自然言語処理 126－13
(1998. 7. 24)

英文契約書における要目の抽出

相良 かおる　　渡邉 勝正

奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県生駒市高山町 8916-5
TEL: 0743-72-5306　　FAX: 0743-72-5309
E-mail: {kaoru-s,watanabe}@is.aist-nara.ac.jp

本稿では、索引語を抽出する際に広く使われている TF.IDF 法を応用して、(1) 英文契約書の書式集に含まれる語または語群 (句、複合語など) の重要度、および、(2) 語間または、ある語と語群間の関連の強さを求める方法を提案する。重要度を用いることで、契約書の内容を表すのに必要な語句を抽出することができる。また、関連の強さを用いることで、詳細な文法規則や品詞情報を用いることなく、ある語または語群の前後に出現する語または語群を推測することが可能となる。(1) の重要度の有効性を評価する実験として、技術取引契約に関する書式集から重要語と重要度の一覧表を作成し、3 種類の契約書について、条項名の推定を行った結果、契約書中の条項名が一覧表に存在する場合、平均 90 %の適合率で条項名を推定することができた。

キーワード ：　情報抽出、TF.IDF 法、英文契約書、重要語、重要度、関連度

Extraction of important terms that reflect the contents of English contracts

Kaoru SAGARA and Katsumasa WATANABE

Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN
TEL: +81-743-72-5306　　FAX: +81-743-72-5309
E-mail: {kaoru-s,watanabe}@is.aist-nara.ac.jp

This paper describes a method for finding the importance degree of terms in English contracts by applying the TF.IDF method. By means of the importance degree of terms, extraction of important terms that reflect the contents of English contracts become possible. Next, an experiment to evaluate the effectiveness of the importance degree of terms is described. We made a list of important terms with the degrees from a collection of 35 kinds of articles of English contracts. Then we made an experiment to predict the article names of three kinds of contracts based on the list of important terms. As a result, the precision of 90 % was obtained. Finally, a method of estimating the correlation between terms is proposed.

keywords ：　Information extraction, TF.IDF method, a contract in English, important terms, importance degree of terms, estimation of the correlation between terms

1 はじめに

長文で複雑な英文契約書 (以下、契約書という) の内容を理解しやすい単文 (主語と述語がそれぞれ一つからなる文) の形で表記することが本研究の目標である。そのためには、契約書の内容を抽出する必要があるが、現在、長文で複雑な契約書の条文を正確に構文解析することは困難である。従って、情報抽出の方法として、テンプレートを使った手法が、構文解析による手法よりも、適切であると考えられる。

そこで、契約書の書式集について、(1) 契約書の持つ特徴、(2) 中学校で学ぶ程度の文法規則、(3) 語句の出現回数を使って、まず、その語句を基に単文の形で内容の抽出を行い、それらを基に情報抽出用のテンプレートを機械的に作成することを考える。

書式集から内容を抽出する際に必要な情報と、それらを求めるのに必要な要素として、以下のものが考えられる。

1. 主語、述語、目的語、補語などの文法関係
   要素：品詞情報、文法規則
2. ある動詞が名詞句等に与える意味的な役割
   要素：前置詞句
3. 抽出した内容の条文中における意味的な役割、重要性
   要素：接続詞、助動詞、語句の意味

品詞情報については、*Eric Brill's rule-based part of speech Tagger*[1] を用いることで、品詞辞書中に未定義語がない場合は約 95 %の正確さで、品詞を求めることができる。しかし、ある動詞のうしろに補語がくるか否か、すなわちある動詞が完全か、不完全かというような詳細な品詞の情報を知ることはできない。

前置詞句、接続詞、および助動詞については、一覧表を作成することが容易であり、また、一度作成したリストが増減する可能性は少ない。なお、本研究で用いる国際事業開発株式会社発行の技術取引に関する書式集 2315 条文に含まれる 114 種の前置詞句と 74 種の接続詞を既に抽出している [2]。

問題は、(1) 長文でかつ複合文である条文 (前述の書式集においては、製品名、住所、会社名などを一語と換算した場合の一文の平均語長は 35 語、最長 167 語である) を機械的に文法規則を用いて構文解析できないことと、(2) 抽出した内容の意味および重要性を知るための情報がないことである。

(1) については、「ある動詞と関連の深い名詞または形容詞」というように、語または語群間の関連の強さを求め、この関連の強さと部分的な文法規則を使って、条文の文法関係を推測することを試みる。

(2) については、語または語群の出現する場所 (条項) と出現回数から TF.IDF 法を応用して、局所性を求め、求めた値を重要度を示す指標として用いて、意味的な役割および重要性を推測することを試みる。図 1 に研究の概要を示す。

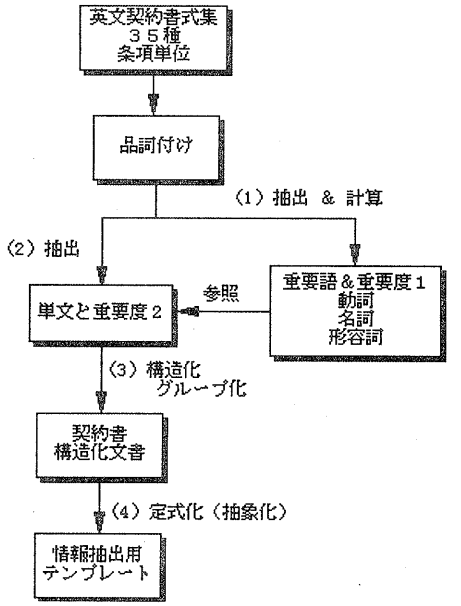

図 1: 研究の概要

本稿では、契約書の構造と内容、語句の重要度および語または語群間の関連の強さについての定義と、重要度を用いた契約書の条項の内容の推定に関する実験と評価について述べる。

2 契約書の特徴

契約書は、条項ごとに内容がまとめられており、大まかな構造化がなされている。条項は、(1) 契約の種類に関係なく、どの契約書にも通常含まれる一般条項と、(2) 契約の内容に依存する固有条項に大別される [3][4][5]。以下に、それぞれの特徴と推測されることを列挙する。

1. 一般条項
   - 文法上の構造や内容が、定式化している場合が多い。
   - 各条項間の関連が希薄である。
     - 未定義語の出現頻度は少ない。
     - その条項に出現する重要語の、他の条項での出現頻度は低い (局所性は高い)。

2. 固有条項

- 他の条項と関連のある条項がある。
- 契約内容に加え、文の構造に作成者の個性が反映される。
  - 文法規則を用いた内容の抽出は困難。
  - 内容の定式化が困難。
  - 未定義語の出現頻度が高い。
  - その条項に出現する重要語の局所性は低い。

契約書に記載されている内容は、以下の4つに大別される。

1. 権利または義務の記述
2. 権利または義務が生じる事由
3. 権利または義務が生じた際の措置
4. 契約内容の詳細

一般に、権利または義務を表す記述は、以下に示す語句 (helping verbs) によって、識別される。但し、"shall"、"may"、および "will" の直後に続く動詞が全て権利または義務を表すとは限らない。

1. shall
2. may
3. will
4. shall be deemed (to)
5. shall have (the/a) right to
6. be entitled to
7. reserve (the/a) right to
8. be permitted to
9. be obligated to
10. be liable to
11. be responsible (to/for)
12. be under the obligation to
13. have (the/an) obligation to

また、権利または義務が発生する条件についての記載は、以下の語句によって識別される。

1. if
2. unless
3. in the event (of/that)
4. in case (of)
5. subject to
6. when
7. upon
8. once
9. provided, however that
10. except (to/for/that)
11. except as provided (herein/hereinafter)
12. notwithstanding
13. but (not/that)
14. save (for/that)
15. in the absent of

したがって、前述の権利または義務を示唆する語句を含み、かつ条件を示唆する語句を含まない節が、権利または義務の記載となる。しかしながら、複合文である条文を機械的に節に分割することは困難である。そこで、本研究では、";"、":"、"("、")" および "." と、場合によっては、"that"、"if"、"which" で条文を分割したものを節として扱う。なお、"," は、"and" または "or" と共に名詞句を列挙する際に多用されることから、本研究では、節を分割する区切り文字として "," は、用いないこととした。

3 語句の重要度

前述のように契約書は、条項ごとに内容がまとめられている。すなわち、ある条項内で多く出現する語を重要語とみなすことができる。そこで、以下の項目のいずれかを満たす、動詞、名詞、および形容詞を重要語と定義する。

1. ある限られた条項に多く出現する。
   判断基準：出現する条項の数、出現回数
2. ある限られた条項にのみ出現する。
   判断基準：出現する条項

重要語の重要の程度を、索引語の自動抽出法として広く使われている TF.IDF 法を応用して数値化する。TF.IDF 法を以下に示す [6]。

$$w_j^i = tf_j^i \cdot log\frac{N}{df_j}$$

w_j^i ：文書 D_i において語 T_j を索引語として採用するかどうかを決定する指標
tf_j^i ： 文書 D_i における語 T_j の出現回数 (語頻度)
N ：文書群に含まれる文書数
df_j ：語 T_j を含む文書数 (文書頻度)

3.1 重要度の定義

ある条項 a_i における語 T_j の重要度 I_j^i は、条項 a_i における語 T_j の出現回数 (語頻度という) を tf_j^i 、条項における平均段落数 P、条項 a_i に含

まれる段落の数を P_i、契約書に含まれる条項の数を N、語 T_j を含む条項数(条項頻度という)を af_j としたとき、以下の式で定義する。

$$I_j^i = \sqrt{\frac{P}{P_i} \cdot tf_j^i \cdot log_2 \frac{N}{af_j}} \quad \cdots\cdots (1)$$

本研究で使用する書式集は、条項毎に例文が段落の形でまとめられている。そこで、段落数の平均 P を当該条文における段落数 P_iで割った比率を用いて、出現回数の正規化を行っている。したがって、$\frac{P}{P_i}$ は、重要度を求めるための必須項目ではない。また、重要度の値を条項の内容を推定する等に利用するために、平方根を用いて、重要度の大きな値間の差を抑えている。求めた重要度は、条項 a_i における語 T_j の頻度 tf_j^i が高く、かつ、その語の条項頻度 af_j が低い場合に、大きな値をとる。全ての条項に出現する語の値 I_j^i は、$af_i = N$ となるので、0となる。

3.2 実験と評価

今回提案する (1) 式の重要度が、契約書の内容を表す重要な語の指標として有効であることを示すために、35 条項における重要語と重要度の一覧表を作成し (556 語、最大値 25.16、最小値 1.86)、2 種類の技術取引契約書と 1 種類の代理店契約書について、重要度の累計を求め、その値から、条項名を推定する実験を行った。重要語の一覧にある条項と同じ内容の条項が契約書にある場合についての、3 種類の契約書における条項名の推定結果の適合率の平均は、約 90 %であった。

条項毎の重要語の一覧を求める手順は以下の通りである。

1. 前述の書式集35種類の条項に含まれる123,644語について、条項毎に原形(語から屈折形態素を除いた語幹)を求め、条項頻度と出現回数を求める。

2. (1) 式により重要度を求める。
(15,166 語、最大値 25.16、最小値 0)

3. 重要度＞ 4 で、かつ 条項頻度＜ 10 の語、および、語頻度＞ 1 で、かつ条項頻度＝ 1 の語を抽出し、重要度を値とする一覧表を作成する。(556 語、最大値 25.16、最小値 1.86)

4. 作成した一覧表を基に、実験対象の契約書について、条項毎に重要度の累計を求める。

表 1 に 23 の条項からなる、ノウハウライセンス契約書の条項の内容を推定した結果を示す。23 条項

表 1: 条項の推定

Know-How License Agreements.

No.	条項名	累計	比率
	WHEREAS clause		
1.	○ Definitions		
	Definition	22	1 / 11
2.	Grant of License		
	Duration	9	1 / 8
3.	Disclose of Know-how		
	Language	7	1 / 17
4.	Company's undertakings		
	Other Assistance	7	1 / 10
5.	Parts etc.		
	Payment	7	1 / 18
6.	× Technical Guidance		
	Technical Training	65	4 / 13
	Technical Guidance	22	3 / 46
7.	Dispatch of Trainees		
	Technical Training	138	5 / 13
8.	Technical License Fee		
	Tax,etc	6	1 / 12
9.	○ Payment		
	Payment	51	6 / 18
10.	Access to Records		
	Audit	25	4 / 26
11.	Charges		
	Tax,etc	26	5 / 12
12.	○ Industrial Property Right		
	Industrial Property Right	132	8 / 55
13.	Improvements made by Company		
	Force Majeure	6	1 / 44
14.	Use of Trademarks		
	Modification,Entire Agreement	11	2 / 14
15.	Warranty		
	Arbitration,Justification	6	1 / 39
16.	Secrecy		
	Confidentiality	9	2 / 14
17.	Term		
	Duration	9	1 / 8
18.	○ Termination		
	Termination,Cancellation	93	14 / 35
19.	Step after Termination		
	Confidentiality	4	1 / 14
20.	○ Arbitration		
	Arbitration,Justification	122	7 / 39
21.	Applicable Law		
	Governing Law	7	1 / 5
22.	○ Language		
	Language	102	4 / 17
23.	○ Headings		
	Heading	59	6 / 14

条項名の上段は、契約書に記載されている条項名である。また、下段は推定された条項名であり、重要度の累計の最も高いものを表示 (一部、2 位まで表示) している。
○：一致
×：不一致
○、×のないものは、一覧表に該当する条項がないものである。

の内、書式集から作成した重要語の一覧表にある条項は8個であり、その内、推定結果が一致したものが7個あった。すなわち、適合率は87 %となる。一致しなかった条項においても、累計結果の2位に正しい条項名が来ている。なお、表1の比率とは、当該条項の一覧表にある重要語の数と契約書の当該条項に出現した重要語の数 (注：出現回数ではない) の比率を表す。

なお、他の2つの契約書における適合率は、それぞれ 89 % (8/9) と 100 % (11/11) であった。

ただし、実験で用いた契約書は、重要語の一覧を求めた書式集と同じ出版社のものである。したがって、今後、他社から出版されている書式集で実験することも必要であると思われる。

3.3 利用方法

本研究で作成した重要語の一覧と作成したプログラムを用いて、実際の契約書の条項間の関係を知ることが可能となる。さらに、契約書に記載されていない項目の提示、または重要語に記載されていない語の提示が可能となる。

4 語間の関連度

本章では、品詞の異なる語間の関連の強さ (関連度という) を考える。なお、同じ品詞間の関連度については、本章および次章で述べる関連度と、"and" または "or" などの等位接続詞で列挙されたか否かの情報などを用いて類似性として数値化することを検討中である。

本研究で扱う品詞の組み合わせ (順序関係のある2つ組) と関連度から推定したい内容を以下に示す。

- 名詞と動詞 (主語と述語)
- 動詞と名詞 (述語と目的格または修飾部)
- 動詞と形容詞 (述語と補語)
- 形容詞と名詞 (形容詞句、副詞句)
- 前置詞と名詞 (形容詞句、副詞句)
- 動詞と前置詞 (動詞句)
- 動詞と副詞 (動詞句)

一般に、文書の種類に依存しない語[1](汎用語という) の出現頻度と専門用語の出現頻度を比較した場合、専門用語の出現頻度が低くなる傾向がある。また、汎用語と組を成す要素の数と専門用語と組を成す要素の数の比較においても、専門用語と組をなす要素の数の方が小さくなる傾向がある。例えば、本研究で使用する書式集に含まれる 36 種類の条項、126,260 語中、語幹 "have" の出現頻度は 504 回、語幹 "deduct(控除する)" の出現頻度は 21 回である。また、重複を許した動詞と名詞の 2 つ組 51,000 組の内、動詞 "have" と 2 つ組を成す名詞の要素数は 449 個、"deduct" と 2 つ組を成す名詞の要素数は 34 個である。

本研究では、語 l と r の 2 つ組の関連の強さに、語 l と組を成す要素数および語 r と組を成す要素数の多寡を加味した値を定義する。すなわち、2 つ組の出現頻度 $f(l,r)$ が同じである 2 つ組 (l,r) が複数ある場合、組を成す要素数が少ない程、関連度が大きくなる (表 2、No.9、14、22 を参照のこと)。
関連度の定義を 4.1 節に示す。

なお、双方の語頻度を加味した指標として、Dice 係数 [9] がある。

[1]ここでの語とは、動詞、名詞、形容詞を意味する。

表 2: 名詞 "royalty" と関連のある動詞

No.	動詞・名詞 (l,r)	関連度	Dice	$f(l,r)$	$\lvert N_l^r\rvert : \lvert N_r^l\rvert$
1	pay, royalty	229.0	0.268	108	145 : 311
2	have, royalty	53.8	0.062	29	145 : 449
3	make, royalty	53.4	0.066	28	145 : 418
4	grant, royalty	53.2	0.061	23	145 : 238
5	agree, royalty	50.3	0.058	26	145 : 402
6	be, royalty	46.8	0.017	26	145 : 485
9	deduct, royalty	40.9	0.049	11	145 : 34
14	authorize, royalty	27.0	0.040	10	145 : 139
22	provide, royalty	21.1	0.025	10	145 : 315

Dice :文献 [9] に記載されている Dice 係数

$Dice(l,r) = \frac{2 \cdot f(l,r)}{f(l) \cdot f(r)}$

$f(l)$、$f(r)$ は、語頻度を表す。

4.1 語間の関連度の定義

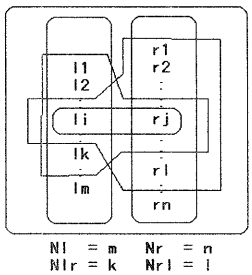

図 2: 語間の関連

英文契約書に出現する語の集合を G とする。

$G = \{(s,t) \mid s$ は文字列、t は品詞 $\}$

いま、2 つの語の組 (l,r) について考える。

lと同じ品詞を持つ語の集合を N_l、

$$N_l = \{(s,t) \mid t = t_l\}$$

rを含み、lと同じ品詞を持つ2つ組の集合 N_l^r、

$$N_l^r = \{((s,t),(s_r,t_r)) \mid t = t_l\}$$

rと同じ品詞を持つ語の集合を N_r、

$$N_r = \{(s,t) \mid t = t_r\}$$

lを含み、rと同じ品詞を持つ2つ組の集合 N_r^l、

$$N_r^l = \{((s_l,t_l),(s,t)) \mid t = t_r\}$$

英文契約書における語lとrの2つ組の出現回数 $f(l,r)$ とするとき、語lとr間の関連の強さ RB(l、r) を以下のとおり定義する。

注：記号 | N | は、有限集合の元の個数を表す。

$$RB(l,r) = f(l,r) \cdot \left(log_2 \frac{|N_l|}{|N_l^r|} + log_2 \frac{|N_r|}{|N_r^l|} \right) \cdots (2)$$

4.2 2つ組の抽出と関連度

長文で複合文である契約書の条文を節に分割することは困難である。単純に、考えられる全ての区切り記号、関係代名詞、関係副詞、接続詞などで、条文を区切った場合、名詞および動詞が必ず含まれる語群よりも、複数の語からなる文の断片の方が多くなってしまう。また、*Brill Tagger* によりある程度の品詞付けができていることから、部分的に文法規則が適応できる状態にあるため、8語または10語というように定数語数で文を分割することも得策ではない。

そこで、今回の実験では、“、”、“：”、“；”、“（”、“）”、“that”、“which”、“if” で、条文を分割したものを節と仮定して処理を行った。

次に問題となるのが、2つ組の求め方である。修飾語句の挿入の多い条文において、「動詞の直前にある名詞が主語である」というような単純な仮定は、意味を成さない。

今回の実験では、以下のように重複を許した2つ組を作成した。

例えば動詞と名詞の語群、

$n1$	$v1$	$v2$	$n2$	$n3$	$n4$

の場合、動詞と名詞からなる2つ組は、$(v1,n2)$、$(v1,n3)$、$(v1,n4)$、$(v2,n2)$、$(v2,n3)$、$(v2,n4)$、の6個、名詞と動詞からなる2つ組は、$(n1,v1)$、$(n1,v2)$ の2個である。

上記の手順により、前述の書式集2315条文の中から、名詞とその後ろに出現する動詞の組、動詞とその後ろに出現する名詞の組、および、動詞とその後ろに出現する形容詞の組、計51,000組についての関連度を求めた（最大値811、最小値3）。

4.3 利用方法

これらの組の中から、意味的に正しくない組を削除し、情報抽出の際に抽出結果の正誤を決める指標として用いる。

5 ある語と語群の関連度

文法上の関係を意識した語群、例えば、主語、述語、目的格の3つ組 (l,c,r) において、ある要素からみた他の要素群との関連の強さの数値化を考える。すなわち、3つ組 (l,c,r) において、主語lと他の要素からなる2つ組 (c,r)、または述語cと他の要素からなる2つ組 (l,r) の関連度の数値化を行う。

本稿では、主語、述語、目的格として求めた、3つ組 (l,c,r) を対象とし、動詞cと他の要素からなる2つ組 (l,r) との関連度について述べる。

表3(company, terminate, license) と、表4(company, cancel, license) は、終了条項において、共に3つ組の出現頻度が1であるが、動詞 “terminate” と組を成す2つ組の要素数は19個、“cancel” と組を成す2つ組の要素数は5個であり、この2つ組を成す要素数の相違が関連度に反映されている。

終了条項において、名詞 “company” および “agreement” と3つ組を成す動詞の一覧を表5に示す。

5.1節に定義を示す.

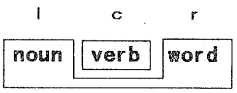

図 3: 語と語群との関連

5.1 語と語群の関連度の定義

英文契約書に出現する語の集合 G とする。

$$G = \{(s,t) \mid s \text{ は文字列}、t \text{ は品詞}\}$$

3つの語の組 (l,c,r) について、それぞれの語の品詞を t_l、t_c、t_r とする。

語lとrのそれぞれの品詞と同じ品詞を持つ3つ組の集合 N_{lr}、

$$N_{lr} = \{((s_1,t_1),(s_2,t_2),(s_3,t_3)) \mid (s_i,t_i) \in G, t_1 = t_l, t_3 = t_r\}$$

語cを含み、語lとrのそれぞれの品詞と同じ品詞を持つ3つ組の集合 N_{lr}^c、

$$N_{lr}^c = \{((s_1,t_1),(s_c,t_c),(s_3,t_3)) \mid (s_i,t_i) \in G, t_1 = t_l, t_3 = t_r\}$$

英文契約書における3つ組 (l,c,r) の出現回数を $f(l,c,r)$ とするとき、関連度 RT(l,c,r) を以下のとおり定義

する。

$$RT(l,c,r) = f(l,c,r) \cdot log_2 \frac{|N_{lr}|}{|N_{lr}^c|} \cdots\cdots (3)$$

5.2 3つ組の抽出と関連度

2章で述べた契約書の特徴を利用して、契約書中に含まれる権利・義務を表す主語、述語、目的格の3つ組の抽出を以下の手順で試みた。

1. *Brill Tagger* を用いて品詞付けを行い、その後品詞辞書との照合および、気付いた範囲で品詞付けの修正を行う。
2. 2章で述べた権利・義務を示唆する語を含み、かつ、条件を示唆する語を含まない条文を求める。
3. 求めた条文の中で、権利・義務を示唆する語の前に出現する名詞で、その名詞の直前に前置詞を含まないものを主語の候補とする。
4. 権利・義務を示唆する語の直後に出現する動詞を述語候補とし、述語候補の直後に出現する名詞を目的格候補とする。
5. 受動態の場合 ("*be*" 動詞の次に過去分詞が現れ、かつ、名詞の前に "*by*" がある場合) は、主語と目的格を入れ換える。
6. 各条項毎に、求めた3つ組の出現回数をもとめ、(3) 式により関連度を計算し、降順に並べ替えを行う。

上記の手順では、権利・義務を示唆する語 (helping verbs) の前に出現した名詞の内、文頭に近いものが主語候補となり、権利・義務を示唆する語の直後に出現した動詞が述語候補となる。また、述語候補の直後に出現した名詞が目的格候補となる。例えば以下の場合、

n_1	n_2	help	v_1	v_2	n_3	n_4

権利・義務を表す3つ組 (n_1, v_1, n_3) が抽出される。

このように、名詞句についての情報、および不完全他動詞か完全他動詞かというような詳細な品詞情報を用いず、機械的に3つ組を求めているため、主語、述語、目的格という文法関係を正しく求められていないという問題がある。

上記の手順で抽出した結果、終了条項において、人手で求めた権利または義務を表す3つ組は51種類 (正解とする) であり、前述の手順では42種類の3つ組が抽出され、その内の正解は26種類であった (26／51)。したがって、再現率は50％、適合率は61％ (26／42) となる。

表 3: 終了条項における "terminate" を含む3つ組

主語	述語	目的格	$I : RT : f(l,c,r)$
○ company	terminate	agreement	6.00: 45.40: 15
○ party	terminate	agreement	5.20: 33.30: 11
999-days	terminate	agreement	2.90: 9.10: 3
notice	terminate	agreement	2.40: 6.10: 2
business	terminate	agreement	1.90: 3.00: 1
assignment	terminate	999-days	1.90: 3.00: 1
○ company	terminate	license	1.90: 3.00: 1
change	terminate	agreement	1.90: 3.00: 1
readjustment	terminate	occurrence	1.90: 3.00: 1
right	terminate	license	1.90: 3.00: 1
right	terminate	obligation	1.90: 3.00: 1
year	terminate	license	1.90: 3.00: 1
day	terminate	agreement	1.90: 3.00: 1
remedy	terminate	agreement	1.90: 3.00: 1

○ : 正しく抽出された3つ組
I : (1) 式で求めた重要度
RT : (3) 式で求めた関連度
$f(i,c,r)$: 出現頻度
N_{lr} : 155 (終了条項内)
$N_{lr}^{terminate}$: 19 (終了条項内)

表 4: 終了条項における "cancel" を含む3つ組

主語	述語	目的格	$I : RT : f(l,c,r)$
○ party	cancel	agreement	2.70: 9.90: 2
licensee	cancel	notice	1.90: 5.00: 1
○ company	cancel	license	1.90: 5.00: 1
○ company	cancel	agreement	1.70: 5.00: 1
licensee	cancel	licensee	1.90: 5.00: 1

○ : 正しく抽出された3つ組
I : (1) 式で求めた重要度
RT : (3) 式で求めた関連度
$f(i,c,r)$: 出現頻度
N_{lr} : 155 (終了条項内)
N_{lr}^{cancel} : 5 (終了条項内)

表 5: 終了条項における "company" と "agreement" を含む3つ組

主語	述語	目的格	$I : RT : f(l,c,r)$
○ company	terminate	agreement	6.0: 45.4: 15
○ company	reinstate	agreement	1.9: 7.3: 1
○ company	cancel	agreement	1.7: 5.0: 1

○ : 正しく抽出された3つ組
I : (1) 式で求めた重要度
RT : (3) 式で求めた関連度
$f(i,c,r)$: 出現頻度
N_{lr} : 155 (終了条項内)
$N_{lr}^{terminate}$: 19 (終了条項内)
N_{lr}^{cancel} : 5 (終了条項内)
$N_{lr}^{reinstate}$: 1 (終了条項内)

5.3 利用方法

(3) 式で求めた関連度は、3 つ組の正誤を判断する指標として用いる。また、情報抽出の際に、動詞から、候補となるテンプレートを見つけ出す指標として利用する。

6 関連研究

語間の関連を示す指標として相互情報量 [10] と Dice 係数 [9] がある。文献 [9] によると、語 l と r の 2 つ組の出現頻度 $f(l,r)$ が同じである組が複数存在した場合、相互情報量は、語 l と r それぞれの出現頻度の多寡に影響を受けない。そして、最も重要な点は、確率を用いている点である。例えば、4.2 節で述べたように "Company may cancel or terminate this agreement" という節の名詞と動詞の 2 つ組は、(company,cancel) と (company,terminate) の 2 つ、動詞と名詞の 2 つ組は (cancel,agreement) と (terminate,agreement) の 2 つというように、本研究では "Company" と "agreement" を重複して処理を行いたい。しかし重複を許してしまうと、確率を用いることが困難となる。

一方、Dice 係数は、語 l および語 r それぞれの出現頻度の多寡が反映される。しかし、表 2 の $(be, royalty)$ の組を見て明らかなように、$f(be) = 2616$、$f(royalty) = 426$ と、極端に出現頻度の高い語を含む 2 つ組においては、2 つ組の出現頻度 $f(l,r) = 26$ が低く評価されてしまう。

本稿で提案する関連度は、2 つ組の出現頻度に影響を与えない範囲で 2 つ組の各要素の汎用性の違いを明示するという点に特徴がある。しかしながら、本研究で提案する関連度は、2 つ組の文法関係が正しく求まっているか否かにより、その値が左右されるという欠点がある。

7 今後の課題

複雑な条文から、全文についての構文解析をせずに内容を抽出することを目的に、4 章および 5 章で述べた関連度の提案を行った。しかし、本稿で提案した関連度を有効なものにするためには、できるだけ正しい 2 つ組または 3 つ組を求めることが必要である。すなわち、条文の文法関係をできるだけ、正しく求めることが、一番の課題である。そこで、節への分割方法、部分的な構文解析についての検討に加え、今後、条文中の類義語句を代表語 1 語に置き換えるなどの前処理についての検討も行う予定である。

加えて、品詞が同じである語間の類似性を数値化する方法についての検討も行う予定である。

8 まとめ

本稿では、本研究の概要と英文契約書の特徴を述べた後、語の重要度とその重要度の有効性を評価するために、契約書式集から求めた重要語の一覧を使って、3 種類の契約書の条項名の推定を行った結果を示した。

次に、品詞の異なる語間の関連度、および語と語群の関連度の数値化についての提案と、その特徴を示した。

謝辞 日頃から御討論いただく、本学、自然言語処理学講座の松本裕治教授、英語教育の Dee A.Worman 教授と、木村晋二助教授、高木 一義助手はじめ渡邊研究室の皆様に感謝します。

参考文献

[1] Eric Brill: Some Advances in Transformation-Based Part of Speech Tagging,(AAAI-94). http://www.cs.jhu.edu/~brill/acadpubs.html

[2] 英和対訳 取引条件表現法辞典 第 2 巻技術取引, 国際事業開発株式会社, 1992.

[3] 日野修男, 出澤秀二, 竹原隆信, 杉浦幸彦, 水谷孝三：英文契約書の知識と実務, 日本実業出版社, 1997.

[4] 岩崎一生：英文契約書 - 作成の理論と実務 -, 同文舘, 1988.

[5] 中村秀雄：新版 英文契約書作成のキーポイント, 社団法人 商事法務研究会, 1996.

[6] 長尾真, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋：言語の科学 9 言語情報処理, 岩波書店, 1998.

[7] S. グリーンバウム,R. クワーク著, 池上嘉彦 他訳：現代英語文法 大学編, 紀伊国屋書店, 1995.

[8] 安井稔編：コンサイス英文法辞典, 三省堂,1996.

[9] Smadja, McKeown, and Hatzivassiloglou : Translating Collocations for Bilingual Lexicons, Computional Linguistics Volume 22, Number 1, 1-38, 1996.

[10] Kenneth Ward Church, Patrick Hanks : Word Association Norms, Mutual Information, Computional Linguistics Volume 16, Number 1, 22-29, March 1990.