、平均して2.3個の法律問題が存在した。

表6の有効な相談文書のカラムにSTEP1の条件を満たした文書数を示した。今回、2,000件中10件はSTEP1の条件を満たさなかったためSTEP2のアノテーションでは除外した。また、“不倫”と“財産分与”、“DV”に関してはSTEP1の評価前から評価後にかけて文書数が増加していた。これは、文書中に“不倫”、“財産分与”、“DV”という単語は出てこないが、別の単語や句でそれらの法律問題について表現している文書があったためである。以下に例を

\*3 <http://www.lancers.jp>

表 6 アノテーションデータに含まれる各法律問題と  
 STEP1 (法律問題同定), STEP2 (著者立場推定) の結果

法律問題名	キーワードにマッチした 相談文書数	法律問題が同定された 相談文書数	加害者	被害者	不明	判別不可
離婚	1,361	1,098	581	290	159	68
慰謝料	655	628	365	215	31	17
調停	640	570	342	158	40	30
別居	518	448	235	143	36	34
裁判	475	337	206	69	31	31
不倫	464	612	248	308	37	19
養育費	390	347	198	109	30	10
婚姻費用	153	148	93	41	9	5
財産分与	141	161	84	43	16	18
DV	123	182	21	140	17	4
総法律相談文書数	2,000	1,990				
総法律問題数	4,920	4,531	2,373	1,516	406	236

示す。

喧嘩の延長で主人から暴力を受けることが月に1回程度あるのですが、今回そのほかの理由もあって離婚を考えています。

この相談から書き手は家庭内暴力 (DV) を受けていることが読み取れる。

STEP2 に関しては、加害者と被害者の量に大きく偏りが出た結果となった。STEP1 を満たした 1,990 文書に含まれる 4,531 件の法律問題に対して STEP2 の評価を行った結果、1,795 文書、4,295 件の法律問題において 2 人以上の作業者の評価が一致した結果となった。236 件の法律問題に関して、作業者 3 人とも違う回答をした原因は、作業者によって各法律問題に関する捉え方の違い、また知識の差が影響していると考えられる。実際にこれらの問題を体験したことのある評価者、または法律に精通している評価者であれば今回のアノテーションに最適であるが、今回取り扱った法律問題の性質上、クラウドソーシングによる一般的な評価者には難易度の高いアノテーションであった可能性がある。

#### 4. 法律相談文書の著者役割推定実験

##### 4.1 実験設定

本研究では、法律相談文書における法律問題同定について、法律相談文書から単語一致でどの程度同定可能か調査した。法律問題同定実験では、表 2 で示した法律問題を表す単語を含む法律相談文書と STEP1 でアノテーションした結果を正解として比較し、Precision, Recall, F1 スコアで評価した。

また、法律相談文書における著者の立場推定を 3 値分類問題として扱い、教師あり学習を用いて自動推定する。線形分類器に線形カーネル SVM、非線形分類器として多

項式カーネル SVM を使用した。多項式カーネルの次元数は 2 次元とした。SVM のツールとして scikit learn (Ver. 0.18.2) を使用し、他クラス分類は one-vs-rest を適用した。教師あり学習の際に使用した素性は、頻度が 3 以上の単語 unigram から trigram までの Bag of N-grams を用いた。

unigram の例

離婚 を 切り出し たら 逆に でっちあげ  
 DV で 訴え られ ました 。

また、法律相談文書を単語分割の際に使用した単語分割器は MeCab<sup>\*4</sup> (Ver. 0.996)、MeCab の辞書は IPADIC (Ver. 2.7.0) を使用した。

著者立場推定実験に使用したデータは STEP2 のアノテーションの結果 1,795 件を使用し、表 6 に示した各法律問題ごとに分類器を学習させた。評価は学習させたデータに対して 10 分割交差検定を行い、Accuracy で評価をした。法律問題タグは STEP1 で付与された人手のタグを正解として用いた。

##### 4.2 実験結果：法律問題同定

評価実験を表 7 の左カラムに示す。慰謝料、離婚、調停、養育費、婚姻費用に関しては単語マッチのみで Precision, Recall とともに高い結果となった。離婚、慰謝料、調停に関しては、実験に用いたデータにおいてこれらの単語を含む相談文書数が多い。また、慰謝料、養育費、婚姻費用に関して、著者が法律相談文書を書く際にこれらの問題を表す同義表現を使用している件数が少なかった。

一方、裁判に関しては他の問題と比較した際に低い Precision, Recall なので、単語マッチだけでは裁判に関する法律問題を同定することは難しい結果となった。これは著者が質問をする時に自分の置かれた状況を詳細に描写する傾向が見受けられ、実際には著者が裁判に全く関係しない

<sup>\*4</sup> <http://taku910.github.io/mecab/>

表 7 法律相談文書の著者推定役割実験結果

評価尺度	法律問題同定タスク			著者立場推定タスク	
	Precision	Recall	F1	線形カーネル SVM Accuracy	多項式カーネル SVM Accuracy
離婚	0.802	0.994	0.887	<b>0.595</b>	0.564
慰謝料	0.921	0.960	0.940	<b>0.710</b>	0.597
調停	0.878	0.986	0.929	0.596	<b>0.633</b>
別居	0.761	0.879	0.816	<b>0.611</b>	0.568
裁判	0.528	0.745	0.618	0.634	<b>0.674</b>
不倫	0.927	0.703	0.799	<b>0.759</b>	0.519
養育費	0.882	0.991	0.934	<b>0.682</b>	0.587
婚姻費用	0.869	0.899	0.884	<b>0.742</b>	0.650
財産分与	0.844	0.739	0.788	0.489	<b>0.590</b>
DV	0.854	0.577	0.689	0.781	<b>0.787</b>

表 8 DV, 不倫, 財産分与を意味する表現の例

DV	家庭内暴力, 暴力を振るわれた, 手を出された, 蹴られた
財産分与	2人で買った家の場合, それを売ったお金を分けてもらえるのか?
不倫	不貞行為, 愛人がいる, 体の関係

表 9 “別居”を含むが著者に関係のない例

例 1	以前交際していた男性に別居中の奥さんがいることがわかったんですが離婚して一緒に来てくれると言っていたので待ってた矢先に・・・
-----	--

場合でも“裁判”, “裁判所”, “裁判官”, という単語を多用することによる例が多数あったことが原因として考えられる.

DV, 不倫, 財産分与については Recall が低い結果となった. その原因として表 7 の DV, 不倫, 財産分与の結果を見ると, “DV”, “不倫”, “財産分与” 以外の表現でそれらの法律問題を表している例が多く見受けられるこのことから, 単純な単語マッチだけではこれらの法律問題を同定できなかったことが考えられる. それらの例を表 8 に示した. また, 別居は Precision が一番低い結果となった. 表 9 に具体例を示したように, 投稿される法律相談文書に“別居”という単語が使われているが, 著者に別居問題が発生していない事例が多いからであったと考えられる.

#### 4.3 実験結果: 著者立場推定

線形カーネル SVM と多項式カーネル SVM での分類結果を表 7 の右カラムに示す. 本研究で取り扱った法律相談文書における著者の役割立場推定問題では, 多項式カーネル SVM より線形カーネル SVM の方が多くの法律問題に関して高い精度を示した.

多項式カーネル SVM の精度が低かった原因として, 本研究では各法律問題ごとに分類器を学習させたため, それぞれのデータ量が少なく, 組み合わせ素性が疎となり有効に機能しなかった可能性が考えられる. 婚姻費用, 財産分

表 10 誤分類した例

失敗例 1	今現在別居中なのですが主人に家賃, 携帯代, 次男の保育料を支払って頂いています. ですがいつかは払って頂けなくなるか不安です. 今後の為にも確保したいのですが今からでも婚姻費用分担の調停をしておいたほうがよいのでしょうか?
失敗例 2	主人から, 毎月, 婚姻費用, 10 万円を貰っています. 私も働いていたら, 主人からの婚姻費用が, 毎月, 減るのでしょうか? どれぐらいに私の月給を抑えれば, 主人からの婚姻費用の 10 万円を, 毎月, 離婚成立するまで, 貰えるのでしょうか?
失敗例 3	主人から離婚をしたいと言われていました. 理由は色々ありますが主人は, 1 人になり楽になりたいと言います.

与, DV に関しては各ラベルが 100 件に満たないのでこれらを分類するにはさらにデータ量を増やす必要がある.

次に線形カーネル SVM で分類した際に誤って分類した例を示す.

表 10 の失敗例 1, 失敗例 2 は, 著者が婚姻費用を請求される側であると誤分類した. これはそれぞれの文書内にある“支払って頂いて”, “貰っています”という表現に影響されている可能性がある. 著者が婚姻費用を請求される側の法律相談文書を調査してみたところ, それらの相談文書には払っている, 貰っているということを意味する表現が多く含まれていることが分かった.

また, 失敗例 2 では著者が離婚を請求する側と誤分類した例である. 法律相談文書の著者の配偶者(パートナー)が主語となり“離婚をしたい”と書いてある場合, 法律問題を表す単語に続き, “したい”と記述されている. これは離婚問題に限らず, 今回対象にした法律問題を分類する際に最も多い誤分類したパターンであった.

#### 5. 終わりに

我々は法律相談文書から特定の法律問題に対し, 法律問題をする側, される側となる著者の役割を推定する新しいタスクを提案した. QA サイトにおける質問を解析し, 質

問者はどのような質問をしているのか、質問者はどの立場で質問をしているのかを自動で判別できるようになる研究が盛んになれば、同じ相談を持つ質問者が過去ログから相談文書を検索する手間を容易にしたり、適切な回答者とのマッチングの精度を向上させることが期待できる。

本研究では、みんなの法律相談に投稿された法律相談文書をアノテーションすることにより、法律相談文書内で言及されている法律問題が著者に関係のある事象であるかをタグ付けし、著者に関係のある法律問題についてする側の立場で文書を書いているか、される側の立場で文書を書いているかをラベル付けした新しいデータセットを作成した。また、作成したデータセットを用いて著者が関わっている法律問題を同定する問題に取り組み、法律問題における著者の立場の推定をした。

今回取り扱った法律相談カテゴリは離婚・男女問題であるが、この加害者、被害者のデータセットを他のカテゴリに分野適用して、全ての法律相談に対して効果があるのか調査する必要がある。

#### 参考文献

- [1] Burger, J. D., Henderson, J., Kim, G. and Zarrella, G.: Discriminating gender on Twitter, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1301–1309 (2011).
- [2] Mukherjee, A. and Liu, B.: Improving gender classification of blog authors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 207–217 (2010).
- [3] Peersman, C., Daelemans, W. and Van Vaerenbergh, L.: Predicting age and gender in online social networks, *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, ACM, pp. 37–44 (2011).
- [4] Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M. and Stein, B.: Overview of the 4th author profiling task at PAN 2016: Cross-genre Evaluations, *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings*, Évora, Portugal, CLEF and CEUR-WS.org, CLEF and CEUR-WS.org (2016).