自然言語処理 128−22
（1998. 11. 6）

英文契約書における内容の抽出
―シソーラス作成のための統計情報を用いた類似度計算―

相良 かおる　　渡邊 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、英文契約書に使われる単語の類似度を求める手法について提案する。本手法は、単語の共起頻度に基づく統計的手法の一種であり、内積を用いて類似度を定義している。本手法の特徴は、英文契約書の書式集の条文から、名詞と動詞、動詞と名詞、形容詞と名詞、前置詞と名詞というように統語構造を意識した2つ組を求め、その2つ組間の関連度をベクトルの要素としている点にある。本手法により、英文契約書の書式集(162,298語)に含まれる名詞898種からなる209,274組のペアについて類似度を求め、数量化IV類によるクラス分けにより、81個のクラスの類概念データを作成した。なお、本研究は、英文契約書の内容抽出のための準備研究である。

キーワード：　英文契約書、単語類似度、統計的手法、類概念、シソーラス

Information Extraction From an English Contract
A statistics-based computation of word similarity for making a thesaurus for 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 proposes an approach for similarity measurement of words that are used in a collection of English contracts. This approach is a statistics-based computation of word similarity by a vector consisting of co-occurrence statistics. Using a vector consisting of the correlation between two-tuple of terms on the basis of syntactic behavior (such as the ordered pair verb,noun) is a feature of this approach. We made similarity data of 209,274 pairs from 898 nouns in the collection of English contracts, and made 81 classes from these similarity data with a multi-dimensional scaling. This work becomes a preparatory work for the information extraction from an English contract.

keywords：　contracts in English, word similarity, statistical model, clustering, thesaurus

1 はじめに

近年、一部の大企業に限られていた国際取引契約の締結が、中小企業やベンチャービジネスの場においても行なわれるようになってきた。それに伴い、多くの英文契約文書（以下、契約書という）が作成され、保管されている。

一方、SGML(Standard Generalized Markup Language)、XML(eXtensible Markup Language)などの出現で、従来紙面で保管されてきた文書を構造化された電子化文書の形で保管することが可能となってきた。

その結果、作成された契約書から重要な情報を見つけ出し、抽出することができれば、SGMLおよびXMLを利用することで、契約書を単なる記録としてではなく、新たな取引契約の締結に向けての重要な情報源として、積極的に利用することができる。

契約書から情報を抽出する際、(1) 何を抽出するのか、(2) どのように抽出するのかという問題が生じる。

(1) については、利用者や状況によって重要な情報が異なる。そこで、契約書の条文を単文(主語と述語がそれぞれ一つからなる文)の形で表記し、その中に含まれる重要語句の重要度の総和を単文の重要性の指標として提示し、その指標の利用者による変更を可能とすることで、必要とする情報を利用者が指定できるしくみを検討している。

したがって、(2) の問題は、長文で複雑な契約書を単文に分割する問題であると換言することができ、これが本研究の主たる目標である。

一般に、契約書における条文は、修飾節および句の挿入が多く、長文で複雑である。例えば、本研究で用いる技術取引契約書の契約書式集 [1] に含まれる 4519 条文 162,298 語 (住所や製品名など契約固有の部分は 1 語として換算) の 1 条文あたりの平均語長は、35 語である。そこで、情報抽出の方法としては、文法規則を用いた構文解析による手法よりも、テンプレートを使った手法の方が適切であると考えられる。

テンプレートを使った情報抽出における問題として、テンプレートの作成、抽出項目（テンプレート・エレメント）の作成、抽出方法がある。本研究では、条文の内容を単文の形式で抽出することから、抽出項目の作成と抽出方法が主たる問題となる。更に、全ての契約書を網羅した抽出項目を作成することは困難であることから、抽出すべき項目の推測および利用者による抽出項目の追加・変更についても重要な問題となる。

本研究では、契約書から情報を抽出するために、以下の 9 種のデータとこれらのデータを追加・変更する手法が必要であると考え、前述の書式集および文献 [2,3,4,5,6] 記載の専門用語を基にデータの作成および作成手順と学習のための追加・変更手法についての検討を進めている。

1. 単語の品詞と原形を求めるための辞書 (5,802 語)
2. 単語および隣接する 3 つ組についての出現頻度データ
   (2 個以上の連続する名詞を含む 3 つ組：1,693 組)
3. 条項における重要語句の一覧 (598 語)
   参照データ：
   - 単語の品詞データ (1)
   - 出現頻度データ (2)
4. 品詞の異なる単語間の関連度データ
   (名詞・動詞、動詞・名詞、形容詞・名詞、前置詞・名詞：全 17,367 組)
   参照データ：
   - 品詞データ (1)
   - 出現頻度データ (2)
5. 同じ品詞を持つ単語間の類似度データ
   (898 種の名詞間の類似度：209,274 組)
   参照データ：
   - 関連度データ (4)
   - 類概念データ (7)(将来)
6. 単文形式の抽出項目 (条文に含まれる内容)
   参照データ：
   - 品詞データ (1)
   - 出現頻度データ (2)
   - 重要語句データ (3)
   - 関連度データ (4)
   - 類似度データ (5)
7. 使われる状況が同じである単語のグループとその代表語からなる類概念データ
   参照データ：
   - 3 つ組の出現頻度データ (2)
   - 類似度データ (5)
8. 複合語や合成語などの連語データ (194 種)
   参照データ：
   - 専門用語データ (9)
   - 3 つ組の出現頻度データ (2)
   - 類似度データ (5)
9. 契約書特有の専門用語一覧 (冗長な表現：108 種、専門用語：937 種)
   参照データ：
   - 文献 [2,3,4,5,6]
   - 3 つ組の出現頻度データ (2)
   - 類似度データ (5)

注）参照データとは、データを作成・変更する際に参照するデータを意味する。

なお、(5) 類似度データは、類概念データ (7) をフィードバックして精選される。

本稿では、(5) の類似度データの作成について、名詞を例に、類似度の定義、類似度の求め方、数量化 IV 類を使ったクラス分けの実験結果と考察について述べる。

なお、本研究では、「契約書式集から抽出項目データを作成する際に必要となるのは、主として名詞間および動詞間の類似度データであり、ここでの "類似性" とは、意味的な類似性ではなく、むしろ条文中で使われる状況の類似性である。」と考えている。すなわち、意味的に類似している複数の単語について、使われる状況が異なる場合、その相違が数値で表現されることの方が、意味的な類似性の数値化よりも重要であると考える。

2 単語間の類似度

単語間の類似度を求める手法は、大きく 2 つに大別される。一つは単語の共起頻度に基づく統計的手法であり [7,8,9]、もう一つはシソーラスを用いる手法 [10] である。なお、シソーラスと統計情報を統合した単語の類似度計算の手法も提案されている [11]。

今回、英文契約書に使われる用語からなるシソーラスをできるだけ機械的に作成することを目標に、品詞の異なる単語の 2 つ組の関連の強さを使った類似度計算の手法を提案する。

関連の強さを求める指標の最も単純なものとして 2 つ組の出現頻度があるが、本研究では、我々が提案した関連度 [12] を用いて類似度を求める。2 つ組を成す要素の汎用性が数値に反映されるという点が、関連度と出現頻度の相違である。

例えば、動詞-名詞の 2 つ組 (authorize,royalty) と (provide,royalty) の書式集における出現頻度は共に 10 回であるが、動詞 "provide" と組を成す名詞の数が 315 個であるのに対して、動詞 "authorize" と組を成す名詞の数は 139 個であり、動詞 "provide" の方が汎用性が高い。我々が提案する関連度は、この汎用性の相違が数値に反映されるという特徴がある。すなわち、(authorize,royalty) と (provide,royalty) の関連度はそれぞれ、27.0 と 21.1 になる。

名詞間の類似度を求めるために用いる 2 つ組は以下の 4 種である。

1. 名詞と動詞
2. 動詞と名詞
3. 形容詞と名詞
4. 前置詞と名詞

同様に動詞間の類似度を求めるために用いる 2 つ組は以下の 4 種である。

1. 名詞と動詞
2. 動詞と名詞
3. 動詞と形容詞
4. 動詞と副詞

すなわち、名詞間の類似度 SIMnn は、名詞と動詞から求めた類似度を SIMnv、動詞と名詞から求めた類似度を SIMvn、形容詞と名詞から求めた類似度を SIMan、前置詞と名詞から求めた類似度を SIMpn としたとき、以下のように定義する。

$$SIMnn = (SIMnv \cdot w1 + SIMvn \cdot w2 + SIMan \cdot w3 + SIMpn \cdot w4) \,/\, 4 \cdots (1)$$

ここで、w1、w2、w3、w4 は、それぞれに係る重みである。本研究では、前置詞と名詞の関連度から求めた類似度は、他の 3 種の類似度に比べて、名詞間の類似度に与える影響が小さいと考え、w1、w2、w3 を 3、w4 を 2 にして計算を行った。

我々は、統語構造を意識した 2 つ組を用い、かつ、出現頻度ではなく関連度をベクトルの要素として用いることで、意味的な類似性についても類似度に反映させることができると考える。このように、統語構造を意識して類似度を求める点が本手法の特徴である。

なお、関連する研究に統語構造を使い、内積ではなく、KL-distance(the Kullback-Leibler distance) を改良して類似度を求めるものがある [8]。

2.1 節では品詞の異なる 2 つ組の抽出方法を、2.2 節では関連度の定義について述べた後、2.3 節では類似度の定義を行なう。

2.1 品詞の異なる2つ組の抽出

本稿で提案する類似度は、品詞の異なる2つ組の関連度を基にしている。したがって類似度の精度は、関連度を求める際の2つ組の係り受けの正しさに大きく依存している。

係り受けの正しい2つ組を求めるためには、複合文である条文を正しい節に分割することが重要となる。しかしながら、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*[13]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、本研究では、“.”、“:”、“;”、“(”、“)”、“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。

以上の方法で2つ組を求めることで、単語については、係り受けの正しい組を網羅することができる。しかし、複合語(合成語)、および慣用句についての考慮がなされていない。

そこで、第1章で述べた契約書特有の専門用語一覧データ(9)と照合し、一致した語句については、ハイフン“−”で連結する前処理を行なった。加えて、受動態および完了形についても、述語動詞の部分をハイフン“−”で連結する前処理を行なった。

2.2 関連度の定義

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つ組の集合を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つ組の集合を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つ組の出現回数を$f(l,r)$とするとき、語lとr間の関連の強さREL(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。

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

2.3 類似度の定義

ある語xkと名詞niとの関連度を関連度の最大値で割ることで正規化した値を改めて$REL(xk,ni)$とする。語xkと同じ品詞tを持つ語と名詞niの2つ組の集合をN_t^{ni}、語xkと同じ品詞tを持つ語と名詞njの2つ組の集合をN_t^{nj}とすると、N_t^{ni}とN_t^{nj}の共通の要素の集合COM_{ni}^{nj}は、

$$COM_{ni}^{nj}=N_t^{ni}\cap N_t^{nj}$$

となる。

名詞niと名詞njの類似度SIMtn(ni,nj)を以下のとおり定義する。

$$SIMtn(ni,nj)=\frac{\sum_{k=1}^{|COM_{ni}^{nj}|}\left(REL(xk,ni)\cdot REL(xk,nj)\right)}{\sqrt{\sum_{k=1}^{|COM_{ni}^{nj}|}\left(REL(xk,ni)\right)^2}\cdot\sqrt{\sum_{k=1}^{|COM_{ni}^{nj}|}\left(REL(xk,nj)\right)^2}}\cdot\frac{2\cdot|COM_{ni}^{nj}|}{|N_t^{ni}|+|N_t^{nj}|}\cdots(3)$$

すなわち、類似度SIMtn(ni,nj)は、語xkと同じ品詞tを持つ語との2つ組の中の共通の組の

関連度をベクトルとした内積の余弦に、全体の組に対する共通の組の比率を掛け合わせたものである。

第3章では、技術取引契約書の書式集4596条文(162,298語)から、名詞の類似度を求める手順および作成した類似度データの一部を示す。

3 名詞間の類似度

3.1 類似度データの作成手順

以下に名詞間の類似度データの作成手順SimNNを示す。

＜手順 SimNN＞

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。
   スペルミス：1種
   タグミス：158種
   条文 ：4,596文
   語数 ：162,298語
2. 専門用語データと一致した複合語句などをハイフン“−”でまとめる。
   専門用語：168種
   外来語：10種
   等位句：16種
   固有名詞を含む名詞句：27種
   条文 ：4,596文
   語数 ：161,621語 (− 0.4 %)
3. 完了形および受動態の述語動詞の部分をまとめる。
   条文 ：4,596文
   語数 ：158,397語 (− 2.4 %)
4. 単語および隣接する3つ組の出現頻度データを求める。
   連続して2つ以上の名詞が並ぶ3つ組：1,693組
5. 名詞の種類および2つ組を求める
   名詞の数：898種
   組合せの数：401,856通り
6. 名詞と動詞、動詞と名詞、形容詞と名詞、前置詞と名詞の2つ組とその出現頻度および2つ組間の距離を求める。
   名詞と動詞：6,380組
   動詞と名詞：7,592組
   形容詞と名詞：814組
   前置詞と名詞：2,568組
   総計：17,367組
7. 前述の式(2)より関連度を求め正規化を行なう。
   注）今回は、最大値が1になるように、0〜1の間で正規化を行なったが、検討が必要
8. 名詞と動詞、動詞と名詞、形容詞と名詞、前置詞と名詞のそれぞれについて、前述の式(3)より名詞間の類似度(401,856組)を求める。
9. 前述の式(1)より最終的な名詞間の類似度を求める。
   類似度の最大値：0.84
   類似度 ＞ 0 の組：209,274組 / 401,856組

3.2 類似度データの結果と考察

表1は、ある名詞と0.3以上の類似度を持つ名詞を類似度の降順にまとめたものの一部である。

項目1.欄の○印は、2つ以上の名詞が隣接する3つ組の出現頻度データに名詞1と2を要素とする組が含まれることを意味する。また、項目2.欄の○印は、名詞1を含む条文に名詞2も出現することを意味する。すなわち、1欄に○が記されているものは、複合語の可能性の高い組であり、2欄のみに○が記されているものは、等位関係にある名詞である可能性が高い組となる。そこで、類似度データから1欄に○のある語を削除し、クラス分けを行なうことで、契約書において使われる状況が類似している名詞の候補をまとめることが可能となる。

今回提案する類似度計算の手法は、シソーラスを作成するためのものであり、当然ながら意味的な要素を考慮していない。したがって、次のような不都合が生じる。

表1における、disclosure(開示)と類似度の高い語はすべて契約書の内容を抽出する上で適切な組合わせではない。すなわちこれらは、類義語ではないし、2.欄に○印がない（同じ条文に出現していない）ことから、述部を構成する名詞群でもない。これらの語の類似度が高く評価されたのは、全ての要素が形容詞(full)、前置詞(in,of)および動詞(be)と2つ組をなし、2.1.2節の(3)式における2つ組全体の総数に対する共通の組の比率が高く評価されたことに加え、第2章(1)式の前置詞と名詞の重みが2、形容詞と名詞の重みが3であることが影響したものと推測される。

表2は、関連度を用いた類似度と出現頻度を用いた類似度について比較した表である。関連度を基にした場合の利点として、汎用的な語との関連度が低くなることから、複合語の候補

表 1: ある名詞と 0.3 以上の類似性を持つ名詞

名詞 1　名詞 2	類似度	1	2
accident 不可効力			
epidemic 伝染病	0.697		○
fire 火災	0.622		○
strike ストライキ	0.618		○
flood 洪水	0.528		○
war 戦争	0.476		○
appropriation 盗用	0.455		○
registration 登記	0.418		
agency 代理	0.402		○
embargo 通商停止	0.400		○
riot 暴動	0.388		○
lockout 工場閉鎖	0.318		○
inability 無力	0.310		○
airmail 航空郵便			
post 郵便	0.562		○
letter 書簡	0.345		
postage 郵送料	0.335		○
telex テレックス	0.335		
telegram 電報	0.327		
cable 電信	0.322		○
disclosure 開示			
content 内容	0.483		
respect 関連	0.456		
copyright 著作権	0.404		
force 効力	0.374		
title 財産所有権	0.364		
exchange 両替、為替			
bank 銀行	0.456	○	○
currency 通貨	0.447		○
rate 率	0.407	○	○
object 目的			
code コード	0.599	○	○
source ソース	0.459		○
section 条			
sentence 文	0.478		○
combination 組合せ	0.474		○
article 条	0.440		
clause 条	0.437		○
workmanship 製品	0.349		
style 形状	0.336		
procedure 手順	0.302		
trade 同業者			
discount 割り引き	0.433	○	○
trademark 商標	0.365		
name 名称	0.341	○	○
secret 秘密	0.328	○	○
mark マーク	0.314	○	○
warranty 保証			
representation 保証, 代理	0.499		○
discussion 検討	0.455		
indemnity 損害補償	0.357		
understanding 協定	0.329		

1. : 3 つ組の出現頻度データに名詞 1 と 2 が含まれる場合に○をマーク
2. : 名詞 1 を含む条文中に名詞 2 が出現する場合に○をマーク

表 2: 関連度を基にした類似度と出現頻度を基にした類似度の比較

case. 名詞 1　名詞 2	類似度	1	2
a. (r) equipment 設備			
machinery 機械設備	0.345		
component 部品	0.342		○
(f) equipment 設備			
replacement 取り替え	0.477	○	○
part 部分	0.351	○	○
machinery 機械設備	0.344		
component 部品	0.340		○
b. (r) none			
(f) subject 主題			
matter 事柄	0.356	○	○
c. (r) none			
(f) company 会社			
licensee 被許諾者	0.444		○
party 当事者	0.317		○

(r) : 関連度
(f) : 出現頻度
none: 該当項目なし
1. : 3 つ組の出現頻度データに名詞 1 と 2 が含まれる場合に○をマーク
2. : 名詞 1 を含む条文中に名詞 2 が出現する場合に○をマーク
注) 類似度が 0.3 以上の組を対象とする。

が選ばれるというケースが少なくなる (equipment replacement parts, subject matter)。一方、同様の理由から、条文中に多く出現する汎用的な名詞同士の項目が落されるケースが生じる (company, licensee, party)。

今回、類似度を求める際に、関連度 (最大値 2,057、平均値 42.6、最頻値 12) と 2 つ組の出現頻度 (最大値 1,034、平均値 5.6、最頻値 2) について、最大値で割るという正規化を行なった。しかし、平均値または最頻値から明らかなように、この正規化は適切ではない。したがって、今回の類似度データから、類似度を求める際のベクトルの要素として、関連度と出現頻度のどちらが妥当であるかについて決定することはできない。

本手法では、語の数を N とした時、類似度のベクトルを計算するのに $O(N^2)$ のメモリ空間を必要とし、なおかつ、(1) 式で明らかなように 4 種類の類似度計算を行なう必要があることから、1 回の名詞間の類似度計算に 1 週間を要した。その結果、考えられる様々なケースについて気軽に実験することが困難であった。

そこで今後、関連のないことが明らかな組合せを排除するなどの工夫をし、関連度の平方根

をとって大きな値間の差を抑えた場合などについて類似度計算を行ない、出現頻度との比較検討を行なう予定である。

4 数量化 IV 類によるクラス分け

前述の手順で作成した類似度データを数量化 IV 類を用いて、クラス分けを行なった。

数量化 IV 類とは、あるn個の対象において、2つずつのペアの間に親近性 $S_{ij}, i \neq j$ (値が大きい程親近性が強い) が定義できるものとするとき、この親近性に基づいて対象に数量を与え、親近性の大きいペアは近くに、親近性の小さいペアは遠くになるように、ユークリッド空間内に位置づけしようという方法である。[14]

今回の実験では、インターネット上で公開されている多変量解析ツール [15] を利用して、0.3 以上の類似度を持つ 446 個の名詞から作成した非類似度行列を基に固有値および固有値ベクトルを求め、クラス分けを行なった。

クラス分けの手順は以下のとおりである。

＜手順 ClassNN ＞

1. 0.3 以上の類似度をもつ名詞のペアを抽出し、対角成分が 0 で、似ていないもの程負の絶対値の大きな値を持つ非類似度行列を求め、固有値および固有ベクトルを求める。このとき、類似度が 0.3 以下のペアに対しては、類似度 0 と置き換えて非類似度行列を求める。
   対象となる名詞：446 種
2. 固有ベクトルの小数点以下 4 桁目を落し、固有値の降順に対応する固有ベクトルをキーに降順ソートを繰り返す。すなわち最も大きな固有値に対応する固有ベクトルを第一キー、2番目に大きな固有値に対応する固有ベクトルを第二キーというようにして、クラスの要素数が 10 個以内になるまで、繰り返す。
3. 2つ以上名詞が連続する3つ組の頻度データと照合し、一致したものを一つにまとめ、要素が2つ以上のものを1クラスとする。(81 クラス,166 語)
4. 同じ条文中に共に出現する要素にマークをつける。
5. クラスの要素について、その妥当性を人手によりチェックする。

4.1 結果と考察

表 3 は、手順 ClassNN の 4 の結果の一部に、第 1 章で述べた、重要語句一覧 (3)、条項

表 3: 名詞の概念クラス

No.	要素	等位	関係
1	a. accountant† 公認会計士 b. counsel† 法律顧問		$a \neq b$
2	a. advertisement 公告 b. promotion 販売促進		$a \subseteq b$
3	a. air mail† b. letter c. airmail† d. cable† e. envelope† f. post† g. postage† h. telegram† i. telex†	a-f a-d b-d c-d d-e d-g d-h d-i	$a \subseteq b$ $c \subseteq b$ $d \subseteq b$ $e \subseteq c$ $e \subseteq d$ $e \equiv f$ $e \equiv g$ $e \equiv h$ $e \equiv i$
4	a. blueprint† 計画 b. plan† 計画		$a \subset b$
5	a. currency† 通貨 b. exchange rate† 為替レート		$a \subseteq b$
6	a. disturbance† 騒動 b. enemy† 敵国 c. inability† 無力	a-b b-c	$a \subset c$ $a \equiv b$
7	a. election† 選挙 b. purpose 意図		$a \subset b$
8	a. endeavour† 努力 b. evidence 証拠 c. instrument† 手段 d. proof 証拠		$a \neq b$ $c \subset b$ $d \subset b$ $c \neq d$
9	a. expiration 満了 b. termination 終了、満期	a-b	$a \subseteq b$
10	a. insurrection† 暴動, 反乱 b. stoppage 支払い停止	a-b	$a \equiv b$
11	a. logotype 意匠文字 b. servicemark 標章、文句	a-b	$a \supseteq b$
12	a. maintenance 維持 扶養 b. recommendation 忠告 c. total† 総計		$a \supset b$ $a \supseteq c$
13	a. object code b. source code	a-b	$a \equiv b$
14	a. product sale 製品販売 b. sale price 販売価格 c. territory sale 地域販売	a-b	$a \supseteq b$ $a \supseteq c$ $b \subseteq c$
15	a. report 報告書 b. statement 計算書	a-b	$a \supseteq b$

†：ある条項に特出する語を表す (重要語)。

等位：同じ条文中に出現する組を表す。

a-b：語 a と b が同じ条文中に出現する。

関係：語が出現する条項の共通性を表す。

$a \equiv b$：語 a と b は出現する条項が同じである。

$a \neq b$：語 a と b が出現する共通の条項がない。

$a \subset b$：語 a の出現する条項数の半数以下の条項に b が出現し、b の出現する条項数は a より多い。

$a \subseteq b$：語 a の出現する条項数の半数以上の条項に b が出現し、b の出現する条項数は a より多い。

単位の出現頻度データ (2) などから得られる情報の概略を付加したものである。この表から、例えば *accountant* と *counsel* は、ある条項における重要語であり、また出現する条項に共通のものはないことから、構文上での使われ方が似ている語である、すなわち、意味が似ているというより抽象的な概念が似ている語であると推測することができる。また、*advertisement* と *promotion* は、*advertisement* が出現する条項の半数以上の条項に *promotion* も出現し、かつ、*promotion* の方が多くの条項に出現する。加えて、同じ条文に共起することがないため、*promotion* は、*accountant* を包含した意味をもつ類義語であると推測される。このようにして、クラス内の語が類義語か、等位関係にある語か、述語を構成する語であるかなどを機械的に推測することで、シソーラスを作成する際の人的労力を削減することが可能となる。しかし、そのためには、類推規則を定める必要がある。

なお、現在のところ、シソーラスのデータ構造については、何も決まっていない。

今回の手順 ClassNN は、インターネット上で公開されているツールを利用しているため、データの変換や操作に占める人的な作業量が大きい。そこで、正規化などを検討して新たに求めた類似度データを再度数量化 IV でクラス分けするとともに、他のクラス分けの手法についても検討し、一連のプログラムでクラス分けができる手順を決める予定である。

5 まとめと今後の課題

本稿では、英文契約書の内容抽出に必要なシソーラスを作成することを目的とした類似度計算の手法を提案し、英文契約書の書式集に含まれる名詞間について類似度をもとめ、クラス分けを行なった。

今後の課題として以下の検討項目がある。

1. 英文契約書の情報抽出に必要なシソーラスのデータ構造
2. シソーラスを用いた類似度計算手法 (動詞間の類似度計算)
3. クラス分けの手法
4. 作成データの自動的または機械的更新 (学習) 手法
5. 書式集から内容の抽出 (抽出項目の作成)
6. 英文契約書からの内容抽出手法

謝辞 インターネット上で、多変量解析ツール Black-Box を公開して下さり、使用方法などを親切に御教示下さった群馬大学の青木繁伸教授、および、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.Waman 教授と、木村晋二助教授、高木 一義助手はじめ渡邉研究室の皆様に感謝します。

参考文献

[1] 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.

[2] 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.

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

[4] 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.

[5] 阿部佳基, 長谷川俊明 : ビジネス法律英語辞典, 日本経済新聞社, 1991.

[6] 宮野準治, 飯泉恵美子 : 英文契約書の基礎知識, The Japan Times, 1997.

[7] Eugene Charniak : Statistical Language Learning, The MIT press, 1993.

[8] Wide R. Hogenhout, Yuji Matsumoto : A Preliminary Study of Word Clustering Based on Syntactic Behavior, Proceedings of the Workshop on Computational Natural Language Learning, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, 16-24, 1997.

[9] Yael Karov, Shimon Edelman : Similarity-based Word Sense Disambiguation, Computional Linguistics Volume 24, Number 1, 41-59, 1998.

[10] Jen Nan Chen, Jason S. Chang : Topical Clustering of MRD Senses Based on Information Retrieval Techniques, Computational Linguistics, Volume 24, Number 1, 61-95, 1998.

[11] 藤井 敦, 徳永建伸, 田中穂積 : シソーラスと統計情報を統合した単語の類似度計算について, 情報処理学会, 自然言語処理研究会,120-8,1997.

[12] 相良かおる, 渡邉勝正 : 英文契約書における要目の抽出, 電子情報通信学会技術研究報告, NLC98-19, 29-36,1998.

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

[14] 田中 豊, 脇本和昌 : 多変量統計解析法, 現代数学社, 1983.

[15] 青木繁伸 : Black-Box 多変量データ解析, http://aoki2.si.gunma-u.ac.jp/BlackBox/BlackBox.html.