3I3-4

ソフトウェア使用許諾書を対象とした重要条項の抽出

Extraction of significant clauses from Software License Agreement

川嶋和希[*1]　野中尋史[*2], 小林暁雄[*1], 太田貴久[*1], 増山繁[*1]
Kazuki Kawashima　Hirofumi Nonaka, Akio Kobayashi, Takahisa Ohta, Shigeru Masuyama

[*1] 豊橋技術科学大学大学院工学研究科情報・知能工学専攻　[*2] 大分工業高等専門学校情報工学科
Toyohashi University of Technology Computer Science and Engineering #1
Oita National College of Technology Information Technology #2

When a software user reads its software license agreement, the user needs characteristic clauses that are difficult to infer its existence based on common sense. In this paper, we focus on characteristic clause's predicates and infer its complement words using machine learning. Then, we infer contents that are attribution of the right destination, prohibited and/or allowed matters, respectively, from the clause's predicates and the complement words. Finally, we compare the contents and previously collected documents' contents and extract the characteristic clauses to make a software user easier to read the software license agreement.

1. 背景

インターネット上でのソフトウェアの配布においては、クリックラップ契約という契約形態がしばしば用いられる。このクリックラップ契約では、ユーザはブラウザ上の『同意する』ボタンをクリックすることで画面上に併せて提示される契約内容に同意したとみなされ、ソフトウェアのダウンロードを提供元から許可される。

ソフトウェア使用許諾書にはユーザの権利や義務が書かれており、後のトラブルを避けるためにも、それらには目を通すべきである。しかしながら、ソフトウェアのダウンロードを急ぐユーザは契約内容をほとんど読まず『同意する』ボタンをクリックしてしまう。これは、契約内容は一般に文章量が多く閲読に時間を要し、読み手であるユーザの負担が大きいためである。そのため、使用許諾書を分析し、膨大な量の条項から特に重要度の高いものを抽出できれば、読み手の負担は軽減されると考えられる。

2. 目的

本研究ではソフトウェア使用許諾書を読むユーザの負担軽減を目的とし、使用許諾書からの重要条項の抽出を試みる。ここで、以下に示す条項を重要条項とする。

- ユーザの指定する検索クエリを含む条項
- 対象許諾書において意味的に特有な条項

ユーザが検索クエリとして指定した「個人情報」や「義務」などの義務や属性を含む条項を抽出することで、多種多様なユーザのニーズへの柔軟な対応が可能となる。また、ソフトウェア使用許諾書全般と比較し、対象とする使用許諾書において意味的に特有である条項を抽出することで「逆アセンブルの禁止」などのソフトウェア使用許諾書において一般的であり、また、推測可能な条項を読み飛ばすことが可能となり、また同時に、通話アプリにおけるアドレス帳データの収集等の当該アプリに特有な条項の把握が容易になる。

上記の重要条項の提示は契約内容の把握を容易にするため、この重要条項の抽出は契約の不履行による紛争発生の抑止に繋がると考えられる。

連絡先：川嶋和希　kawashima@la.cs.tut.ac.jp

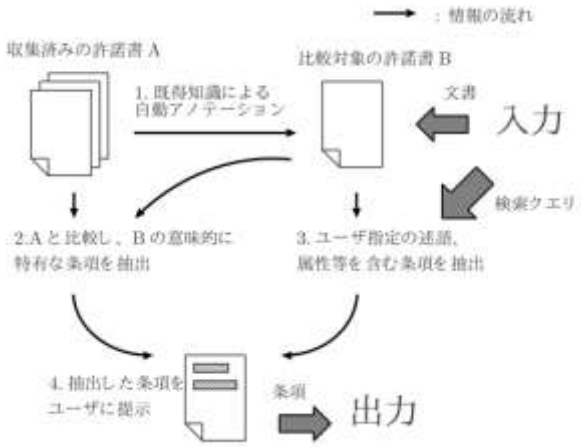

図 1　システム概念

3. 関連研究

島津は法令の制定や施行のための計算機支援としての法令工学の必要性を論じている[1]。法令文書における論理的矛盾の検出除去に加え、難解な法令の一般人にも読みやすい文章への変換を行っている。

法令文書のみに限らず、契約文書に対しても同様のアプローチが求められる。契約文書の分析支援としては、川端らによる英文契約文書の分析支援の研究[2]が挙げられる。川端らの手法では、契約文書作成の熟練者があらかじめ定めた重要語句を用いて、文書中の重要箇所を同定している。

同じく契約文書を対象としたものとして、待井らの契約文書を対象とした評価支援システムの開発が挙げられる[3]。しかしながら、文書中の要注意箇所を契約ナレッジに基づいて同定しており、読み手の盲点を突く条項には対応できない問題がある。

契約文書を対象とした研究そのものが少ない現状に加え、川端らの手法も待井らの手法も分析対象が企業間の英文契約文書に留まっている。

対象文書を限定し、その分野における常識を利用した意味解析としては向仲らの研究が挙げられる[4]。しかしながら、対象文書は技術文書に留まっており、日本語の契約文書を対象とはしていない。

畑山らは重要語句を用いた要約文の自動生成を行っている[5]。畑山らの手法では、格フレームを用いて重要語句を自動抽

出しているが、同様の意味をもつ語句の汎化は行っておらず、文の意味の特有さを分析する上で語句の抽出に同手法を用いるのは不適切である。

中原らは単文の意味を格フレームで表現し、その集合としての文の意味解析を行っている[6]。

柴田らも格フレームを用いて事態間知識を自動で獲得している[7]。例えば、「財布を拾って警察に届ける」という文は事態「財布を拾う」と「警察に届ける」のペアから成っている。この「拾う」と「届ける」は述語「拾う」のヲ格が「財布」で、述語「届ける」のニ格が「警察」のときにそれらの述語が共起しやすく、主にその共起情報と「ので」「ために」といった手掛かり表現から事態間の知識を獲得している。

しかしながら、中原らと柴田らの手法は事態間知識の獲得に留まっており、その知識の特有さの解析は行っていない。

4. 提案手法

本研究における条項抽出の概念を図 1 に示す。この既得知識のアノテーションのため、本研究では条項抽出の前処理として、条項内の述語に対し格フレームを付与し、インスタンスと述語、許諾書そのものに対し属性を付与する。

4.1 前処理

(1) 格フレーム付与

この前処理では、まず条項の述語に格フレームを付与する。そして、ソフトウェア使用許諾書の文中から格フレームの補語を抽出し、その補完された格フレームを条項の意味として扱う。以下はあるソフトウェア使用許諾書から抜き出した条項である。

> サトーは、本ソフトウェア製品を本契約書の条項内容を承諾した日本国内における居住者に限り、本ソフトウェア製品に対応するサトー製品の利用を条件として本ソフトウェア製品を使用する非独占的権利を許諾します。

例えば、この条項においては、述語「許諾」に格フレーム[主格,対格,与格]が付与され、文中の「サトー」「非独占的権利」「居住者」がその補語となる。そして、それぞれの格と補語の対応から、この格フレームは「サトー」が「居住者」に「非独占的権利」を「許諾」するという情報を保持する。ただし、「居住者」や「非独占的権利」は単体では情報が不足している。そのため、「居住」の与格としての「日本国内」の抽出や、「権利」に対する述語「使用」の対応づけ、さらに「使用」の格フレーム補完などを行い、それらのインスタンスに情報を付加する必要がある。

ここで述べた述語の種類は一般に膨大な数に及ぶが、本研究では対象文書をソフトウェア使用許諾書に限定しているため、後述のように、出現する述語の種類は限られる。そこで、頻出するそれらの述語と格フレームの対を保存するテーブルを人手で作成し、後の自動アノテーションに使用する。

(2) 属性付与

格フレームの補語は、「ユーザ」「弊社」「ソフトウェア」などのインスタンスである。しかしながら、同様の意味を持つインスタンスであっても、「ユーザ」「利用者」「お客さま」などの表記ゆれがソフトウェア使用許諾書には存在する。これらを全く異なる意味のインスタンスとして扱うと、同様の意味の条項を異なる意味の条項と扱ってしまい、その意味の特有さを判断する上で障害となる。そこで、表記ゆれの影響の軽減を目的としてインスタンスに対し属性を付与する。例えば、「ユーザ」「利用者」「お客さま」であれば属性【利用者】、「本ソフトウェア」「本製品」であれば属性【提供物】が付与される。この属性付与により表記ゆれの影響が軽減され、たとえば、「ユーザの権利」と「利用者の権利」のような同様の意味を持つ条項のまとめ上げが容易になる。ソフトウェア使用許諾書におけるインスタンスの契約上の位置づけを表すこれらを深層の属性として扱う。

深層の属性に加えて、ソフトウェア使用許諾書に限定せず、インスタンスがより一般的に属するカテゴリを表層の属性として扱う。例えば、インスタンス「ユーザ」「居住者」「第三者」に対し属性【人物】が、また「権利」「条件」に対し属性【抽象物】が付与される。そして、述語「使用」について、属性【製品】が属性【人物】を「使用」することをあり得ないと仮定すれば、先述の格フレームの補完時に、この表層の属性をフィルタとして用いることができ、格フレーム補完を自動化した際に、その精度向上に繋がる。

属性の付与対象はインスタンスに限らない。属性は述語と許諾書全体に対しても付与される。述語に対しては文中の手掛かり表現から属性【義務】【可能】【否定】などを付与し、格フレームと補語のみでは不足する情報を保持する。許諾書全体に対しては属性【有料ソフト】【個人向けソフト】【企業名】を付与し、対象許諾書をそれらに限定した統計処理を可能にする。

述語と格フレームの対と同様に、ソフトウェア使用許諾書において出現するインスタンスの種類は限定されると考えられる。そこで、これらのインスタンスと属性の対についてもテーブルを作成し、後の自動アノテーションに使用する。

4.2 手法

図 1 のシステム概要に示すように、本研究の条項抽出システムは、ユーザからの検索クエリとソフトウェア使用許諾書を入力とする。この検索クエリは、検索対象が単純な場合に用いられる文字列でもよいし、条項の述語の属性や、格とインスタンスの組合せであってもよい。さらに、ユーザが検索を必要としないならば、検索クエリは入力しなくともよい。ただし、属性やインスタンスといった本システム特有の要素を検索に用いる場合、それらを一覧から選択式にするなどのユーザの入力支援が必要になる。

入力された使用許諾書からは、文字列を抽出し条項に分割、そして、既に収集されたアノテーション済みの許諾書を用いて自動アノテーションを行う。この自動アノテーションでは、条項に対する係り受け解析結果と前処理の段階で作成した頻出語のテーブルを参照し、対象許諾書から述語とその格フレームを補完するインスタンスを抽出する。

自動アノテーション終了後、その文書における各述語の格フレームとその補語を収集済みの許諾書のものと比較し、意味的に特有な条項を抽出する。例えば、ソフトウェア使用許諾書において稀にしか出現しない述語とその格フレームの補語の組合せが対象許諾書に存在すれば、それは、意味的に特有な条項となる。また、事前に作成したテーブルに存在しない語句で構成された条項があれば、それも意味的に特有な条項となる。

検索クエリが入力された場合は、対象許諾書からそれらを含んだ条項を抽出する。単純な文字列検索でなく、属性や格とインスタンス、またその組合せで検索する場合は、例えば主格インスタンスが属性【利用者】、述語の属性【義務】と絞り込むことで利用者の義務を抽出することができる。この場合、主格を属性【利用者】に限定することで企業側の義務を排除した検索が可能となっている。

条項抽出の終了後、それらを元の対象許諾書上でハイライトする。抽出された条項以外を排除して提示しないのは、関連した内容の条項がその前後に存在しうるためである。

5. 頻出語の抽出

前処理、インスタンスへの属性の付与と、述語への格フレームの付与を行う前に、まず、インスタンスと述語をそれぞれ抽出する。そこで、文書の収集とテキストの抽出、ファイルの XML 形式への統一を行い、それに含まれる条項から付与対象となるインスタンスと述語の候補を抽出した。

文書収集においては、インターネット上から、HTML 形式と PDF 形式のソフトウェア使用許諾書を合わせて約 600 収集した。そして、不要な HTML タグや PDF ファイルにおけるページ番号等のノイズを除去した本文のみを抽出し、それらの形式を XML に統一した。

それらの XML ファイルの条項データに対し MeCab を用いた形態素解析を行い、述語とインスタンスの候補となる名詞を抽出した(表1)。述語として名詞を抽出するのは、ソフトウェア使用許諾書においては、「使う」を「使用する」と表現するように、述語が名詞の形で現れるためである。

表 1 抽出した頻出語の例(名詞)

述語	使用、契約、許諾、利用、提供、同意、複製
インスタンス	ソフトウェア、お客さま、製品、弊社、サービス

同時に、約 600 のソフトウェア使用許諾書において出現数が上位 95%を占める名詞は 1739 であり、上位 68%については 215 という結果が得られた。

この抽出により、ソフトウェア使用許諾書においてはごく一部の種類の述語、インスタンスが偏って出現することが確認できた。

6. まとめ

ソフトウェア使用許諾書からの条項の抽出基準となる重要度を得るため、格フレームとその補語を用いた条項の意味の推定手法を考案した。そして、ソフトウェア使用許諾書の読み手の負担軽減のためには、その格フレームと補語のアノテーションを自動化する必要があり、その前準備としての手動アノテーションの妥当性を頻出語の種類数の面から検証した。

検証の結果、ソフトウェア使用許諾書における頻出語の種類数は数百から千数百に収まり、手動アノテーションは可能という結論が得られた。今後は、述語とインスタンスの分類と、その手動アノテーション、述語の格フレームテーブルの作成を行い、その結果を用いた自動アノテーションの精度について検証を行う。

参考文献

[1] 島津明:法令文書の言語解析, 電子情報通信学会技術研究報告. 言語理解とコミュニケーション, 2010 年.
[2] 川端薫 他:英文契約文書の分析を支援するための取り組み, プロジェクトマネジメント学会 2009 年度 春季研究発表大会予稿集, 2009 年.
[3] 待井君吉 他:英文契約書評価システムの開発, 情報処理学会研究報告, 2010 年.
[4] 向仲コウ:技術文書の機械翻訳における常識と文脈情報の利用, 情報処理学会論文誌, 1990 年.
[5] 畑山満美子 他:重要語句抽出による新聞記事要約, 情報処理学会研究報告. 自然言語処理研究会報告, 2001 年.
[6] 中原大輔 他:意味構造抽出によるテキストマイニングと知識利用, 情報科学技術フォーラム一般講演論文集, 2002 年.
[7] 柴田知秀 黒橋禎夫:述語項構造の共起情報と格フレームを用いた事態間知識の自動獲得, 情報処理学会研究報告. 自然言語処理研究会報告, 2011 年.