NAIST-IS-DT9661013

博士論文

英文契約書の電子文書化に関する研究

相良 かおる

2000年2月7日

奈良先端科学技術大学院大学
情報科学研究科 情報システム学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である。

相良 かおる

審査委員： 渡邉 勝正 教授
松本 裕治 教授
植村 俊亮 教授

英文契約書の電子文書化に関する研究*

相良 かおる

内容梗概

毎年，日本企業によって数百社の海外子会社が新規設立されている．また，英語を母国語としない国においても現地人とのコミュニケーションに英語が使われる場合が多いとの報告がある．従って，海外子会社を介した英文契約書の数も増加していると推測される．一方，英国，ドイツでは世界各国に点在する契約書の共有を目的とした分散管理システムの研究が報告され，また，契約書からの知識獲得の必要性が議論されている．日本語と英語間の契約書の機械翻訳の要求の中，契約書の機械翻訳が困難であるとの報告もある．

本論文では英文契約書の構造文書化について，(1) 構造文書化に必要な専門用語などの辞書類の作成，(2) 電子化における索引語の作成ならびに情報検索の支援を目的とした，(a) 語の基本形，(b) 条項の内容を表すための重要語，(c) 語の類似度計算とクラスタリングについて述べる．最後に (3) 契約書の構造文書化に必要な条文の解析方法について述べる．

(a)「語の基本形」は，対象とする語から語形を変化させる屈折接辞を取り除いた「語幹」から更に語の品詞を変える派生接辞を取り除いたものをいう.「語の基本形」を導入することで，本研究で作成した用語辞書に含まれる 5,804 種の語を 2,854 種の基本形にまとめることができる．本研究では,「語の基本形」を求めるために 97 種類の接尾辞から品詞を変える接尾辞 85 種類を抽出し，語形を変換させる 262 種類の規則を作成している．

(b) 契約文書には，条項ごとに内容がまとめられているという特徴がある．そこで，索引語を抽出する際に使われる TF.IDF 法を応用して，技術取引に関する

*奈良先端科学技術大学院大学 情報科学研究科 情報システム学専攻 博士論文, NAIST-IS-DT9661013, 2000年 2月 7日.

契約書式集から，34 条項を対象に，重要語を求めた．そして，異なるテキストデータにおいて条項名の推定を行った．その結果，出版社の異なる書式集から抽出したテストデータにおいても 79 %の正しさで条項名の推定ができ，かつ「語の基本形」を導入することで，平均 85 %の正解を得ることを確認した．従って，本研究で作成した重要語は，実際の国際取引において作成される契約文書の条項の内容を推定するために，または，情報検索する際のキーワードとして有用であることが分かる．

(c) 自然言語においては，同じ意味を表すために，幾通りもの表現が可能である．そのため，類義語辞書 (thesaurus) は情報検索を行う際に重要なデータとなる．そこで，類義語辞書を作成する際の人的作業量を減らすために，関連度を用いた類似度計算の提案を行った．

関連度とは，条文を解析する際，複数の係り受けの中から妥当な係り受けを決定するために用いることを目的に提案したものである．

関連度を用いた名詞間の類似度は，(名詞，動詞)，(動詞，名詞)，(形容詞，名詞)，(前置詞，名詞) の二つ組の各々において，共通の組の関連度をベクトルとした内積の余弦を求め，全体の組に対する共通の組の比率を掛け合わせた後，重み付けを行い，求めた 4 種の値を加算したものである．

このようにして求めた語 x と y の類似度 (x,y) は，反射律と対称律を満たす．そこで，類似度を用いた相似関係行列を作成し，推移律を満たすファジイ同値関係を導き，全体集合を同値類に分類するプログラムを作成した．

そして，用語辞書に含まれる 894 種の名詞について，283 種の名詞からなる 99 個の同値類を求め，人手で意味を確認し，84 種の名詞からなる 34 個の類義クラスを求めた．

(3) 長文で複雑な契約書の条文を完全に統語解析することは困難である．本研究では，条文から統語構造を抽出する手法を提案し，試作プログラムを実装した．ここでの「統語構造の抽出」とは，一つの解析木を求める統語解析とは異なり，(1) 修飾－被修飾関係，および (2) 主部・述部を独立に抽出することをいう．本手法は，契約文書の条文に Brill's Rule Based Tagger を用いて品詞タグを付加した後，94 種類の規則を用いて意味的にまとまりのある語句に分割し，統語構造およ

び語義を注釈として付加した注釈付き解析データを作成する．本手法の特徴は，動詞を中心とするパターン情報を用いて解析する方法と，文法規則による解析の二つの手法を併用して解析を行っている点である．また，解析に使用する規則，データおよび解析結果はテキストデータであり，追加・修正することができる．

キーワード

英文契約書, TF.IDF 法, 構造化文書, 類似度, クラスタリング, 情報抽出

Studies on Structured Documents from Contracts in English*

Kaoru SAGARA

Abstract

Hundreds of overseas subsidiaries are newly established Japanese enterprises every year.

In these circumstances, English is a medium of communications even in Japan where English is not the official language. Inevitably, English contract documents (Contracts) are made voluminously in international transactions. In Britain and Germany, there is a report of the decentralized control system that aims to share Contracts scattered in the world. Moreover, the necessity of the knowledge acquisition from Contracts is reported. In addition, a report said that it is actually difficult to implement a system of machine translations of Contracts between the Japanese language and English language though the system is necessary.

The purpose of this thesis is to consider characteristics of Contracts and to show four proposals that aim at supporting information retrieval from a structured and digitized Contract (Digitized Contract) and aim at making the Digitized Contract.

I first introduce the base of a word (BSWORD). Secondly, I propose an extraction method for the important words that reflect the content of articles. Thirdly, I propose a calculation method of words' similarity from vectors consisting of co-occurrence statistics, and word clustering. Lastly, I propose a semi-automatic analytical method of a sentence in Contracts necessary to structurize them.

*Doctor's Thesis, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9661013, February 7, 2000.

(1) The BSWORD means the character string of the remainder removing the derivational suffix that changes the part of speech from a stem of a word. A "stem of a word" is the remainder removing the inflectional suffix from the targeted word. Use of the BSWORD reduces 5,804 variant word forms in the special terms' dictionary to 2,854 common BSWORD forms.

(2) A Contract is divided into various articles by their content. Therefore, by applying the TF.IDF method, I made an important words' list that reflects the content of 34 kinds of articles. The TF.IDF method is used often to extract the index word. Next, I did a presumption experiment of article's names with the list, by using Contracts included in books of eight publishers. By using the list of "raw words", the correct answer rate of presumption was 79 % (60 article's names out of 76 kinds are presumed correctly), and by using the BSWORD forms, that was 85 %.

From these results, the list is evidently effective as a words' list that shows the characteristic of articles. In addition, clearly BSWORD is more effective than "raw word" as the key word of information retrieval.

(3) There are some synonymous expressions in natural language, and for this reason, the thesaurus is effective to retrieve the associated information from sentences in Contracts.

In this proposal, the similarity (x,y) between word x and word y satisfies the reflexive law and the symmetric law. Therefore, the similarity relation matrix by using the words' similarity leads to a fuzzy equivalent relation that satisfies the transitive law. Words in Contracts can be classified into the equivalent classes by the fuzzy equivalent relation.

(4) A long and complex sentence is difficult to syntactically analyze in any parser.

In the proposal method, first, a digitized contract document is analyzed with Brill's Rule Based Tagger to annotate the part of speech. Next, the digitized document is divided into chunks that are significant word and phrases. Lastly,

the chunks with annotations of the syntaxtic structure and tags of the meaning of the chunks are made.

This method has the following characteristics.

(1) The method uses an analysis by grammatical rules together with an analysis by a template of a sequence of words.

(2) An analytical result is gained, even if the analysis is imperfect.

(3) Rules and data used for the analysis, and the analytical results are plain texts. They can be modified and corrected.

Keywords:

contracts in English, TF.IDF method, structured electronic documents, word similarity, clustering, information extraction

目 次

図目次

表目次

第1章
はじめに

本章では，研究の背景と研究のアプローチについて説明する．さらに，筆者の目標とする英文契約書の構造文書化についての構想を示す．

1.1 日本企業の海外進出と契約書

文献 [1][2] によれば，1997 年度に日本企業は 520 社の海外法人を新規に設立しており，これは前年度比 253 社減と 2 年連続減少しているものの，海外で経営活動を行う企業は増加している．また，これら海外で経営活動を行っている企業を対象とした「将来の経営計画」についてのアンケート調査では,「事業の多角化を図る」または「現在の事業領域で事業拡大を図る」と回答した企業が 63 %であり，現地法人企業の積極的な経営姿勢がうかがわれる．

表 1.1 海外子会社の数

多国籍企業	製造子会社	販売子会社	その他	合計
1974 年 (37 社)	310 社 (8 社)	105 社 (3 社)	—	—
1994 年 (149 社)	2,018 社 (14 社)	1,263 社 (8 社)	574 社	3,855 社 (26 社)

注：( ) 内は親会社 1 社あたりの平均海外子会社数.
その他の海外子会社には，統括または持株会社，研究開発会社などがある.
出所：文献 [3] pp.29 を一部変更の上引用.

表 1.1は，日本の多国籍企業の持つ海外子会社数の調査結果である．なお，ここでの多国籍企業とは，(1) 製造業である，(2) 大企業である，(3) 海外 5 箇所以上に製造子会社を持っている，の三つの要件を満たす企業をいう [3]．この表か

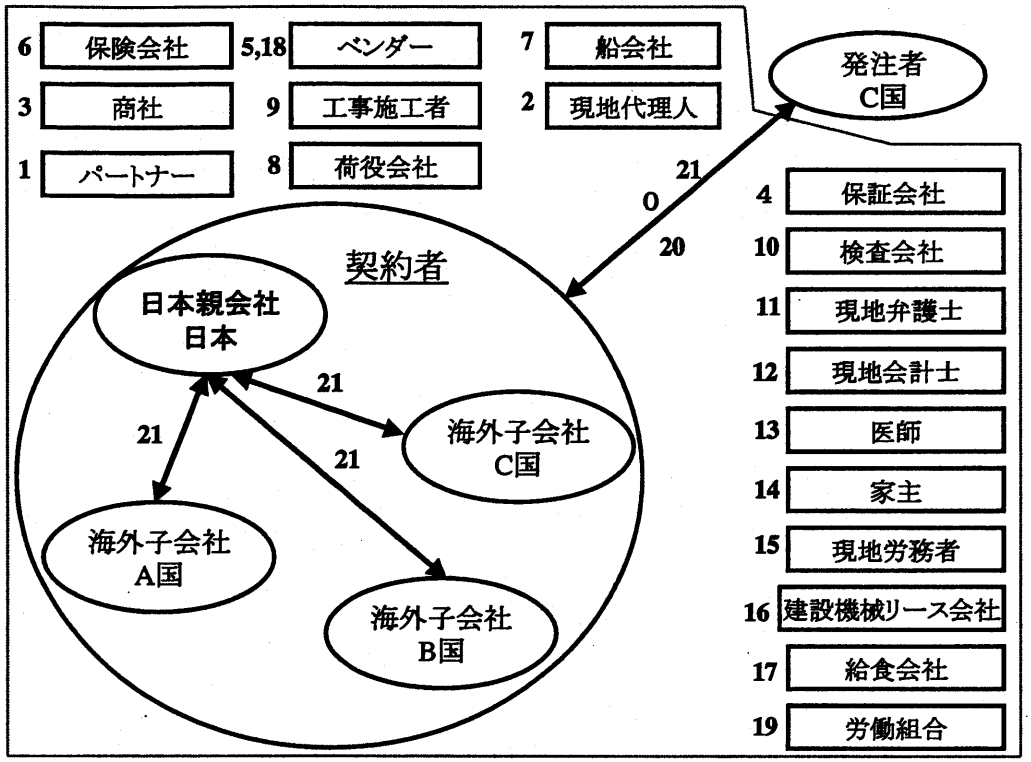

図 1.1 海外のプラント建設契約の履行に必要な諸契約

ら，1974 年から 1994 年の 20 年間で多国籍企業が 37 社から 149 社へと 4 倍に増加し，また親会社の持つ海外子会社の数も製造子会社では平均 8 社から 14 社に，販売子会社では 3 社から 8 社に増加していることがわかる．

親会社は，海外に製造子会社を設立し現地生産を行う．その際，親会社は海外子会社に技術を供与する．海外子会社は親会社から供与される技術について技術ライセンス契約を結び，供与された技術を使って生産を行う．

図 1.1と表 1.2は，海外でプラント建設契約を履行する際に派生する諸契約を示したものである．この例では，一つのプラント建設契約に派生して 21 種類の契約が履行されることがわかる．

すなわち，表 1.1と，図 1.1および表 1.2から，近年多国籍企業の増加に伴い，多くの契約書が作成されていることがわかる．

表 1.2 主契約と派生する諸契約

No.	契約の種類	相手方
0	主契約 (プラント建設契約)	発注者
1	consortium ／ joint venture 契約	パートナー
2	代理店契約	現地代理店
3	業務協定書	商社
4	保証受託契約書	銀行，保険会社
5	機材発注契約 (注文書)	機材供給者
6	保険契約	保険会社
7	運送契約	船会社
8	荷役契約	荷役会社
9	下請会社	工事施行会社
10	検査契約	検査会社
11	リーガル・サービス契約	弁護士
12	アカウンティング・サービス契約	会計士
13	医療サービス契約	医師
14	家屋賃貸借契約	家主
15	雇用契約	労務者，従業員
16	建設機械リース契約	建設機械保有会社
17	給食契約	給食会社
18	vendor supervisor の派遣契約	機材製作者
19	労働協約	労働組合
20	主契約内容の変更	注文者
21	ライセンス契約	ライセンサー

注：表 1.2の番号は図 1.1の番号と対応している．

出所：文献 [4] pp.38 を一部変更の上引用．

表 1.3 海外子会社の基本共通言語

共通言語	アメリカ	イギリス	ドイツ	シンガポール	台湾
英語	327 社	85	46	65	3
日本語，英語，現地語	—	—	3	1	24
日本語と英語	10 社	0	0	5	1
日本語	0 社	0	0	1	16
現地語	—	—	10	0	3
合計	337 社	85	59	72	47

出所：文献 [3] pp.225 を一部変更の上引用.

次に，表 1.3は海外子会社において，日本人と現地人が意思の疎通を図る上で使われる言語について調査したものである．英語を母国語としない，ドイツ，シンガポール，台湾においても，基本共通言語に英語が含まれる割合がドイツ 83 % (49 ／ 59)，シンガポール 99 % (71 ／ 72)，台湾 60 % (28 ／ 47) と半数を越えていることがわかる．このことから，国際取引契約の締結により作成される英文契約文書の割合が高いであろうことが推測される．

1.2 契約文書の重要性

契約書は，契約当事者双方の意図を正確・明確に記録し，当事者間の紛争や誤解を防ぎ，取引目的の達成のために作成される．とくに国際取引契約においては，政府や関係当局の許可の取得や届出を得るために，または銀行や取引先の支援・協力を得るために，契約書が必要となる．また，会社の経理・財務上においても契約書は必要である [13][18]．このように，契約書には取引契約に関する重要な情報が含まれている．従って，契約書は重要文書として軽易なもので 3 年，重要なものは永久保存される．

企業活動の中で作成される文書には，定式化され，記入項目があらかじめ決められている帳票と，社交文書，連絡文書，法令，社内規定文書，契約書などの自

表 1.4 企業におけるコンピュータの利用分野 (複数回答) と処理形態

利用分野	部門	業務	割合%	集中	分散	なし	不明
財務・会計管理	経理	定型	68.6	83.3	11.7	2.7	2.2
給与管理	人事・労務	定型	68.6	76.1	20.2	2.2	1.4
販売管理	営業	定型	64.9	72.4	13	12.4	2.3
仕入管理	仕入・購買	定型	52.2	69.3	9.9	16.4	4.5
受発注管理	仕入・購買，営業	定型	48.0	—	—	—	—
在庫管理	仕入・購買	定型	43.6	65.9	13.1	16.5	4.5
原価管理	製造	定型	40.1	61.6	12.7	19.8	5.9
生産・工程管理	製造	定型	37.6	44.2	15.9	31.6	8.2
経営計画・企画・調査	総括	非定型	25.2	31.2	42.9	17.2	8.6
品質管理	製造	定型	12.9	23.8	26.9	37.2	12.1
物流管理	仕入・購買，営業	定型	11.6	—	—	—	—
研究開発	技術開発	非定型	10.6	6.9	47.7	33.8	11.6
情報伝達	全部門	非定型	8.4	8.9	61.2	18.5	11.6
その他			6.4	—	—	—	—

出所：文献 [5] の図表 3-1-47 および図表 3-1-13 と，文献 [6]pp.95 を基に作成

資料 1：中小企業事業団「情報化推進アドバイス研究事業報告書」1995 年度

資料 2：JISA ユーザーアンケート調査 1996 年度

注："—" は，データのないことを示す.

然言語で書かれた一般文書がある．帳票は定型業務の中で作成され，入力項目が決っていることから，情報化が進められている．

表1.4は，企業におけるコンピュータの利用分野と処理形態についてアンケート調査を行った結果である[5][6]．定型業務については，集中管理が容易であり，その結果，コンピュータの利用が進んでいるのに対し，非定型業務については，集中管理が困難であり，コンピュータの利用が低いことがわかる．すなわち帳票管理については，情報化が定着していることがわかる．また，生産・工程管理，品質管理，研究開発については，集中管理と分散管理の両処理に該当しない割合が高い．これは，人的資源に頼る部分が多く，コンピュータを使って管理することが困難であることを示している．なお，経営計画・企画・調査については，分散管理の割合が高く，コンピュータを利用する割合が25.2％と高くはない．しかしながら，集中管理にも分散管理にも該当しない処理も，17.2％と低い．このことは，分散している情報を共有化できれば，コンピュータを利用する割合が増すことを示唆している．そして，経営計画・企画・調査において重要な文書は契約書であり，また，そこに含まれる情報は，経理，営業，仕入・購買など複数の分野にとっても重要である．

一方，一般文書のファイリングシステム(文書管理)については，(1)紙による管理，(2)カセットファイルやマイクロフィルムによる管理，(3)電子文書化による管理がある．現在のファイリングコンサルティングが，文書分類の基準の決定，重複保管書類の削減を主たる目的としていることから，実際には，電子文書化による管理の割合は低いと思われる．なお，ファイリングの目的は，(1)保管場所の節約，(2)原本保存，(3)情報の共有化，(4)検索の容易化にある．

一般文書の電子化を阻む理由としては，スキャナーの精度などの技術的な問題に加え，非定型業務であることから，検索頻度が少なく，また関連部署も限られているため，投資に見合った効果が得られないという問題がある．さらに一般文書には重要文書が含まれるため，機密保持の問題も電子化を阻む要因となる．

表1.5は，INSPEC(Information Services for Physics, Electronics and Computing)データベースで以下のキーワードを用いて文献検索を行った結果22件の概要である．

表 1.5 INSPECデータベース検索結果

	内　容	件数
a	光ディスクと関係データベースを用いた契約文書のイメージ圧縮格納システム	1
b	契約書類の分散管理による情報の共有化	2
c	知識の獲得をするための工事契約文書の形式化の重要性	1
d	契約文書の分析の重要性	1
e	契約文書の重要性および完全な契約書の草案	8
f	契約書類，特許の書類の日英機械翻訳の要求	1
g	契約書類に関係のない論文	8

＜キーワード＞

(1) contract note

(2) English contracts

(3) contract document

(4) written contract

(5) contract sheets

(6) deed and contract

表1.5において，実装されているものはａとｂに関する３件であり，その他は提案である．この表から，契約書は重要な経営資源となることが推測される．また，契約書の分析，知識の獲得の重要性についての議論がなされ，契約書に含まれる情報の共有化の試みがなされているものの，契約書の条文を解析する具体的な手法についての研究がほとんど行われていないことがわかる．

1.3　電子化に向けての研究アプローチ

本節では，英文契約書 (以下, 契約文書という) の電子化について述べる．従来の紙面による文書管理では，文書名，作成部署・作成者，作成日付を索引として管理を行う．パーソナルコンピュータを使った文書管理では，索引部分と格納場

所をデータベース化し，原本は紙面で保存している．原本も電子文書化することを前提にした文書管理については，以下の形態が考えられる．

(1) 索引語 + 電子文書

(a) 索引語 = 文書名，作成部署・作成者，作成日付

(b) 索引語 = (a) + 文書の内容を表す重要語

(c) 索引語 = (a) + 文書の要約

(2) 構造化文書 = タグ付き電子文書

(a) は，従来の紙面による文書管理を電子化したものである．

(b) の実装には，内容を表す重要語を求める必要があり，その一つの手法として TF.IDF 法がある [8].

(c) または (2) の実装には，情報を抽出する必要がある．文書から情報を抽出するためには，(i) 全文を文法規則を使って統語解析した後，意味解析を行う手法と，(ii) テンプレートを使って抽出する手法が考えられる．そして，これら自然言語を対象とした統語解析，情報抽出に関する研究も進められている．

しかしながら，契約文書における条文は，修飾節および句の挿入が多く，長文で複雑である．従って，これらの条文を正確に統語解析することは困難であるし，また，テンプレートを用いた情報の抽出においては，網羅性の高いテンプレートを作成することは困難である．

なお，SGML(Standard Generalized Markup Language)，XML(eXtensible Markup Language) などの出現で，契約文書の解析ができれば，これらを用いた構造化文書が可能となる．

以上のことから上記契約文書の電子化は，(a) → (b) → (2) → (c) の順位で実装の難易度が増すと考えられる．従って，契約文書の電子化に向けての適切な研究アプローチは以下のようになる．

(1) 契約文書をコンピュータ処理するための専門用語などの辞書類の作成と，契約書の統語的，内容的な特徴分析 (本章，表 1.6の参照辞書・データ，第 2 章)

(2) 情報検索を支援するシソーラスの作成,「語の基本形」の提案 (第 3 章，第 5 章)

(3) 索引語の作成および情報検索を支援する契約文書の内容を表す重要語一覧の作成 (第 4 章)

(4) 文書の要約および構造化を支援する契約文書の解析手法の提案 (第 6 章)

1.4 本研究の全体構想

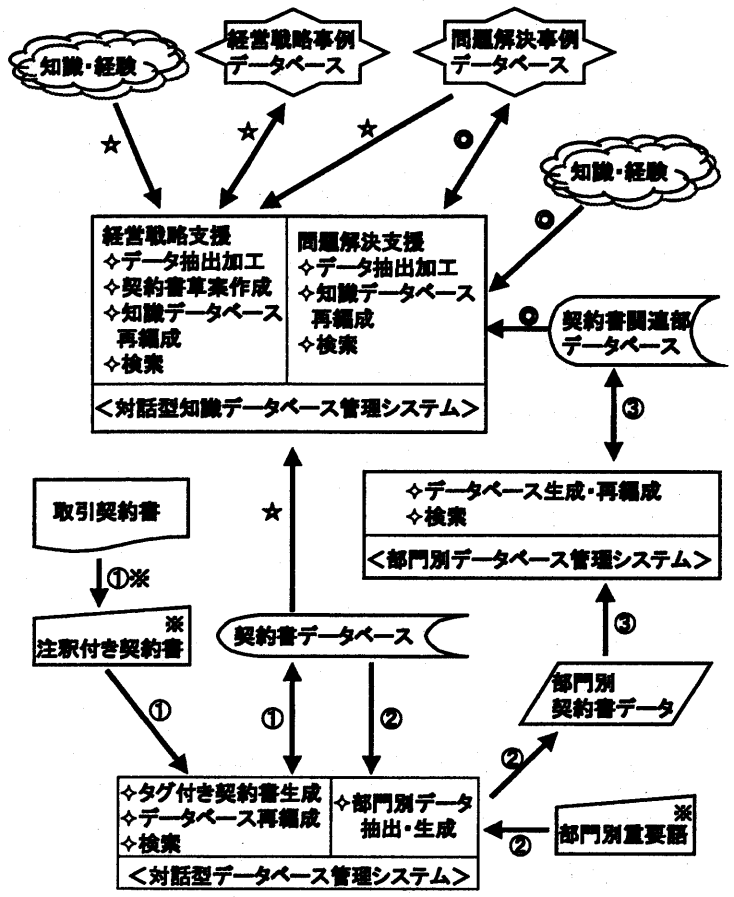

図 1.2 企業で発生する契約文書の構造化データベース例

図 1.2は，本研究で想定している契約文書のデータベースの構成図である．※印部分，すなわち「注釈付き契約書の作成」と「部門別重要語の作成」が本研究で扱う部分であり，その詳細を含めた本研究の全体構想を表 1.6に示す．なお，図 1.2の「部門別重要語」は表 1.6では「重要語一覧」と表現している．

表 1.6 本研究の構想

入出力データ	処　理	参照規則	参照辞書・データ
英文契約書	1. 品詞タグ付け	タグ付け規則†	品詞辞書†
	2. 登録処理		
↓			
品詞付き契約書	3. 品詞タグの修正	(a) 品詞タグ修正規則	
	4. 複合語・関連語句の連結	(b) 連結語句のパターン	
	5. 構成語句分割と分類タグ付け (第 6 章)	(c) 分割規則	(1) 注釈変換データ
↓			
注釈付き契約書	6. 構成語句間の関連付けと注釈の付加 (第 6 章)	(d) 語句間の関連付け規則	(2) 動詞と前置詞の関連データ (3) 語の等位関係データ (4) 動詞・主語・述部のパターン (文献 [22])
	7. 構造化 (XML または SGML 利用)	(e) タグ付け規則	
↓			
構造化契約書データベース	8. 情報検索		(5) 用語辞書 (6) 類義語辞書 (第 5 章) (7) 重要語一覧 (第 3，4 章)
	9. 特徴抽出		

注 1：本研究では，Brill's Rule Based Tagger を利用して品詞タグ付けを行う．

従って，タグ付け規則および品詞辞書は付随のものを利用する．

但し，品詞辞書については，(5) 用語辞書と照合し，未登録語 552 種を追加している．

注 2："→" はデータの参照を，"←" はデータの更新を意味する．

本研究で作成し使用するデータおよび規則は以下のとおりである．実例については付録 A および付録 B を参照のこと．

＜参照辞書・データ＞

(1) 注釈変換データ：

前置詞（群前置詞），名詞（複合名詞）計 681 セット

条文を分割して求めた意味のある語のまとまり (本論文では構成語句という) の意味を表す注釈データ.

(2) 動詞と前置詞の関連データ： 809 セット

構成語句間の関係を求める際の参照データ.

書式集から，語 x ∈動詞，語 y ∈前置詞とするとき，二つ組 (x,y) の出現頻度と，(x, 前置詞) および (動詞,y) の頻度を基に自動生成したものであり，不適切なものが含まれる.

(3) 語（句）の等位関係データ：計 70 セット

構成語句間の等位関係を調べるための参照データ.

(4) 動詞・主語・述部のパターン：計 146 パターン

書式集の秘密保持条項 [7] に含まれる 95 条文から抽出した権利・義務の記述に関する 57 種の動詞を含む 146 種のパターン [22].

(5) 用語辞書：計 5,806 語

書式集 [7] に出現する用語の辞書.

(6) 類義語辞書：計 34 セット

類義語辞書の作成については，第 5 章で詳述する.

(7) 重要語一覧：34 条項について計 637 語

構造化契約書データベースから特徴を抽出する，または抽出した内容の重要度を調べる際に参照するリスト.

技術取引契約書で使われる 34 種類の条項毎について，重要語句をまとめたもの．重要度は，TF・IDF を応用して数値化している．詳細は，第 4 章で述べる.

＜参照規則＞

(a) 品詞タグ修正規則 ：計 183 パターン

文献 [9] によれば，品詞辞書に未定義語が存在しない場合，95 %の正しさで品詞付けされる．本研究では，前述 (5) の用語辞書と品詞付けツールに含まれる品詞辞書との併合を行っており，未定義語はない．従って，書式集 [7] に含まれる約 20 万語のテキストデータの内，不適切な品詞タグが付加された語が約 1 万語あると推測される．本研究では，不適切なタグが発見される毎に手作業で修正規則 (Perl の正規表現) の追加，更新を行う．

(b) 連結語句パターン ：計 403 パターン

文献 [10] ～ [21] より契約文書特有の，専門用語 937 種および冗長な表現 108 種の一覧を求めている．連結語句パターンは，これらの中から，複合語，群動詞，群前置詞，等位関係の語句を抽出し，Perl の正規表現による連結規則を自動生成したものと，受動態および完了形の動詞を一つにまとめる規則，並びに本研究中に気が付いて追加したものからなる．

(c) 構成語句分割規則：計 94 種類

助動詞および前置詞の前後，代名詞および冠詞の後ろ，＜形容詞＋前置詞＞の前，名詞および動詞の間，関係代名詞および関係副詞の前後，特殊記号の前後，等位接続詞の前後を分割する規則を経験則により人手で作成している．第 6 章で詳述する．

(d) 構成語句間の関連付けプログラム：計 7 種類

(i) 構成語句間の等位関係を求めるプログラム

(ii) 形容詞と構成語句の関係を求めるプログラム

(iii) 前置詞と構成語句の関係を求めるプログラム

(iv) 相関接続の関係を求めるプログラム

(v) 前置詞を含む構成語句と被修飾構成語句の関係を求めるプログラム

(vi) 述部をまとめるプログラム

(vii) 単文を求めるプログラム

第6章で詳述する.

(e) SGML, XML または GDA などのタグ付け規則（今後の検討項目）

1.5 本論文の概要

本論文は以下のとおり構成される.

第2章 契約文書の特徴

第3章 語の基本形

第4章 語句の重要度計算と評価

第5章 語句の類似度計算とクラスタリング

第6章 条文の統語構造の抽出

第7章 本論文のまとめ

第2章では，英文契約書の特徴について述べる．第3章では，構文が異なる類義の条文に柔軟に対応するために提案した「語の基本形」について述べる．次に第4章では，TF・IDF を利用した条項における語の重要度計算について説明し，技術取引に関する 34 条項について求めた重要語一覧 (表 1.6の (7)) を使って，本研究で提案する「語の基本形」と「重要度」の有効性を明らかにする．第5章では類義語辞書 (表 1.6の (6)) を作成支援するために考案した類似度計算とクラスタリングについて述べる．以上第3章～5章で述べる内容は，電子文書化を行う際の索引語および情報検索の際のキーワードに関するものである．第6章では，契約文書の構造化に関して提案した,「注釈付き契約文書」の作成 (表 1.6の処理 3 ～ 6) について述べる．そして第7章でまとめる.

第 2 章
契約文書の特徴

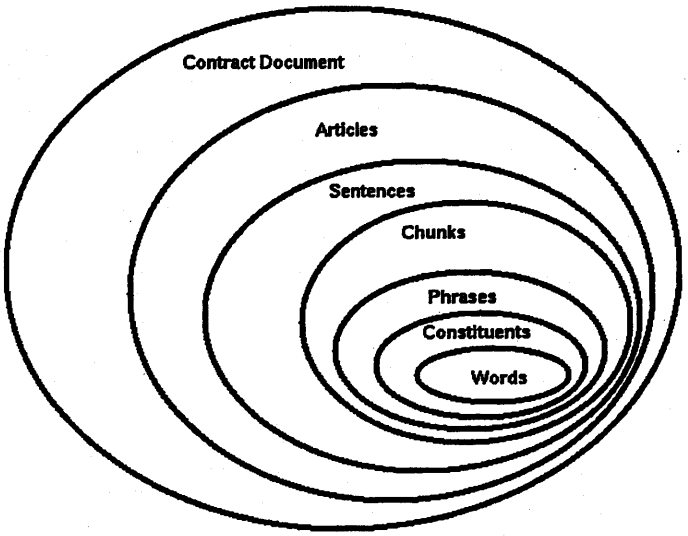

図 2.1 契約文書の構造

2.1 条文の特徴

契約文書に含まれる条文は，以下の特徴を持つ．

(1) 条文は長文である．

(2) 条文は多くの修飾語句，または修飾節を含み複雑である．

(3) 契約文書に出現する語句は，限られている．

表 2.1 文の比較

	契約文書	W.S.J
サンプル数	4,677 文	4,678 文
(1) 平均語数／文	50 語	21 語
(2) 動詞の数 (現在形, 分詞形, 過去形) ／文	6 語	3 語
サンプル数	162,298 語	163,330 語
(3) 単語の種類 (文字列の異なる語)	5,784 種	11,611 種

これら (1) ～ (3) の特徴は，表 2.1の *Wall Street Journal* (WSJ) の品詞タグ付きコーパスからランダムに抽出した文と本研究で用いる書式集 [7] との比較結果より明らかである．なお，英文を対象とした統語解析において，WSJ のコーパスが使われる場合が多い．しかしながら，WSJ に含まれる全てのデータを正確に適正な時間で統語解析できる統語解析ツールは見当たらない．

他の特徴として，契約文書に含まれる条文は，平叙文のみであり，疑問文や命令文，そして感嘆文などは含まれない．

2.2 条項の特徴

一般に契約書は，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．表 2.2は，書式集 [7] に含まれるライセンス契約書の論理的構造を示したものである．この例において，ライセンス契約に関する条項 2 ～ 23 は，内容の上で四つのグループにまとめられる．

「条項ごとに内容がまとめられている」という特徴を活かして，書式集 [7] に含まれる 34 種類の条項ごとに，TF.IDF 法を応用した重要語一覧を作成し，書式集 [7] を含む 8 社から出版されている文献 [10] ～ [16] から抽出した例文を対象に，その条項名の推定を行い，約 80 %の正しさで条項名を推定できることを確認している．詳細は第 4 章で述べている．

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 (表 2.2の † 印がある条項) と，(2) 契約書固有の条項に大別される．以下に，それぞ

表 2.2 ライセンス契約書の条項例

表題†	Know-How License Agreement:ノウハウ・ライセンス契約書
前文†	WITNESSETH (頭書，前文)
Article 1†	Definitions (定義)
♡ Article 2	Grant of License (実施権の付与)
♡ Article 3	Disclose of Know-how (ノウハウの開示)
♡ Article 4	Company's undertakings (会社の約束事項)
♡ Article 5	Parts etc. (部品等)
♡ Article 6	Technical Guidance (技術指導)
♡ Article 7	Dispatch of Trainees (訓練生の派遣)
♡ Article 8	Technical License Fee (技術実施料)
♣ Article 9	Payment (支払い)
♣ Article 10	Access to Records (記録の閲覧)
♣ Article 11	Charges (課徴金)
◇ Article 12	Industrial Property Right (工業所有権)
◇ Article 13	Improvements made by Company (会社による改良)
◇ Article 14	Use of Trademarks (商標の使用)
◇ Article 15	Warranty (保証)
◇ Article 16	Secrecy (秘密)
◇ Article 17†	Term (期間)
◇ Article 18†	Termination (契約解除)
◇ Article 19	Step after Termination (契約終了後の措置)
♠ Article 20†	Arbitration (仲裁)
♠ Article 21†	Applicable Law (適用法)
♠ Article 22†	Language (言語)
♠ Article 23†	Headings (表題)
末尾文言†	IN WITNESS WHEREOF (後文)
署名欄†	

† ： 契約書の種類に関係なく記載される項目
♡ ： ライセンス契約の主題
♣ ： 支払いに関する条項
◇ ： ライセンスの付帯条件
♠ ： 付随的事項

れの特徴と，特徴から推測されることを列挙する．

(1) 一般条項

- 文例が書式集としてまとめられているように，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．
  - 内容の定式化がしやすい．
  - 出現する語句がある程度限定されている．
  - 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 (局所性が高い)．

(2) 契約書固有条項

- 各条項間が密接に関連している．
- 契約内容および，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である．
  - 内容の定式化が困難である．
  - 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．

固有条項より一般条項の方が定式化されていることは，第 4 章の表 4.3の番号 6 と番号 7 において，一般条項における条項名の推定結果の正解率が，固有条項のそれと比べて高いことから結論できる．

一般に，情報抽出において固有名詞（社名，住所，製品名など）の抽出が大きな課題となる．契約文書においては，契約締結の日付けや，当事者に関する情報（社名，住所，商号，設立準拠法）は「前文」にまとめて記載され，契約文書中で使われる契約用語や製品名等は「定義条項」で定義されており，ある程度定式化がなされている．本研究では，これら前文および定義条項を解析の対象から外している．

2.3 契約書の内容

契約書に記載されている内容は，以下の四つに大別される．

(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

一般に契約文書は能動態で表現することが推奨される．その場合,「人に，権利・義務が，生じる」,「人が，行為を，する」などの「主語，動詞，目的格 (S,V,O)」に,「何時」,「何処で」,「何のために」などの各種限定条件を表す修飾語句または修飾節 (以下，これらをまとめて "**Modifier**" という) が付加された,「S + V + O + Modifier」の構文が多く存在する．

能動態が推奨されるのは，受動態には以下のような欠点があるためである [18].

(1) 能動態に比較して文が長くなる．

(2) 意味の明瞭さを欠く場合が多い．

一方，次のような場合は，受動態が用いられる [18][19][23].

(1) 行為者の結果が重要であって，行為者は重要ではなく省略しても差し支えない場合

The summons and complaint were served on January 19th.

(1 月 19 日に召喚状と告訴状を受け取った)

(2) 行為者が不明または特定できない場合

The ledgers were stolen.

(元帳が盗まれた)

(3) 主語が長く，かつ動詞の部分を強調したい場合

There shall be collected a penalty in the amount of $ 100 for each violation.

(各々の違反につき $ 100 の違約金が課される)

(4) 行為者など強調すべき要素 ("the laws of the State of California") を文尾に持ってくる場合

(注：一般に英語では文尾に新しい情報や重要な情報が盛り込まれる)

This Agreement shall be governed by the laws of the State of California.

(本契約は，カリフォルニア州法に準拠する)

(5) 事故・病気などの表現など英語特有の受動表現

The 10:20 train was derailed, and a lot of people were injured.

(10 時 20 分発の列車が脱線し，多くの負傷者が出た)

文献 [10] ～ [14]，[17] ～ [21] によれば，原則として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，現実には，誤用される場合もあり，"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

加えて，合意事項として当事者間の権利または義務を記述する場合があり，この場合は以下の語句によって識別することができる．

(1) agree (to/that)
(2) undertake (to/that)

なお，書式集 [7] の内，34 種類の条項における 2,315 条文の中で，前述の権利・義務を示唆する語句を含む条文は 1,704 条文ある．

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

そこで本研究では，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を，権利または義務の記載とみなす．

第 3 章
語の基本形

本章では，本研究で導入する「語の基本形」について述べる．

自然言語において，一つの意味を表すのに幾通りもの表現が可能である．そこで一般に情報検索を行う際，これらの同義文に柔軟に対応するために「語幹」などの「抽象化された語」を導入する．

表 3.1は，同義文を分析したものである．同義文には，ある語と同じ品詞の同義語句で語を置き換えた記述 (表 3.1 (1)，文例 1) や，語根に接辞が付加された品詞の異なる同義語を用いた記述 (表 3.1 (4)，文例 4) などが考えられる．

表 3.1 同じ意味を表す文

番号	使用語	語根	品詞	接辞	構文
(1)	同義語句	異	同		同
(2)		異	異		異
(3)	語形変化	同	同 (動詞)	屈折	異
(4)		同	異	派生	異

注：番号は以下の文例の番号に対応している

＜文例＞

いずれの当事者も本契約を譲渡してはならない．

Either party shall not **assign** this Agreement.

(1) Either party shall not **convey** this Agreement.

(2) This Agreement shall not **be conveyed** by either party.

(3) This Agreement shall not **be assigned** by either party.

(4) This Agreement shall not **be assignable** by either party.

No **assignment** of this Agreement shall be made by either party.

3.1 語句の定義

はじめに本章で扱う語句の定義を行う [24][25].

語根 (root)

語からすべての屈折接辞と派生接辞を取り除いたもの

語根 "**assign**" = 語 "**assign-or-s**" − 派生接辞 "**-or**" − 屈折接辞 "**-s**"

語幹 (stem)

語から屈折接辞を取り除いたもの

語幹 "**assignor**" = 語 "**assign-or-s**" − 屈折接辞 "**-s**"

基体 (base)

屈折または派生接辞を受ける形式

語 "**assign-or-s**" において，屈折接辞 "**-s**" が付加される基体は，"**assignor**" であり，語 "**assign-or**" において，派生接辞 "**-or**" が付加される基体は，"**assign**" である.

基本形 (BSWORD)

語から屈折接辞と，品詞を変える派生接尾辞のみ (接頭辞は含まない) を取り除いた基体を「基本形」いう.

語根，語幹，基本形の関係 (文字列の長さ) は次の通りである.

語幹 ⊆ 基本形 ⊆ 語根

語の構造　　語の構造は以下の 7 式で表される.

型	形態	具体例
a. X → X Y^{af}	屈折	(assign-ed)
b. X → Y X^{af}	接尾辞付加	(assign-ment)
c. X → Y^{af} X	接頭辞付加	(un-faithful)
d. X → X^{af} Y	特別な接頭辞付加	(en-able)
e. X → Y X	複合	(assignment clause)
f. X → X Y	特別な複合	(parties hereto)
g. X → Y	転換	(assign 譲渡する：譲受人)

X, Y ：名詞，形容詞，動詞などの 品詞を表す包括記号

X^{af}, Y^{af} ：接辞

前述の 7 式中，接辞が語の品詞 X を決定するものは，b 式の接尾辞 X^{af} と d 式の特別な接頭辞 X^{af} のみである．また，品詞を決定する特別な接頭辞は，**en-**，**a-**，**be-**，**de-**，**dis-**，**out-**，**un-**などである [29].

接辞の分類

接辞：(1) 屈折接辞：構文上での文法関係を示すもの
(2) 派生接辞：a) 接頭辞 (53 種類)：i) 品詞を変えるもの (6 種類)
ii) 意味を変えるもの (53 種類)
b) 接尾辞 (97 種類)：i) 品詞を変えるもの (85 種類)
ii) 意味を変えるもの (41 種類)

注：( ) 内は，文献 [29] より抽出したもの

再現率 (recall)

$$\text{再現率} = \frac{\text{検索された該当文書数}}{\text{全文書中の該当文書数}}$$

適合率 (precision)

$$\text{適合率} = \frac{\text{検索された該当文書数}}{\text{検索された文書数}}$$

3.2 語幹または語根を求めるアルゴリズム

本研究と関連するものに英語のキーワード抽出に用いられる “Stemming Algorithms” がある．“Stemming Algorithms” とは，語幹または語根を求めるアルゴリズムをいう．

一般に英語における情報検索では，語幹が用いられる．“Stemming” は，検索効率を高めるためと索引ファイルの規模を節約するために用いられる．

なお，検索効率を評価する指標として「再現率」と「適合率」が使われる．過不足なく検索結果が得られた場合，再現率と適合率は共に 1 となる．

文献 [26] では “Stemming” の手法として，(1)Affix Removal，(2)Successor Variety，(3)n-gram が紹介されている．なお，(2) は語幹を生成するために大規模なコーパスを必要とする．そして (3) は，正確には語幹を生成するアルゴリズムではない．

(1) の実装方法としては，語と語幹の対応表を用いる方法と，規則を用いて語幹を求める方法がある．文献 [26] で取り上げられている “The Porter algorithm(1980)” では，3 種の屈折接辞と 22 種の接尾辞についての 62 の規則を用いて語幹 (weak stemming) と語根 (strong stemming) を求めている．

そして，同文献に記載されている評価実験の結果では，“The Porter algorithm” による “weak stemming” は，再現率を著しく増加させ，かつ適合率の著しい減少は見られないのに対し，“strong stemming” は，適合率の著しい減少が見られ，検索効率は，weak stemmed > unstemmed > strong stemmed の順であるとの報告がある．

そして，Java 言語で作成された “The Porter algorithm” によるツールが Web 上で公開されている (文献 [27] より取得可能)．この Java 版の Stemming プログラミングは，数，程度，大きさなどを表す接頭辞 (“kilo”,“micro”,“milli”,“intra” ,“ultra” ,“mega”,“nano”,“pico”,“pseudo”) を削除する機能を持つ．なお，文献 [29] には，53 種類の接頭辞が記載されており，これらは全て基の語の意味を変えるものである．

3.3 語の抽象化の階層

本研究では，以下の 3 種の抽象化クラスを導入し，用途に応じてこれらのクラスを使い分ける．

階層 <抽象クラス>	<付加される要素>	<自動作成>
(1) 語幹クラス		
同じ語幹の集合	屈折接辞	+++
(2) 基本形クラス		
同じ基本形の集合	品詞を変える派生接辞	++
(3) シソーラス		
同じ意味の集合	語の意味	+

語の抽象化を考える際，複数の語の意味が類似している同義性とは対照的に，一つの語が複数の意味を持つ多義性の問題がある．語とその意味が1対1に対応していれば，上位にある抽象化クラスを用いる程，複数の同義の文書処理に柔軟に適応できることがわかる．しかしながら，実際には語とその意味の関係は，1対1とは限らない．

表3.2は，書式集[7]に含まれる4,519条文(162,298語)に出現する「遂行」という意味を持つ語の抽象化クラスの階層を示したものである．同義語のクラスの中には，契約書に使われる意味に限定したとしてもなお複数の意味を持つものがある．例えば，語“discharge”は,「遂行」という意味以外に「負債などの弁済」や「契約などの取消し」などの意味がある．このように，シソーラスのクラスでは，複数のクラスに同じ語が含まれる場合が生じる．これは，適合率を減少させる原因となる．したがって，上位の抽象クラスが文書処理において有効であるとは限らない．加えて，意味による分類を機械的に行うこと，すなわち自動生成することは困難である．

また，3.2節のとおり語根による検索においても適合率を著しく減少させるとの報告がある．そこで，本研究では，条文に含まれる語の意味を推測するために「基本形クラス」という抽象化クラスを導入する．3.1節にあるように「基本形」は，語から品詞を変える接尾辞を取り除いたものである．従って,「基本形クラス」は同じ意味と語根を持つ語の集合となり，複数の基本形クラスを用いることで，シソーラスの代用が可能となる．なお，接頭辞を削除の対象から外したのは，接頭辞のほとんどが語の意味を変えるものである(3.1節，語の構造c式)という理由からである．

このような基本形クラスにおいても多義性の問題が生じる場合がある．例えば，“-ive”は名詞を形容詞にする，または動詞を形容詞にする接尾辞であるが，“executive”には「遂行上の」という形容詞の意味の他に「重役」などの,「遂行する」という意味とは直接関係のない意味もある．

しかしながら,「同じ語根を持つ」という制約により，多義性の問題はシソーラスに比べて低減される．

表 3.2 「遂行」という意味を含む語の例

シソーラス	基本形	語幹	語	品詞
EXECUTE	execute	execute	execute	vt.
			executed	pp.pt.
(代表語)		execution	execution	n.
		executive	executive	a.n.
			executives	pl.
		executory	executory	a.
	fulfill	fulfill	fulfill	vt.
			fulfil	vt.
		fulfillment	fulfillment	n.
			fulfillments	pl.
			fulfilment	n.
			fulfilments	pl.
	その他の同義語 (基本形と句)： carry out, enforce, *discharge*, perform, administer, accomplish, achieve, complete, effectuate, consummate, realize			

vt.：他動詞，　pp.：過去分詞，　pt.：過去形
n.：名詞，　pl.：複数形，　a.：形容詞

3.4 「語の基本形」の作成

既に，技術取引に関する契約書式集 [7] に含まれる 4,519 条文 162,298 語から，5,804 種の単語について，品詞と語幹 (3,601 種) を求めるための辞書を作成している．

また，シソーラスの作成については，分類に掛かる人手を減らすために単語 (語幹) の共起頻度に基づく類似度計算とファジィ関係行列の推移的閉包によるクラスタリングの手法の提案を行っている．これらの詳細は第 5 章で述べている．

基本形クラスについては，文献 [29] に記載の 97 種類の接尾辞から品詞を変える 85 種類を抽出し，形容詞→名詞 (78 種)，形容詞→動詞 (1 種)，動詞→名詞 (19 種)，名詞→形容詞 (26 種)，動詞→形容詞 (24 種)，名詞→動詞 (114 種)，合計 262 種類の語形変換テーブルを作成している．そこで用語辞書 5,804 語から，固有名詞と，"-" で連結された合成語を除いた 5,550 語の語幹 2,984 語を対象に，語形変換テーブルの内，変換後の語長が基の語よりも短くなるものを使って基本形を求め，人手により見直しを行い，5,804 語を 2,854 の基本形クラスに分類している．人手により見直しする際，例えば "inform" でまとめられている "information" や "informal" については別の基本形クラスにするなどの意味的な分類処理を行っている．

表 3.3は，品詞を変える 85 種類の接尾辞による語形変換テーブルの一部である．

表 3.3 品詞を変える語形成変換テーブル

形容詞 → 名詞		動詞 → 名詞		動詞 → 形容詞		名詞 → 動詞	
like		ative	ate	en		able	
able		ciate	ce	ify		ation	
al		itate	ity	ivate	ive	ductory	duce
cious	cy	itify	ty	lsify	lse	ency	
getic	gy	otify	ote	stify	stic	gnition	gnize
icable	y	tirize	tire	tten	t	position	pose

なお，接尾辞による語形成については，(1) 接尾辞付加の順序，(2) 付加できる接尾辞の数，ならびに (3) 疑似接尾辞の存在などの規則がある [29].

3.5 Java 版 Porter Stemming と本手法の比較

表 3.4は，公開されている Java 版の Porter Stemming プログラム [27] と本手法の処理の違いをまとめたものである．また，表 3.5は，本研究で作成した用語辞書に含まれる単語を入力とした，処理結果の比較である．

表 3.4 Porter Stemming と本手法との処理の比較

	Porter	文献 [29]	本手法	備 考
出力	語根		基本形	Porter：全ての接辞の削除を目標.
入力	単語		語幹	本手法では屈折接辞の削除は行わない.
屈折接辞	△		−	△：処理が完全ではないことを意味する.
接頭辞	9 種	(全)53 種	−	−：処理しないことを意味する.
接尾辞	(全)23 種	(全)97 種 (品詞)85 種 (意味)41 種	(品詞)71 種	本手法では，意味を変えずに品詞のみを変える接尾辞 68 種類に “ly”,“ish”,“age” を加えた 71 種類を対象にした 254 種類の規則を利用. 85 種類による規則は 262 種類.
規則	(全)51 種	(品詞)262 種	(品詞)254 種	

表 3.5 Porter Stemming と本手法との処理結果

	入力語数	処理件数	正解数
Porter	(raw)5,550 語	1,990 語 (36 %)	341 語 (17 %)
	(stem)2,984 語	1,846 語 (62 %)	256 語 (14 %)
本手法	(stem)2,984 語	2,706 語 (91 %)	2,046 語 (76 %)

処理件数：入力文字列 ≠ 出力文字列であるものを処理件数とする.
正解数：語幹の集合 (66,083 語) に出力文字列が含まれる場合に正解とする.
語幹の集合は，Brill Tagger で利用する品詞辞書と本研究で作成した用語辞書から語幹を抽出したものである.

3.6 まとめ

本章では，同義文における検索効率を高めるため，または検索ファイルの規模を節約するために,「語の基本形」の提案を行った.「語の基本形」は，品詞を変える派生接辞のみを削除するため，全ての派生接辞を削除した語根に比べて意味的な制約が強い．そして,「語の基本形」を求めるために品詞を変える接尾辞を対象に 262 種類の規則を作成している．

加えて第 3.5 節では，Web 上で公開されている Porter Stemming との処理の比較を行った．なお，Porter Stemming の正解率が低いのは，対象とする接尾辞が少ない (23 種類) ことに起因している．すなわち，本研究で作成した語形変換テーブルを Porter Stemming のプログラムに含まれる変換テーブルと置き換えることで，Porter Stemming の精度を上げることができると思われる．

第4章
語句の重要度計算と評価

本章では，TF.IDF 法を応用した条項の特徴を表す重要語について述べる．

今回，技術取引に関する契約書式集から，14 種類の一般条項と 20 種類の固有条項の計 34 条項 (全 3,526 条文，142,716 語) を対象に，語の抽象化を行い，抽象化した語について重要度を求めた．そしてこれらの重要度から，各々の条項の内容を表す上で重要な語を定め一覧表を作成し，その有効性を評価するための実験を行った．

4.1 重要語の定義

契約書は，条項ごとに内容がまとめられている．また，本研究で対象とするデータは，いろいろな用例についての模範例を集めたものであることから，条項ごとに書かれている内容や構文についての信頼性は高い．従って，ある条項内で多く出現する語をその条項における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語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

4.2 TF.IDF 法

出現頻度の局所性を考慮して索引語を決定する方法として，TF.IDF 法 [8] がある．TF.IDF 法は，索引語の自動抽出法として広く使われている．そこで，本研究

では，TF.IDF 法を応用した重要度を用いる．TF.IDF 法は以下の式で表される．

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

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)

4.3 重要度の定義

本研究では，重要度を以下のとおり定める．

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，条項に含まれる条項例の平均 P，条項 a_i に含まれる条項例の数を 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．

$$I_j^i = \sqrt{\frac{P}{P_i} \cdot tf_j^i \cdot \log_2 \frac{N}{af_j}} \qquad (4.1)$$

本研究で対象とするデータは，条項の種類によりデータ量にばら付きがある．例えば，契約書の中で使われる語句の定義を記述する「定義」条項に関しては，136 種類の条項例があるのに対して，契約書に使用する言語を定める「言語」条項は，12 種類である．そこで，条項例の平均 P を当該条項の例の数 P_i で割った比率を用いて，出現回数の正規化を行う．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．平方根を求めている理由は，重要度の値を条項の内容を推定する等に利用する上で，重要度の大きな値間の差を抑える必要があるためである．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，大きな値を取る．全ての条項に散らばって出現する語の値は，0 となる．

4.4 条項における重要語の作成

条項の特徴を表す重要語があらかじめ判っていれば，条項の内容を推測することが容易となる．加えて，重要語の出現状態から，関連のある条項を推測することが可能となる．

そこで，技術取引に関する契約書式集に含まれる 34 種の条項について，各条項の特徴を表すのに重要な基本形を以下の手順で定め，一覧表を作成した．

＜手順 MKEYWD ＞

(1) 前述の書式集 34 種類の条項に含まれる 142,716 語 (4,489 種) について，重要語になり得ない 292 種の語を除いた 4,197 種の語について，条項毎に基本形を求め，条項頻度と出現回数を求める．
基本形：2,123 種

(2) 4.3節記載の (4.1) 式により重要度を求める．
基本形：11,149 データ，最大値 25.2，最小値 0

(3) 重要語の数 (34 条項の総数) を 1,000 個以内にするために，データを重要度の降順にソートし，データの 20 分の 1 番目の重要度を求める．
基本形：(11,149 ／ 20) 番目の重要度 4.2

(4) 重要度＞ 4.2 で，かつ 条項頻度＜ 27(＝ 34 × 0.8) の語，および，語頻度＞ 1 で，かつ条項頻度＝ 1 の語を抽出し，条項ごとに分類し，重要度を値とする一覧表を作成する．一覧表のデータ構造は以下の通りである．

条項名識別子：重要語：重要度：語頻度：条項頻度

表 4.1は，基本形，語幹および基の語，それぞれを基にした秘密保持条項における重要語の一覧である．

表 4.1 秘密保持条項における重要語一覧

番号	基本形	重	頻	条	語幹	重	頻	条	基の語	重	頻	条
1	confident	8.3	47	10	confidential	8.1	34	7	confidential	8.1	34	7
2	disclose	7.8	65	16	information	6.9	96	23	information	6.6	87	23
3	employee	6.2	28	11	disclose	6.8	50	16	disclose	6.5	26	9
4	inform	6.2	96	25	employee	6.2	28	11	employees	6.3	27	10
5	strict	6.2	13	3	treat	5.5	12	4	secret	5.9	12	3
6	secret	5.7	30	14	secret	5.5	20	10	treat	5.8	10	2
7	treat	4.8	12	7	strictly	5.1	9	3	disclosed	5.3	22	12
8	know-how	4.7	33	20	disclosure	4.9	15	9	know-how	5.2	31	17
9	precaution	4.6	5	1	confidence	4.8	10	5	strictly	5.1	9	3
10	subcontractor	4.5	7	3	know-how	4.7	33	20	confidence	4.8	10	5
11	divulge	4.5	6	2	precaution	4.6	5	1	hold	4.8	10	5
12	leakage	3.6	3	1	secrecy	4.6	10	6	prevent	4.8	11	6
13	recipient	2.9	2	1	divulge	4.5	6	2	disclosure	4.8	14	9
14	unpublished	2.9	2	1	leakage	3.6	3	1	precautions	4.6	5	1
15	necessity	2.9	2	1	recipient	2.9	2	1	secrecy	4.3	9	6
16					necessity	2.9	2	1	sublicensee	4.3	8	5
17					nondisclosure	2.9	2	1	benefiting	3.6	3	1
18					unpublished	2.9	2	1	learnt	3.6	3	1
19	注)								leakage	3.6	3	1
20	重：重要度								codes	2.9	2	1
21	頻：出現頻度								knowledges	2.9	2	1
22	条：条項頻度								object	2.9	2	1
23	(語の出現する条項数)								necessity	2.9	2	1
24									Sub-Licensee	2.9	2	1
25									non-disclosure	2.9	2	1
26									unpublished	2.9	2	1
27									recipient's	2.9	2	1
28									divulging	2.9	2	1

表 4.2 テストデータ

出版社名	秘密保持条項		ノウハウライセンス契約書		
			一般条項	固有条項	
	例数	語数	例数	例数	語数
国際事業開発	10	1,237	13	13	3,172
日本実業出版社	4	459	13	12	2,527
The Japan Times	1	242	10	10(-1)	2,713
東京布井出版	7	347	36	35(-1)	8,412
若竹出版	2	88			
日経文庫	2	220			
中央経済社	13	923			
民事法研究会	3	1,183			
計	42	4,699			

注：35(-1) は，35 条項の内，重要語一覧に載っていない条項が 1 条項あることを意味している．

4.5 評価実験

本研究では以下のような仮説を立て，これらを立証する．

＜仮説＞

(1) 文の特徴を抽出する上で基本形クラスは有効である．すなわち，語幹や基の語を用いる場合に比べて少ないデータで同義の多様なテキストデータに柔軟に対応できる．

(2) 英文契約書という厳密性と網羅性が追求される文においては，その書式がある程度定式化されており，モデル書式集から作成した重要語の汎用性は高い．

(3) 一般条項における重要語は固有条項における重要語よりも汎用性が高い．

4.6 実験の手順

上記の仮説を確かめるために，以下の手順で評価実験を行った．

(1) 4.4節の手順 MKEYWD により，基本形，語幹，基の語それぞれについて重要度を求め，34 条項の特徴を表す重要語の一覧を作成する．

(2) 重要語の一覧を作成する際に使用した書式集と同じ出版社を含む 8 社から出版されている文献 [7]，[10] ～ [16] からテストデータを作成する (表 4.2)．

(3) 3 種類の重要語一覧とテストデータに含まれる語との照合を行い，一致した語の条項名と重要度の累計を取る．

(4) 重要度の累計をキーに降順にソートを行い，最も得点の大きい条項名をテストデータの条項名と推定する．

(5) 実際の条項名と，推定された条項名を人手によりチェックする．

なお，条項名には色々な表現が使われている．これらについては，条項名の意味が同じであれば，同じ条項名とみなす．

例) 秘密保持 :

(a)Confidentiality

(b)Secrecy

(c)Secrecy of Obligation

4.7 考察

表 4.1において，基本形による重要語の数は，基の語を用いた場合の 53%(15/28) であり，これは語幹を用いた場合の 64%(18/28) よりも縮小率が大きい．従って，基本形による重要度が基の語や語幹による重要度と同じ精度で条項の特徴を表すことができれば，基本形の方が有効であることは明らかである．

なお基本形の有効性については，抽象化による副作用についての考察も必要である．例えば，表 4.1において，番号 17 および番号 25 にある “nondisclosure(非

開示)" という語は，基本形のリストには見当たらない．これは，基本形へと抽象化を行った結果，条項頻度が 2 以上となり，手順 MKEYWD (4) の条件を満たさなくなったためにおきた情報落ちである．そこで，抽象化による情報落ちを防ぐために (4) の「条項頻度＝1」という条件を「条項頻度＜3」に緩めて重要語の抽出を行った．すると，新たに "nondisclosure"，"code"，"oral"，"competition"，"improper"，"subcontract" の 6 語が重要語として加わった．しかしながら，これら 6 語の内で「秘密保持」条項の特徴を表す重要語として適当だと思われるのは，"nondisclosure"1 語である．従って，抽象化による情報落ちを防ぐために手順 (4) の条件を緩めることは，得策ではないと思われる．そこで，これらの副作用については，手作業でリストに追加し対処するものとする．なお今回の実験では，手順 MKEYWD で求めたリストをそのまま用いた．

表 4.2は，評価実験に用いたテストデータの詳細である．本研究では，国際事業開発株式会社 [7] の書式集を用いて研究をすすめている．そこで，テストデータは同じ会社からのものに加えて異なる 7 社の出版社 [10] ～ [16] から抽出している．具体的には，42 例 (全 141 条文，4,699 語) の秘密保持条項と，3 社の文献 [7]，[11]，[14] に記載されているノウハウライセンス契約書の 3 文例 (全 71 条項，462 条文，8,412 語) のテストデータを作成している．なお，文献 [11] のライセンス契約書 20 条項の内，1 条項 (政府承認申請条項) については 34 条項のリストに存在しない条項であった．

表 4.3は，実験結果をまとめたものである．

表 4.3の番号 10 の結果から，基の語による重要語の 60%の重要語 (番号 3) で，語幹や基の語よりも高い推定結果が得られている (85%＞ 79%，85%＞ 80%) ことから，仮説 (1) の基本形の有効性が示されている．更に，表 4.3番号 8，9 の結果から，基本形の方が多様な文に柔軟に対応できることがわかる．すなわち，基の語においては，同じ出版社のテストデータを用いた場合の推定結果 (83%) より他社のものを用いた場合 (79%) の方が低くなっているのに対し，基本形の場合は他社のものを用いた場合 (86%) の方が同社のものを用いた場合 (83%) より高くなっている．このことから，仮説 (1) の語の抽象化は有効であるといえる．

仮説 (2) についても，3 種類のノウハウライセンス契約書全 71 条項の内，重要

表 4.3 実験結果 (1)

番号		基本形	語幹	基の語
1.	書式集に含まれる語種	2,123	2,717	4,489
	基の語を 1 とした場合の比率	0.5	0.6	1.0
2.	34 条項に含まれる語の重要度データ	11,149	12,305	15,745
	基の語を 1 とした場合の比率	0.7	0.8	1.0
3.	34 条項の特徴を表す重要語数	637	763	1,079
	基の語を 1 とした場合の比率	0.6	0.7	1.0
4.	秘密保持条項の特徴を表す重要語	15	18	28
	基の語を 1 とした場合の比率	0.5	0.6	1.0
5.	秘密保持条項における条項名の推定結果	93%	81%	83%
	(正解数/全条項数) (不正解数)	39/42 3	34/42 8	35/42 7
6.	一般条項における推定結果	86%	89%	89%
	(正解数/全条項数) (不正解数)	31/36 5	32/36 4	32/36 4
7.	固有条項における推定結果	74%	68%	68%
	(正解数/全条項数) (不正解数)	25/34 9	23/34 11	23/34 11
8.	他社出版のテストデータにおける	86%	79%	79%
	推定結果 (正解数/全条項数) (不正解数)	65/76 9	60/76 16	60/76 16
9.	書式集と同じ出版社のテストデータにおける	83%	81%	83%
	推定結果 (正解数/全条項数) (不正解数)	30/36 6	29/36 7	30/36 6
10.	全テストデータにおける推定結果	85%	79%	80%
	番号 8 + 9 (正解数/全条項数) (不正解数)	95/112 17	89/112 23	90/112 22

注：番号 7 において，重要語リストに含まれない条項 (表 4.4番号 26) は，対象からはずしている．

語のリストに含まれない条項が一つであること (表 4.2) に加え，表 4.3番号 8，9 より仮説が成り立つといえる．なぜならば，同社のテストデータの約 2 倍の条項例を用いても基本形については同社のテストデータ以上の推定結果が得られており (83%→ 86%)，また語幹および基の語においても他社の条項名の推定において 79%の推定結果が得られている．仮説 (3) については，表 4.3番号 6(一般条項) と 7(固有条項) より仮説が成立すると結論できる．

表 4.4は，正しく推定されなかったものをまとめたものである．番号 1 ～ 9 は，42 の秘密保持条項における推定結果であり，表 4.3の番号 5 に対応している．番号 10 ～ 27 は 3 種類のノウハウライセンス契約書における推定結果で表 4.3の番号 6，7 に対応している．番号 28 は，表 4.3の番号 10 の不正解数に 34 の重要語リストに含まれない 1 条項 (表 4.4の番号 26) を加えたものである．

基本形の場合，正しく推定されなかったもの，すなわち推定の 1 位が正しい条項名でなかったものが 18 条項あり，候補の中に正しい条項名が含まれていたものが 16 条項あった．加えて，その 16 条の内，正しい条項名が候補の 2 位にあるものが 8 条項あった (番号 29)．

今回用いたテストデータは，英文契約書の例文から抜粋したものである．すなわち，内容の信頼性は高い．更に，表 4.3の結果から，基本形を用いた重要度計算が有効であることがわかる．

これらのことから，表 4.4の番号 28，29 の結果は,「条項名とその条項に記載されている内容とのずれ」を示唆していると解釈することができる．

具体的には，表 4.4番号 4 は，秘密保持条項に関する用例の中の「守秘手続に関わる監査」について記載された条文であり，その推定結果は，1 位が「監査」，2 位が「秘密保持条項」になっている．同様に，番号 6 は，条項名は「秘密保持」であるが，その内容は著作権の供給者の規定および文書の使用範囲を規定するもので，無効となった条項を他の条項から分離する，すなわち条項を規定する「分離可能性」が 1 位に,「著作権 (copyright)」という重要語から,「工業所有権」が 2 位に推定されている．

以上のことから，その条項に付けられた条項名の意味と，条項に記載されている内容にずれがある場合に，条項名の推定が正しく行われないことがわかる．

表 4.4 実験結果 (2)

出版社名 条項名	番号	基本形 推定結果	正解	語幹 推定結果	正解	基の語 推定結果	正解
国際事業開発		該当有:1	無:0	該当有:2	無:0	該当有:1	無:0
秘密保持	1.	研究成果	2	研究成果	2	秘密保持	○
	2.	秘密保持	○	競合・地域制限	2	実施許諾，技術譲渡	2
中央経済社		該当有:1	無:1	該当有:2	無:1	該当有:1	無:2
秘密保持	3.	秘密保持	○	技術情報	2	秘密保持	○
	4.	監査	2	監査	3	仲裁，裁判管轄	*(4)
	5.	秘密保持	○	秘密保持	○	工業所有権	3
	6.	分離性	*(2)	分離性	*(2)	工業所有権	*(1)
東京布井出版		該当有:0	無:0	該当有:2	無:0	該当有:2	無:0
秘密保持	7.	秘密保持	○	工業所有権	2	工業所有権	2
	8.	秘密保持	○	技術情報	2	研究成果	2
民事法研究会		該当有:0	無:0	該当有:1	無:0	該当有:1	無:0
	9.	秘密保持	○	終了後の措置	2	終了後の措置	4
国際事業開発		該当有:5	無:0	該当有:4	無:1	該当有:4	無:1
◇ 定義	10.	技術情報	4	技術情報	4	原材料等	3
◆ 競合	11.	契約期間	5	契約期間	*(6)	契約期間	*(4)
◆ 技術情報の開示	12.	技術情報	○	技術援助	2	技術援助	3
◆ 改良	13.	研究成果	2	研究成果	2	研究成果	2
◆ 宣伝	14.	対価	2	対価	8	技術訓練	5
◇ 不可抗力	15.	通知	3	不可抗力	○	不可抗力	○
日本実業出版社		該当有:4	無:0	該当有:4	無:1	該当有:4	無:1
◇ 定義	16.	対価	2	対価	4	原材料等	2
◆ 付与	17.	実施許諾，技術譲渡	○	秘密保持	2	実施許諾，技術譲渡	○
◆ 競合	18.	契約期間	5	契約期間	*(6)	契約期間	*(4)
◆ 技術情報の開示	19.	技術情報	○	技術情報	○	技術援助	2
◆ 改良	20.	研究成果	2	研究成果	2	研究成果	2
◆ 宣伝	21.	対価	2	対価	7	対価	4
The Japan Times		該当有:5	無:1	該当有:5	無:1	該当有:5	無:1
◇ 定義	22.	対価	12	研究成果	14	商標，標章等	10
◆ ノウハウの開示	23.	秘密保持	4	秘密保持	4	秘密保持	4
◆ 技術援助	24.	技術訓練	3	技術訓練	3	技術指導	3
◆ 改良	25.	研究成果	2	研究成果	2	研究成果	2
◆ 政府承認申請	26.	不可抗力	**	不可抗力	**	不可抗力	**
◇ 契約期間	27.	対価	6	対価	10	対価	3
該当有/全 該当無/全	28.	16 /18(89%)	2/18	20 /24(83%)	4/24	18 /23(78%)	5/23
2位に正解があるもの	29.	50%	8/16	55%	11/20	44%	8/18

注 1) *(n)：条項名の全候補 n 個の中に該当する条項名がない

注 2) **：重要語のリスト (34 条項) に該当する条項名がない

注 3) ○：正しく推定されたもの, n：正しい条項名の推定順位

注 4) ◇：一般条項，◆：固有条項

なお今回，定義条項を一般条項としている．定義条項では，各条項で使われる語の定義を行っているため，各条項の重要語が出現し，推定の正解率が低くなる．したがって，今回の実験においては，全体の正解率を下げる要因となっている．

4.8 まとめ

本章では，同義の多様な英文のテキストデータに柔軟に対応できるように導入した「語の基本形 (第 3 章に詳述)」を用いた重要度計算について説明し，その有効性を示した．

具体的には，英文契約書の書式集から，34 の条項の内容を表す重要語の一覧を求め，112 の条項例について条項名の推定を行い，85%の正しさで条項名を推定できることを確認した．

また，表 4.4の結果から，重要語のリストを用いた条項の推定が，条項に付けられた見出しと条項に記載された内容とのずれを示唆する機能を持つことが分かった．

一般に契約書では，見出しと条項に記載されている内容との意味的な違いから生じる誤解を避けるために，「見出し (Headings)」条項で，「見出しは契約の解釈上，無意味である」ことを規定している．これらのことから，契約書のモデル文例集から作成した条項単位での重要語リストは，「契約書の草案の見直しを支援する」という新たな機能を持つことがわかった．

今回契約書式集から作成した重用語リストの利用方法を以下にまとめる．

- 実際の国際取引で作成された英文契約書を電子化部門別に分散化する場合の，条項の内容の推定
- データベースに含まれる過去の契約書の情報を検索する際のキーワード
- 契約草案を作成した際の，または提示された相手方作成の草案の内容のチェック

第 5 章
語句の類似度計算とクラスタリング

本章では，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，および，類義語辞書の作成支援としてファジィ関係行列を使ったクラスタリングについて説明する．

5.1 単語間の類似度

ある領域内に共起する頻度が大きい単語のペアは関係が強いとするとき，その関係の強さを数値化する方法として，次のような尺度がある．

(1) 相互情報量 [30] [35]

特徴：十分な出現回数を持つ二つ組に有効な手法である．単語の組の出現回数が低くなる程値が大きくなる．

$$MI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)} \tag{5.1}$$

(2) $DICE$ 係数 [31]

特徴：同時出現回数 f_{XY} の大小に関係なく，独立出現回数と同時出現回数の相対比により値が決まる．同時出現回数が過小評価される．

$$DICE(X,Y) = \frac{2 \cdot f_{XY}}{f_X + f_Y} \tag{5.2}$$

(3) 重み付き $DICE$ 係数 [32]

定義：$DICE$ 係数に同時出現回数を重み付けしたもの．

$$WDICE(X,Y) = (\log_2 f_{XY}) \cdot \frac{2 \cdot f_{XY}}{f_X + f_Y} \tag{5.3}$$

また，統計情報に基づく尺度において，単語の類似度を数値化する方法として，内積 [33] と KL-distance [34][36] がある．この場合の尺度としては，前述の (1) ～ (3) の他，TF・IDF や順序対の出現回数などが使われる．

これらの手法を使って，大量の形態素解析または構文解析が施されたコーパスから自動的にワードクラスタリングを行う研究が報告されている．[33] ～ [37]

ところで本研究のように，契約書という特異な分野に含まれる語句の類概念データを作成する際，大量の解析済みのコーパスを見つけること，または作成することは困難である．すなわち，(a) 隣接する単語の文字列から統計的に類似度を計算するにはデータ量が少ない，更に，(b) 構文解析を行うには条文が複雑であり，(c) 専門分野に特化していることから，シソーラスを用いた類似度計算は困難である．

本研究では，長文で複雑な条文から単文を生成する際の統語関係を推測することを目的として，語間の関連の強さを表す関連度を定義し，関連度データを作成した．

更にこの関連度を用いた，語間の類似度についての提案を行なった．

5.2 関連度

関連度とは，品詞のタグ付けを行った契約書式集から統語関係を意識した語の二つ組を抽出し，二つ組の出現回数を基に，二つ組を構成する各要素の汎用性（ここでの汎用性とは，「“get” は “gain” より汎用性が高い語である」という時の意味である）を加味した数値である．次節では，語の二つ組の抽出方法について述べる．

5.2.1 二つ組の抽出

係り受けの正しい二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である．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容易ではない．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's Rule Based Tagger*[9] 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用できる状態にあるため，ある語の前後 8 語または 10 語というようにウィンドウを設け，2 つ組を求め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する．

なお，条文において，三つ以上の語句を “and”，“or” で列挙するケースが多いため，区切り記号から “,” を除いている．

次に問題となるのが，二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，意味をなさない．

そこで，以下のように重複を許した二つ組を使って関連度を求める．

例えば，動詞，形容詞と名詞からなる語群，

$n1$	$v1$	$adj1$	$n2$	$n3$	$n4$

の場合，動詞と名詞からなる二つ組として $(v1, n2)$,$(v1, n3)$,$(v1, n4)$ を，名詞と動詞からなる二つ組として $(n1, v1)$ を，形容詞と名詞からなる二つ組として $(adj1, n2)$,$(adj1, n3)$,$(adj1, n4)$ を抽出する．

以上の方法で二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の網羅性を高めることができる．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，連語データと照合し，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．

このようにして求めた17,354組の二つ組について関連度を求めた．次節では，関連度の定義を行う．

5.2.2 関連度の定義

いま，契約文書に出現する二つの語の組 (l,r) について考える．

$$l=(s_l,t_l),\quad r=(s_r,t_r)$$

なお,s は文字列,t は品詞を表す．

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 と同じ品詞を持つ二つ組の集合を 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 と同じ品詞を持つ二つ組の集合を 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 の二つ組の出現回数を $f(l,r)$，二つ組 (l,r) の N_l と N_r の積の最大値を max_{lr} とするとき（本実験では，(動詞，名詞）の二つ組の動詞の種類と名詞の種類の積 $836\times834=697,224$ の定数），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 N の元の個数を表す．

$$REL(l,r)=f(l,r)\cdot\left(1.0+\frac{\log_2\frac{|N_l|}{|N_l^r|}+\log_2\frac{|N_r|}{|N_r^l|}}{\log_2\left(max_{lr}\right)}\right) \tag{5.4}$$

すなわち関連度とは，同時出現回数 $f(l,r)$ に，各単語の汎用性の相違 $\left(\log_2\frac{|N_l|}{|N_l^r|}+\log_2\frac{|N_r|}{|N_r^l|}\right)$ の正規化した値を，重みとして掛けたものである．

5.2.3 重み付き $DICE$ 係数と関連度の比較

表5.1は，重み付き $DICE$ 係数と関連度を比較した表である．

表 5.1 重み付き $DICE$ 係数と関連度の比較

No.	word pair	$WDICE$	REL	f_{lr}	N_r^l	N_l^r	f_l	f_r
1	agreement, cancel	0.0173	9.79	7	240	12	2253	25
	agreement, commence	0.0171	9.55	7	240	19	2253	47
	agreement, provide	0.0157	8.81	7	240	92	2253	238
2	render, arbitrator	1.0022	20.94	13	27	7	54	42
	sell, product	1.0087	188.04	150	110	209	405	1745
	manufacture, product	1.0203	225.79	179	99	209	881	1745
	bear, expense	1.0565	58.02	40	37	44	88	315
	take, action	1.0566	39.87	27	52	22	130	113
	use, trademark	1.0853	110.61	80	88	46	670	262
	assume, responsibility	1.1270	24.27	15	19	9	42	62
3	grant, patent	0.3144	41.9	31	65	93	395	582
	grant, sublicense	0.5656	42.9	28	65	8	395	81

番号1は，出現回数 $f(l,r)$(表5.1では，f_{lr}) が同じ二つ組の例である．番号2は，1.0に近い値の重み付き $DICE$ 係数を持つ二つ組の例を，番号3は，二つ組の出現回数の順位と関連度の順位が逆転している例を示している．

番号1の例において，名詞･動詞の二つ組(agreement,cancel)，(agreement,commence)，(agreement,provide)は，同時出現回数が共に7回である．しかし，関連度においては，12個の(名詞，cancel)の組合せと92個の(名詞，provide)の組合せの違いが値に反映されている．すなわち，“cancel”より汎用性の高い動詞“provide”を含む二つ組を低く評価している．

一方，重み付き $DICE$ 係数においても，“cancel”の25回の出現回数と“provide”の238回の出現回数の違いが値に反映されている．

番号2の例では，重み付き $DICE$ 係数と出現回数との相違がわかる．すなわち，重み付き $DICE$ 係数は同時出現回数と独立出現回数の相対比に重きが置かれている．

番号3の例は，関連度と出現回数の相違を示したものである．関連度は，同時出現回数と対応した値であるが，語の汎用性にバラツキがある場合は，その度合を反映している．

重み付き $DICE$ 係数と関連度の相違点は，関連度が同時出現回数に各単語の汎用性の相違を加味しているのに対し，重み付き $DICE$ 係数は独立出現回数と同時出現回数の相対比に重みとして同時出現回数の情報 ($\log_2 f_{xy}$) を加味している点にある．

同時出現回数と独立出現回数のどちらに重きを置くかの判断は，その利用目的に依存する．

本研究で対象とするテキストデータは，契約書式集から作成されており，不適切な語句は含まれない．現在のところ，品詞タグ付け，および複合語などをまとめる処理は，完全ではないが，これらが完全になされたテキストデータにおいては，統語関係の正しい二つ組の同時出現回数の持つ意味が大きくなる．そこで，出現回数の持つ意味を重視し，かつ「汎用性の低い語は重要語である可能性が高い」という仮定の基に，同時出現回数が同じである場合は，独立出現回数の低い方の価が高くなるような関連度の定義を行っている．

表 5.2 重み付き $DICE$ 係数と関連度の統計データ

	max	min	mode	mean
$WDICE$	7.3821	0.0005	0.0011	0.0797
REL	1125.47	2.37	2.64	7.82
f_{xy}	1034	2	2	5.61

注：但し，二つ組の出現回数が 2 回以上のものを対象としている．

この関連度は，長文で複雑な条文から，単文を生成する際の統語関係を推定する指標として用いる．

表 5.2は，重み付き $DICE$ 係数と関連度および同時出現回数の統計データである．同時出現回数の最頻値（$mode$）が 2，平均が 5.61 から，尺度として相互情報量を用いるには，テキストデータの規模が小さいことがわかる．

5.3 名詞間の類似度

5.3.1 類似度の定義

ある語 w と名詞 n との関連度を $REL(w,n)$ とする．語 w と同じ品詞 t を持つ語と名詞 i の二つ組の集合を N_t^i，語 w と同じ品詞 t を持つ語と名詞 j の二つ組の集合を N_t^jとすると,N_t^iと N_t^j の共通の要素の集合 COM_i^jは，

$$COM_i^j = N_t^i \cap N_t^j \tag{5.5}$$

となる．

ここで，類似度 $SIMtt(i,j)$ は，語 w と同じ品詞 t を持つ語との二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とする．

名詞 i と名詞 j の類似度 $SIMtt(i,j)$ の定義を以下に示す．

$$com = |\,COM_i^j\,| \tag{5.6}$$

$$SIMtt(i,j) = \frac{\sum_1^{com}\Big(REL(w,i)\cdot REL(w,j)\Big)}{\sqrt{\sum_1^{com}\Big(REL(w,i)\Big)^2}\cdot\sqrt{\sum_1^{com}\Big(REL(w,j)\Big)^2}}\cdot\frac{2\cdot com}{|N_t^i|+|N_t^j|} \tag{5.7}$$

本研究では，以下の4種類の二つ組を使って名詞間の類似度を求める．

(1) 名詞と動詞 (主語と述語)

(2) 動詞と名詞 (述語と目的格)

(3) 形容詞と名詞

(4) 前置詞と名詞 (場所や時間を表す修飾部)

すなわち，名詞間 (x,y) の類似度 $SIMnn(x,y)$ 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．

$$SIMnn(x,y) =$$
$$(SIMnv\cdot w1 + SIMvn\cdot w2 + SIMan\cdot w3 + SIMpn\cdot w4)\,/\,4 \tag{5.8}$$

ここで，$w1$，$w2$，$w3$，$w4$ は，それぞれに係る重みである．

なお，契約書において，場所や時間，理由などを表す前置詞と名詞が結びついた修飾部が多く出現することから，名詞間の類似度を求める要素として，前置詞と名詞の二つ組を加えている．しかしながら，前置詞と名詞の組においては，前置詞の多義性の問題 (例えば，前置詞 'for' は期間を表す場合と理由を表す場合がある) と，前置詞の係り受けの問題，すなわち，前置詞が前の動詞に係るのか後ろの名詞に係るのかの決定が困難であるという問題から，他の3種の重み $w1$，$w2$，$w3$ を3にし，$w4$ を2にして，計算を行った．

5.3.2 類似度データの作成手順

技術取引契約書の書式集 4,596 条文 (162,298 語)[7] を用いた，名詞間の類似度データ作成手順 SimNN を以下に示す.

＜手順 SimNN＞

(1) *Brill's Rule Based 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) 名詞の種類および二つ組を求める
　名詞の数：894 種
　組合せの数：399,171 通り

(5) 名詞と動詞，動詞と名詞，形容詞と名詞，前置詞と名詞の二つ組とその出現回数を求める.
　名詞と動詞：6,417 組

動詞と名詞：7,655 組
形容詞と名詞：818 組
前置詞と名詞：2,569 組
総計：17,459 組

(6) 5.2.2節の (5.4) 式より関連度を求める．

(7) 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，5.3.1節の (5.7) 式より名詞間の類似度を求める．

(8) 5.3.1節の (5.8) 式より最終的な名詞間の類似度を求める．
類似度の最大値：0.841
類似度 $>$ 0 の組：209,274 組 / 399,171 組

5.3.3 類似度データの結果と考察

表 5.3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．

欄 1は，書式集から連続する 3 語を抽出し，出現頻度をもとめた「三つ組の頻度データ」との照合により得た，名詞 1 と 2 の隣接関係を示している．すなわち，(1)b.“mail” の欄 1 の “cb” は，b.“mail” が c.“air” の後ろに隣接して出現していることを示す．すなわち，“air mail” は，複合語の可能性の高い組か，または述部を構成する名詞群を意味している．また，欄 2 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．そして，欄 2 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，類似度データから隣接関係にある語を削除または連語としてまとめた後に，クラス分けを行うことで，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．

表 5.3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，

表 5.3 ある名詞と 0.3 以上の類似度を持つ名詞

名詞 1	名詞 2	類似度	1	2
(1) airmail 航空郵便				
	a. post 郵便	0.559		○
	b. mail 郵便	0.455	cb	○
	c. air 空	0.422	cb,cd	○
	d. letter 書簡	0.345	cd	
	e. postage 郵送料	0.336		○
	f. telex テレックス	0.336		
	g. telegram 電報	0.327		
	h. cable 電信	0.323		○
(2) disclosure 開示				
	a. content 内容	0.483		
	b. respect 関連	0.456		
	c. copyright 著作権	0.406		
	d. force 効力	0.375		
	e. title 財産所有権	0.364		
(3) currency 通貨				
	a. exchange 両替，為替	0.447	ab	○
	b. bank 銀行	0.438	ab	○
(4) money 金銭				
	a. duration 存続期間	0.317		
	b. class 部類	0.303		
	c. endeavour 努力	0.303		
	d. voucher 商品券 領収証	0.303		
	e. remuneration 報償	0.303		
(5) trademark 商標				
	a. mark マーク	0.430	ca	○
	b. invention 発明	0.387		
	c. trade 商業	0.368	ca,cf	
	d. design デザイン	0.328		
	e. patent 特許	0.316		
	f. name 名称	0.316	cf	

欄 1 ： 隣接する名詞の対

欄 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

欄 2 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” と二つ組をなし，5.3.1節の (5.7) 式における二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，(5.8) 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．

“money”(金銭) と類似度が高い組についても，共通の組が前置詞 “in”，“of”，および，汎用性の高い動詞 “provide”，“pay” との組であることが原因であると推測される．

これらの問題において，“in force”，“in respect of” などの慣用句をまとめる処理を徹底させることで，不適切な前置詞と名詞の組を減らすことができる．また，品詞タグ付けの修正，および複合語などをまとめる前処理を徹底させることで，統語関係の正しくない二つ組を減らすことができる．そこでこれらの処理を徹底させた後，必要であれば重みについての検討を行う．

一方，意味の上では類似していると思われる “money” と “currency” の類似性が低いことが，類似度によりわかる．これは，“money” が形容詞を持たずに汎用性の高い動詞 “pay”，“provide” と組をなすのに対して，“currency” は，“local”，“foreign”，“settled”，“prevailing” などの多様な形容詞と組をなし，また，“designate”，“indicate”，“remit” などの汎用性の低い動詞と組をなすことに起因している．すなわち，“currency” を含む単文において，“currency” を “money” で置き換えることが出来ないという推測が可能となる．

以上の結果から，本研究で提案する類似度は，(1) 複合語の可能性を調べる，(2) 語の等位関係を調べる，(3) 類義語での置き換えの可能性を調べる際に利用できることがわかる．

なお，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する．かつ，(5.8) 式で明らかなように 4 種類の類似度計算を行う必要があることから，1 回の名詞間の類似度計算に 3 日を要するという問題がある．

5.4 クラスター分析

ここでは文献 [39][40][41] をもとにクラスター分析の概要を述べる.

あるデータの構造または規則性を知る上で有効な方法としてクラスター分析がある. クラスター分析とは, データを構成している個体または属性を何らかの基準によって, いくつかの均質なグループ (クラスター) に分類する手法を総称したものである.

クラスター分析は, 階層的クラスター分析と非階層的クラスター分析に大別される. これらのクラスター分析には, 分類されたクラスターの結果がどの程度の意味を持つかの仮説検定を行う議論は含まれていない. すなわち, 多変量解析のような統計的なモデルを持たない. 従って分類の妥当性は, 得られた結果の解釈により成される.

5.4.1 階層的クラスター分析

階層的クラスター分析は, 結果的に樹形図 (デンドログラム) が得られる方法で, とくにクラスター数を定めず, 対象の階層構造を求め, 目的に応じて分類する方法である. 階層クラスター分析には, クラスターが形成されていく過程が階層的な構造を持ち, その形成過程におけるクラスター間の類似度または距離が, 一つ前の段階で求めた類似度または距離によって計算されるという特徴がある. 階層的クラスター分析の基本的なアルゴリズムは以下の四つの手順から成る.

＜手順 1 ＞

総個体数を N とする.

各個体間相互の類似度あるいは距離を計算する.

初期状態として, N 個の個体それぞれが一つのクラスターを形成しているものと考える. すなわち, クラスターの個数 M を M ＝ N とする.

＜手順 2 ＞

M 個のクラスターの中で最も類似度の大きい, または距離の小さい対を求めて, それを一つのクラスターにまとめる.

M を M − 1 とし, M ＞ 1 なら手順 3 に, そうでなければ手順 4 の処理を行う.

＜手順 3 ＞

新しく融合されたクラスターと他のクラスターとの類似度または距離を計算し，手順2の処理を行う．

＜手順4＞

結果を出力し，終了する．

個体間の類似度または距離を計算する際の尺度として，以下のものがある．

(1) 間隔尺度 (連続量で表されるデータ)

(2) 名義尺度 (カテゴリー的データ)

(3) 順序尺度 (大小関係にのみ意味のある数値データ)

そして，類似度または距離を定義する方法として以下のものがある．

(1) 間隔尺度

a. ユークリッド距離

b. 重み付きユークリッド距離

c. ミンコフスキー距離

d. マハラノビス距離

e. 内積による類似度

(2) 名義尺度

各個体について 0,1 の値を取る場合，個体間の取る 0,1 の値の組み合わせとその頻度を用いて類似度を定義する手法に以下のものがある．

a. 類似比

b. 一致係数

c. Russel-Rao の係数

d. Rogers-Tanimoto の係数

e. Hamman の係数

f. ファイ係数

(3) 順序尺度

順位で表現される値を，連続的な値に変換し，間隔尺度の手法で定義する方法がある．また，順位を表す値をそのまま用いて，対応する個体の順位の一致度を数量化する方法として以下のものがある．

a. Spearman の順位相関係数

b. Kendall の順位相関係数

次にクラスター間の類似度または距離の定義方法とその性質について述べる.

(1) 最短距離法 (nearest neighbor method)

クラスター A と B の個体の全ての組み合わせについて距離を求め, 距離の最小値をクラスター間の距離とし, クラスター間の距離の最小のものを融合する.

接近している二つのクラスター間の分類能力は落ちるが, 線状 (曲線状) をなすクラスターの検出能力に優れている. 反面, この長い線状のクラスターを生み出す性質によって, クラスターの相対する端にある個体が著しい非類似性を示すこととなる.

(2) 最長距離法 (furthest neighbor method)

(1) とは逆に, クラスター間の距離として二つのクラスターに属する個体間の距離の最大値を取る. クラスターが融合されるたびに各クラスター間の距離が大きくなる. (1) に比べてクラスターの分離能力は大きくなる. ただし, クラスター間の相違に関する解釈は困難である.

(3) メジアン法 (median method)

(1) と (2) の中間的手法であり, (1) と (2) の欠点が緩和される分, 長所も現れない.

(4) 重心法 (centroid method)

クラスター間の距離として, 各クラスターの重心間の距離を用いる. 但し, 個体間の距離としてユークリッドの 2 乗距離を用いる. 融合された後に再計算した距離の最小値が融合される前の最小値より小さい場合が生じる, 従って樹形図を描く際に線が交差することがある.

(5) 群平均法 (group average method)

二つのクラスター間の全ての個体の組み合わせについて距離を求め, その平均値をもとに融合を行う. 従って, 極端な値に依存してクラスターを定めるものではないため, クラスター内の最小または最大の類似性についての解釈ができない.

(6) ウォード法 (Ward method)

異なるクラスターを一つにまとめる際に失われる情報 (情報損失量) をクラスター間の距離と定義し，情報損失量の最小のものを融合する．

5.4.2 非階層的クラスター分析

階層的クラスター分析は，全ての個体間の類似行列を算出し記憶しておく必要があることから，分類される個体数が多くなるにつれて，計算量および記憶量が膨大となり，実行不可能となる．従って，多数のデータを対象とした場合にはそれらを逐次分類する手法が必要となる．データを逐次分類する目的のために，非階層的クラスター分析が用いられる．

非階層的クラスター分析の代表的な手法には，K-平均法 (K-means method) と ISODATA 法がある．

非階層的クラスター分析の基本的な考え方を以下に示す．

(a) 初期条件として，クラスターの核となる個体 (シード点) を決定する．

(b) 逐次個体と各シード点間の距離を計算し，その距離をもとに分類を行う．

(c) ある収束条件を設定し，条件を満たすまでシード点の変更を繰り返す．

(1) K-平均法 (K-means method)

これは J.B.MacQueen によって提案された手法である．最初にシード点の個数，クラスター間の距離の閾値，クラスターと個体間の距離の閾値を定める必要がある．K-平均法は，収束が達成されるまでシード点の変更を繰り返す必要はなく，シード点の変更は 1 回のみであり，ISODATA 法に比べて簡便な手法である．

クラスターの数を K，総個体数を N とすると，K-平均法における計算量と比較量は共に $O(K^2)$ となる．一方，階層的クラスター分析における計算量と比較量は共に $O(N^2)$ となる．一般に K < N の関係が成り立つことから，K-平均法は階層的クラスター分析に比べて僅かな仕事量で処理を行う．しかしながら，初期値として K の値を定義する必要があり，このことは分析者の大きな負担となる．

(2) ISODATA 法

この手法は G.H.Ball & D.J.Hall によって開発されたもので，収束条件，処理

を繰り返す限度数，作成されるクラスター数の最小値，クラスター内の個体数の最小値，クラスター間の距離の閾値など，最初に7種類のパラメータを設定する必要がある．

なお，本研究では非階層的クラスター分析を行うほどデータ数が大きくないため，階層的クラスター分析により分類を行う．

5.5 ファジィ2項関係行列

本章では，ファジイ関係行列の推移的閉包によるクラスタリングについて述べる [42][43].

推移的閉包によるクラスタリングは，階層的クラスタリングの最短距離法に等価である．しかし，最短距離法によるクラスタリングでは，クラスターの個数が一つずつ減っていくのに対して，推移的閉包によるクラスタリングでは，一つずつ減るとは限らないという違いがある [44].

また，階層的クラスタリングによる手法は，クラスター間の類似度の設定方法，および出力結果である樹形図からクラスを求める際のカットするレベルの決定などの問題が生じる．

本章で提案する手順 ClassNN は，閾値の初期値としてグレードの平均値を用い，変位として 0.01 を加算していくことで，機械的にクラスの生成を行う．また，クラスター間の距離を再計算するなどの必要がない．

5.5.1 定義と手順

手順 SimNN で作成した名詞の類似度 $SIM_{nn}(x,y)$ (5.3.1(5.8)式) を $\mu_s(x,y)$ と改め，ファジィ関係 R_s とし，$\mu_s(x,y), x \in N, y \in N$ と表す．$\mu_s(x,y)$ は x と y がファジィ関係 Rs に属する度合 (grade of membership) を表し (以下，グレードという)，μ_s をメンバーシップ関数 (membership function) という．

ここで，N が $N = \{m_1, m_2, m_3, ..., m_n\}$ なる有限集合であるとすると，ファジィ関係 Rs は n × n 正方行列によって表すことができ，これをファジィ行列 (fuzzy matrix) という．なお，メンバーシップ関数 μ_s は区間 [0,1] の値を取るこ

とから，ファジィ行列の成分も [0,1] の間の値を取る．

また，$\mu_s(x,x) = 1$ すなわち反射律を満たし，かつ，$\mu_s(x,y) = \mu_s(y,x)$ であることから対称律を満たす．このように，反射律と対称律を満たすファジィ関係 R_s を相似関係 (resemblance relation) という．

今，相似関係 Rs からファジィ同値関係 (equivalence relation) を導き，名詞からなる全体集合 N を同値類で分解することを考える．

ファジィ同値関係の定義を以下に示す．

[定義 1] ファジィ同値関係とは，次の性質を満足するファジィ関係 S のことである．

(1) 反射律 (reflexive)：

$$\mu_S(x,x) = 1 \tag{5.9}$$

(2) 対称律 (symmetric)：

$$\mu_S(x,y) = \mu_S(y,x) \tag{5.10}$$

(3) 推移律 (transitive)：

$$\mu_S(x,z) \geq \vee_y\{\mu_S(x,y) \wedge \mu_S(y,z)\} \tag{5.11}$$

$$a \vee b : max\{a,b\}$$

$$a \wedge b : min\{a,b\}$$

すでに，相似関係 Rs は，反射律と対称律を満たしていることから，相似関係 Rs を推移的にすることを考える．

[定義 2] R を任意のファジィ関係とするとき，次式で表されるファジィ関係 $\hat{R}$ は推移的である．

$$\hat{R} = R \cup R^2 \cup ... = \cup_{j=1}^{\infty} R^j \tag{5.12}$$

$R \cup S$:

$$\mu_{R \cup S}(x,y) = max\{\mu_R(x,y), \mu_S(x,y)\} = \mu_R(x,y) \vee \mu_S(x,y)$$

ここで,

$$R^j = \underbrace{R \circ R \circ \ldots \circ R}_{j} \tag{5.13}$$

$R \circ S$:

$$\mu_{R \circ S}(x,z) = max\ min\{\mu_R(x,y), \mu_S(y,z)\} = \vee\{\mu_R(x,y) \wedge \mu_S(y,z)\}$$

である.

この関係 $\hat{R}$ を R の推移的閉包 (transitive closure) という.

R が推移的であるという条件は式 (5.11) より,

$$R = \hat{R} = \vee_{j=1}^{\infty} R^j \tag{5.14}$$

と表せる.

本研究では,

$$R(x,y) \leq \vee_{i=1}^{n}\{R(x,m_i) \wedge R(m_i,y)\} \tag{5.15}$$

となる $R(x,y)$ の値を, 右辺の値まで増加することを繰り返すことで推移律を満たす行列を求める.

以下にクラス分けの手順を示す.

＜手順 ClassNN＞

(1) 類似度データより, 相似関係行列を求める.

対象となる名詞 894 種： 894 × 894 行列

(2) 推移律を満たす同値関係行列を求める．
Max を $\vee_{i=1}^{n} R(x,m_i) \wedge R(m_i,y)$ としたとき，条件 (5.14) 式を満たすまで，$R(x,y) \leq Max$ となる $R(x,y)$ の値を Max で置き換える．

(3) 閾値の初期値としてグレードの平均値をセットする．

(4) $R(x,y)$ から，グレードが閾値以上の要素を 1，閾値未満の要素を 0 とする行列を求める．

(5) 行番号と行のパターンの索引テーブルを作成する．一行に含まれる 1 の総数を第一キーに，一行の 0 と 1 のパターンを第二キーにして，降順に行を並べ換える．

(6) 次に列番号と列のパターンの索引テーブルを作成する．一列に含まれる 1 の総数を第一キーに，一列の 0 と 1 のパターンを第二キーにして降順に列を並べ換える．

(7) 閾値以上のグレードの数が同じものを同値類とする．

(8) 要素数が二つ以上ある同値類の数が最大となるように，閾値に変位 0.01 を加算しながら，処理 (4) ～ (7) を繰り返す．該当する閾値が複数ある場合は，値の小さい閾値による同値類を採択する．

(9) 同値類の要素数が 20 以上ある場合，要素数が 9 以上の同値類の要素を対象に処理 (1) ～ (8) を繰り返す

5.5.2 結果と考察

表 5.4は，手順 ClassNN による同値類およびその要素数の推移を表したものである．第 1 回目のクラス分けでは，Class 1 の要素数が 20 以上ある．従って手順 ClassNN 番号 9 の条件を満たすため，要素数が 9 以上の Class 1 と Class 2 を対象に 2 回目のクラス分けを行った．同様に 2 回目のクラス分けにおいても，Class 1 と Class 2 の要素数が 20 以上あるため，要素数が 9 以上の Class 1，Class 2，Class 3 を対象にクラス分けを行った．このように今回の場合，3 回のクラス分け

表 5.4 同値類の結果（99 クラス，283 語）

phase 1 : 894 nouns		phase 2 : 207 nouns		phase 3 : 97 nouns	
class No. (#)	# elements	class No. (#)	# elements	class No. (#)	# elements
1	194	1	54	1 (1)	15
				2 (1)	14
				3 (1)	7
		2	34	4 (1)	6
				5 - 6 (2)	5
				7 (1)	4
2	13	3	9	8 (1)	3
				9 - 11 (2)	2
		4 - 6 (3)	4	11 classes	65 nouns
		7 - 9 (3)	3		
		10 - 20 (11)	2		
3 - 5 (3)	6	17 classes	140 nouns		
6 (1)	5				
7 - 11 (5)	4				
12 -19 (8)	3				
20 - 73 (54)	2				
71 classes	382 nouns				

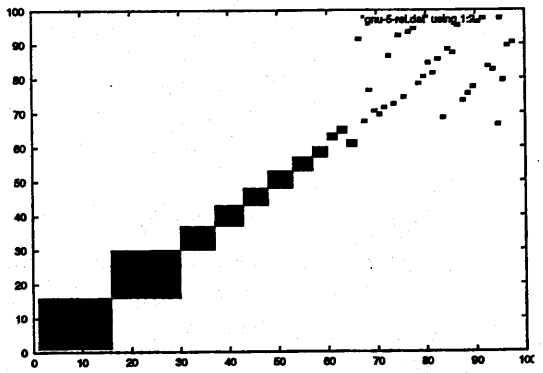

図 5.1 表 5.4 の 3 回目の分類結果 (11 クラス)

により，283 個の名詞からなる 99 個の同値類が求まった．なお，図 5.1は 3 回目の分類結果を図示したものである．

表 5.5と表 5.6は，同値類の要素について三つ組の頻度データ (c) と照合し，隣接する要素を一つにまとめ，新たな同値類を作成したものである．同値類に含まれる要素について，技術取引きに関する契約書式集に含まれる 35 個の条項毎の出現頻度データを持つ単語頻度データ (f) と照合し，出現する条項数を欄 1 に，共通に出現する条項数を欄 2 に記している．

番号 4 の “license company”，“licensee company” は，隣接関係を調べて，新たに作成した同値類である．この例は，複合語を意味するのではなく，述部を構成する要素である．現在，複合語であるか，述部の構成要素であるかを機械的に分類する仕組みはない．なお，番号 12 の “air mail”，“air letter”，番号 13 の “exchange rate”，“exchange bank”，“remittance rate” も隣接関係があり，一つにまとめた例である．これらは，複合語の例である．

一方，表 5.6は，クラスの要素内において，共通に出現する条項がない場合である．表 5.5と表 5.6を比較すると，共通する条項のない表 5.6の同値類は，類義関係のない要素となっていることがわかる．

以下に求めた同値類から名詞の類概念データを作成する手順を示す．

表 5.5 名詞の同値類 (a) 共通して出現する条項が一つ以上ある

No.		elements	1	2
1		advertisement 広告	5	4
		promotion 販売促進	9	4
2		approval 同意・承認	22	20
		consent 同意	27	20
3		breach 違反・不履行	16	5
		protection 保護	8	5
4	統	license company 許諾，会社	33	32
	統	licensee company 実施権社，会社	32	32
5		condition 条件	25	24
		term 条件・期日	29	24
6		cost 費用	21	18
		expense 費用	22	18
7		court 裁判所	8	7
		jurisdiction 管轄区	8	7
8		delay 遅延	11	4
		failure 不履行	10	4
9		discussion 討議	6	1
		indemnity 賠償	4	1
		negotiation 交渉・譲渡	6	1
		representation 陳述・代理	5	1
		understanding 協定・了解	4	1
		warranty 根拠・保証	5	1
10	等	airfare 航空運賃	1	1
		lodge 宿泊	1	1
11	反	morning 朝	1	1
		night 晩	1	1
12		air letter 航空便	5	0
		airmail 航空郵便	3	1
		cable 海底電信	4	1
		air mail 郵便	6	1
		post 郵便	2	1
		postage 郵便料金	1	1
		telegram 電報	1	1
		telex テレックス	2	1
13		excahnge bank 為替銀行	5	2
		currency 通貨	4	2
		exchange rate 為替レート	6	3
		bank rate 公定歩合	5	2
		remittance rate 送金率	5	2
		transfer 為替・振替	15	3

欄 1：出現する条項の数
欄 2：要素内において，出現する共通の条項数
<マーク>

統：統語関係
等：等位関係
反：反意関係

表 5.6 名詞の同値類 (b) 共通に出現する条項が全くない

No.	elements	1	2
17	assignor 譲渡人	1	
	belief 信頼	1	
	class 階級	3	
	remuneration 報酬	2	
18	attachment 差し押え	4	
	breakdown 破損・明細書	2	
	investigation 調査	1	
	lockout 締め出し	1	
	voucher 保証人・証拠物件	1	
19	care 管理・注意	2	
	endeavor 努力	5	
	objection 異義	2	
	repair 修理・賠償	4	
20	collaboration 協力	4	
	expert 専門家	1	
21	departure 退去	4	
	designee 被指名人	3	
	lump 一括	3	
	selection 選択	3	

＜類概念データの作成手順＞

(1) 手順 ClassNN で得られた 99 個の同値類の要素について，書式集に含まれる単語の頻度データと照合して，出現する条項名を求め，共通して出現する条項のない同値類を削除する．

該当する同値類：14 個

(2) 85 個の同値類について，三つ組頻度データと照合して，隣接する要素を一つにまとめる．その結果，同値類の要素数が 1 つになったものを削除する．

該当する同値類：14 個

(3) 71 個の同値類について，書式集のデータ (α) と照合し，同じ条文に出現する要素にマークを付ける．

(4) 人手によりチェックする．

表 5.7は，分類した結果の内訳である．

番号 1 の関連のない語が同値類となる原因は，係り受けの正しくない二つ組の影響である．二つ組の係り受けを正しくするためには，品詞タグの誤りをなくす，複合語や等位関係にある語などの関連語をまとめるなどの前処理を徹底させる必要がある．また，番号 2，3 は，複合語や等位関係にある語をまとめることで解消される．

したがって，今回の同値類の結果を基に表 1.6の (b) 連語語句のパターンの更新を行い，二つ組の係り受けの精度を高めて行くことでこれらの問題は改善される．

なお，これらの問題点は，全て類似度を求める際に生じるものであり，クラスタリングの手法における問題点は，今のところ見受けられない．

次節では，関連度の代わりに重み付き $DICE$ 係数および同時出現回数を使って類似度を求め，同値類を求めた結果との比較を行う．

表 5.7 同値類の詳細

No		# classes
1	要素間に関連がない (出現する条項に共通のものがないもの 14，意味的に関連のないもの 17)	31
2	複合語または，述部を構成する関係 (要素の数が 2 個で，二つの要素は隣接して出現する)	14
3	等位関係 (同じ条文中に出現するが，隣接関係になく，意味的な類似性はない)	8
4	述部を構成する関係 (同じ条文中に出現するが，隣接関係になく，意味的な類似性はない)	9
5	意味的に類似性がある	34
6	反意関係にある	3
		計 99 クラス

5.6 重み付き *DICE* 係数および同時出現回数から求めた類似度との比較

表5.8は，関連度をベクトルとした類似度から求めた同値類と，重み付き *DICE* 係数および同時出現回数を基にした類似度を使って，同様の手順 ClassNN で求めた同値類の比較を行ったものである．関連度による同値類 99 クラスと同時出現回数による同値類 98 クラスには，共通するクラスが 97 クラスあり，近似しているのに対し，重み付き *DICE* 係数による同値類 87 クラスとの間には，共通するクラスが 74 クラスであり，大きな相違があることがわかる．

表 5.9は，表 5.8の † 印の詳細である．すなわち，関連度と同時出現回数との同値類の相違を示したものである．“court”（法廷）と “jurisdiction”（裁判管轄）は，類義関係にあるが，“technician”（専門家）とは，類義関係はなく，かつ，出現する条項に共通のものがないことから，関連度による同値類の方が妥当である．なお，“sale price”（販売価格）は複合語である．

同様に表 5.10は，関連度と重み付き *DICE* 係数との同値類の相違の一部を示したものである．表 5.8において，関連度による同値類とは共通の要素を持たない 4 個の同値類の内，類概念データとして妥当なものが 2 個あるものの，全体的

表 5.8 同値類の比較

	$=R$	$>R$	$<R$	$\neq R$	sum	R
f_{xy}	97	1†	0	0	98	1†
(妥当な同値類の数)		R1				
$WDICE$	74	2	7	4	87	18
(妥当な同値類の数)		R1	R5	D2		R8

$=R$：関連度による同値類と，要素および要素数が同じ同値類の数
$>R$：関連度による同値類と比較して，要素数が関連度の同値類より多い同値類の数
$<R$：要素数が関連度の同値類より少ない同値類の数
$\neq R$：該当する同値類が関連度による同値類にはない
R ：関連度による同値類にのみ存在する同値類の数
Dn：重み付き DICE 係数における妥当な同値類の数
Rn：関連度における妥当な同値類の数

表 5.9 関連度と同時出現回数の比較

	f_{xy}		REL	
	要素	条項頻度	要素	条項頻度
$>R$	court	8	court	8
	jurisdiction	8	jurisdiction	8
	technician	4		
	共通の出現条項数	0	共通の出現条項数	7
R			price	17
			sale	25
			共通の出現条項数	16

注：表 5.8の † 印の同値類の詳細

表 5.10 関連度と重み付き *DICE* 係数の比較

		WDICE		*REL*	
NO.		要素	条項頻度	要素	条項頻度
1	$< R$	air	5	air	5
		airmail	3	airmail	3
		cable	4	cable	4
		post	2	letter	14
		postage	1	mail	6
		telegram	1	post	2
				postage	1
				telegram	1
				telex	2
		共通の出現条項数	1	共通の出現条項数	1
2		bank	5	bank	5
		currency	4	currency	4
		exchange	6	exchange	6
		rate	13	rate	13
				remittance	6
				transfer	15
		共通の出現条項数	2	共通の出現条項数	3
3	$> R$	allowance	5	insurance	6
		insurance	6	set	23
		line	3		
		set	23		
		共通の出現条項数	2	共通の出現条項数	4
4		amendment	7	amendment	7
		locality	1	extension	16
		optionor	0		
		proprietor	7		
		umpire	1		
		共通の出現条項数	0	共通の出現条項数	2
5	$\neq R$	account	14		
		book	5		
		record	11		
		共通の出現条項数	3		
6		benefit	16		
		director	8		
		共通の出現条項数	5		
7	R			condition	25
				term	29
				共通の出現条項数	24
8				cost	21
				expense	22
				共通の出現条項数	18

には，関連度による同値類が99クラスに対して重み付きDICE係数が74クラスと同値類の数が少ないこと，表5.10にみるように妥当な同値類が関連度の方に多いことから，本章で提案する類似度の尺度としては，関連度の方が適している．

5.7 まとめ

本章では，

(1) 品詞の異なる二つ組の関連度の提案

(2) 関連度を用いた類似度の提案

(3) ファジィ関係行列を用いたクラスタリングの提案

(4) クラスタリングの結果を用いた類概念データの作成手順の提案

について述べた．

関連度においては，「契約書式集から求めた統語関係の正しい二つ組の出現回数は，実際の契約書の条文を解析する際の貴重な情報となる」との考えから，二つ組の同時出現回数に重きをおいた関連度を提案し，重み付きDICE係数との比較を行った．

類似度に関する評価については，当初，電子化されている *Roget's Thesaurus* との照合を考え，調査を行った．しかし，契約文書に含まれる語の意味が制限されていること (例えば，契約書における 'party' という語は，「当事者」という意味でのみ使われる)，専門用語が多く使われていることから，*Roget's Thesaurus* の類義語クラスに該当する語が含まれない場合や，*Roget's Thesaurus* の類義語クラスの要素が，契約文書用の類義語クラスとして適切ではない場合があり，*Roget's Thesaurus* を使った評価は適切でないとの結論に至った．そこで，類似度については，今後の研究において類似度データを利用していく中で評価を行う予定である．

なお，本研究で提案した類似度計算は，複合語や慣用句を一つにまとめる前処理を施した品詞タグ付きのテキストデータを対象としている．本手法は，疑問文や命令文，および感嘆詞などの冗長な語句が多く挿入された英文には不向きであるが，使用される語句がある程度限定されている前処理済みの平叙文には適用できると考えている．

第6章
条文の統語構造の抽出

本章では，条文から統語構造を抽出する手法の提案を行う．平叙文からなる契約文書における「統語構造の抽出」とは，(1) 修飾・被修飾関係の抽出，(2) 主部・述部の抽出をいう．

本手法は，条文を意味のある語のまとまり (本研究ではこれを「構成語句」という) に分割し，品詞および語義を表す分類タグを付加する．更に，構成語句間の関連付けを行い，関連付けられた構成語句のまとまりに，統語構造を注釈の形で付加する．

一般に，統語構造を抽出する処理を統語解析という.「統語解析」は，文法規則と辞書を用いて文の統語構造を明らかにする．すでに,「統語解析」に関する多くの提案がなされている．文献 [45][46][47] などは「統語解析ツール」の一部である．

しかしながら，自然言語を対象とした統語解析において，文法，辞書共にその形式 (記述方法) および記述すべき内容は定まっていない．

第2章の表2.1で示したように，契約文書の条文は長文で複雑である．また，文書の構造化を行う際，たとえばXMLなどのタグ付けを行う場合の統語解析において，1本の解析木を求めることは必ずしも必要ではない．むしろ，修飾語の修飾する範囲を限定するなどの部分的な解析結果の正しさを保証することの方が重要であると思われる．そこで本手法は，動詞を中心とするパターン情報と，文法規則を併用して7種類の統語構造の抽出を行う．

表 6.1 解析の手法

最小の単位	処理	本研究での処理	最小の単位
単語	(1) 形態素解析	(A) 品詞タグ付け	
	↓	↓	
		(B) 品詞タグの修正	
		↓	
		(C) 語句の連結	
	(2) 統語解析	↓	
	↓	(D) 構成語句分割と分類タグ付け	構成語句
		↓	
	(3) 意味解析		
		(E) 構成語句間の関連付けと注釈付け	
	↓	↓	
	(4) 構造化	(F) 構造化	

6.1 解析の手法

表 6.1は，自然言語処理における一般的な解析手法を組み合わせて条文の解析を行う場合と，本研究で提案する手法とを示したものである．

(1) ～ (4) の処理は互いに独立している．そして，(1) で条文を単語 (形態素) に分割した後，(2) で単語の品詞情報と統語規則を用いて，単語→句→節というように関連付けがなされる．(3) では，(2) で求めた単語間の関係について意味的整合性のチェックを行う．そして (4) では求めた解析結果を参考にして，人手によってタグ付けを行い，構造化文書を作成する．

一方,「契約文書に含まれる語句はある程度限定されている (表 2.1参照)」という特徴を生かした本手法では，(C) の段階で一部の複合語や慣用句の連結を行う．現在，403 種の連結規則により，条文にある語句を約 90 %にまとめ上げている (表 6.6(C) の平均比率を参照のこと)．契約書に頻出する等位語句や慣用句など

を (C) で連結処理することで，(E) で生じる構造的曖昧性を減らすことができる．(D) では，異なる機能を持つ単語間の境界を見つけて，条文を意味のある語句のまとまり (構成語句) に分割する．その後，構成語句間の関連付けを行い，注釈を付加する．本章では (D) および (E) について述べる．

6.2 用語の定義

本章で用いる語句を以下に定義する．

群動詞 ＜動詞＋副詞＞，＜動詞＋前置詞＞などの形で，全体で一つの動詞と同じ働きをするものをいう．句動詞 (phrasal verb) と同義．

群前置詞 2 語以上が集まって一つの前置詞として働くものをいう．

句 複数の単語が集まって，名詞，形容詞，副詞に相当する働きをするものを一般に句という．本章ではさらに，群動詞，動詞と他の語が慣用的に結びついた動詞を含む慣用語句，動詞に助動詞が付加されたもの，および完了時制，進行形，受動態を作るために “have”，“be” が付加された複合動詞，を併せた**広義の句**を**句**という．

構成語句 意味的にまとまった単語列をいう．

単語，構成語句，ならびに句間の包含関係は，以下のとおりである．

単語 $\subseteq$ 構成語句 $\subseteq$ 句

主辞 構成語句または句の中で，中心となる部分をいう．以下の構成語句において太字の部分が主辞である．

名詞 ： **prior written consent**

動詞 ： **is received**

Modifier 時間，場所，条件などの各種限定条件を表現する修飾語，修飾句，修飾節をまとめて Modifier という．

構造的曖昧性 構成語句間の関係を表す階層的構造が複数ある場合，これを「構造的に曖昧である」という．

相関接続詞 不連続に共起する一対の語句が対応して一つの接続詞として機能するものをいう．相関接続詞は等位相関接続詞と従属相関接続詞に大別される．

例 “both **A** and **B**” ； A，B は構成語句，句，または節

述部　平叙文は，(1) 主題となる部分「…が，…は」と，(2) 主題について述べる部分「…する，…である」から成り立っている．この (1) の部分を主部，(2) の部分を述部という．

文法機能的制約　主語と動詞間の人称や数の一致，時制の一致などをいう．

helping verb　権利または義務を表す助動詞などの語句をいう．詳細は 2.3節を参照のこと．

6.3　構成語句への分割

本研究では，単語から構成語句をまとめ上げるのではなく，異なる機能を持つ構成語句間の境界を見付けて分割を行う．具体的には条文を，

- 助動詞および前置詞の前後，
- 代名詞および冠詞の後，
- 「形容詞＋前置詞」の前，
- 連続する名詞と動詞，または動詞と名詞の間，
- 関係代名詞および関係副詞の前後，
- 特殊記号の前後，
- 等位接続詞の前後，

で分割し，構成語句を求める．

構成語句への分割は以下の手順で行う．

(1) 構成語句への分割に先立ち，表 6.1の処理 (A) ～ (C) までの前処理を行う．

(2) 次に，“not” 否定と動詞，または “not” 否定と比較級の連結，重要な意味を持たない副詞の削除を行う．

(3) 「94 種類の分割規則 (付録 B.3 参照のこと)」を用いて，構成語句への分割を行い，「注釈変換データ (付録 A.1 参照のこと)」を参照して，分類タグを付加する．

契約文書を理解する上で，修飾語と被修飾語の関係は，重要な問題となる．本研究では，“a” などの限定詞や，“any”，“all”，“each” などの不定代名詞，“this”，“such” などの指示代名詞の形容詞的用法については，一般の形容詞より被修飾の範囲が広くなるように構成語句の分割を行っている．例えば以下の条文において，

“...to use all or part of the trademark on any natural food product, beverage or nutritional supplement”[14]$^{pp.79}$

表 6.2 解釈 1

	(1)	(2)	構成語句
1		m	any
2	:	:	natural food product
3	:	:	beverage
4	c	:	or
5	:	:	nutritional supplement

注) “:” は，影響する構成語句を表す．
(1) 等位関係においては，“or” の影響を受ける構成語句が番号 2，3，5 であることを示している．また，(2) 形容詞との関係においては，“any” が修飾する構成語句が番号 2，3，4，5 であることを意味する．

表6.2と 6.3の構造的曖昧性が生じるが，構成語句の分割結果は表6.2となる．なお，表6.2の解釈 1 は表 6.3の解釈 2 より商標 (trademark) の使用範囲を広げる結果となる．

表 6.3 解釈 2

	(1)	(2)	構成語句
1		m	any natural
2	:	:	food product
3	:	:	beverage
4	c	:	or
5	:	:	nutritional supplement

本手法の汎用性を確認するために *Wall Street Journal* の品詞タグ付きコーパスからランダムに抽出した文例30種(725語)について，前処理表6.1(A) ～ (C)を施し，705種の語句にまとめたものを構成語句に分割した．その結果得られた594種の構成語句は,「構成語句⊆句」の関係を満たしていた．

表6.4，表6.5の処理5は，条文を構成語句に分割した結果である．

6.4 分類タグ

分割した構成語句に，品詞情報と語義を表す分類タグを付加する．分類タグの形式は以下のとおりである．

(1) 「構成語句の品詞－主辞」: 表6.4の番号1

(2) 「構成語句の品詞－主辞の代表」

(3) 「構成語句の品詞－主辞の概念」: 表6.4の番号25

語義は，語句と語義の対応表である「注釈変換データ」を用いて求める．等位接続により複数の主辞がある場合は，代表語を指定する．また，注釈変換データに主辞の概念を表す語が指定されている場合は，その概念を指定する．なお，概念は大文字で表す．表6.4の番号25は，複数の主辞があり，かつ概念の指定がある場合の例である．(2)については表6.4，表6.5に該当する例がないが,「主辞の代表」とは，等位接続で複数の名詞が列挙されているとき，注釈変換データに指定がある場合はそれを，ない場合は等位接続詞の直前の語を意味する．

6.5 構成語句間の関連付け

構成語句に分割し，分類タグを付加したデータを基に，以下に示す構成語句間の関連を求める．

(a) 等位関係

(b) 修飾と被修飾の関係

(c) 前置詞と構成語句の関係

<例文 1 (72 語) >

Technical Information including all information, data, specification, drawing or any of them given by Company shall be considered absolutely confidential and shall not be disclosed to any other person, firm or corporation by Licensee without a prior written consent of Company, and Licensee shall require its employees to treat such matters as confidential not only during the life but also after the termination of this Agreement[7].

表 6.4 処理 5 と処理 6 の解析結果 1

No.	処理 5： 構成語句分割		処理 6： 構成語句関連付け										
	分類タグ	構成語句	分類タグ	1	2	3	4	5	6	7	8	9	10
1	NP-information	Technical=Information/NN	NP-information										S
2	VBG-including	including/VBG	VBG-DO								vm		
3	JJ-all	all/JJ	NP-information			m					d		
4	NP-information	information/NN	+	:		:					d		
5	SYM	,/SYM	+	:		:					d		
6	NP-data	data/NNS	+	:		:					d		
7	SYM	,/SYM	+	:		:					d		
8	NP-specification	specification/NN	+	:		:					d		
9	SYM	,/SYM	+	:		:					d		
10	VBG-drawing	drawing/VBG	+	:		:					d		
11	CC-or	or/CC	+	c		:					d		
12	PRP	any_of_them/PRP	+	:		:					d		
13	VBN-given	given/VBN	VBN-M									vm	
14	PP-by	by/IN	BY-PERSON				p					m	
15	NP-PERSON	Company/NNP	+				:					m	
16	MD-shall	shall/MD	MD-shall										H
17	VBV-considered	be_considered/VBV	VBV-considered									v	V
18	RB-absolutely	absolutely/RB	RB-absolutely										
19	JJ-confidential	confidential/JJ	JJ-confidential									c	C
20	CC-and	and/CC	CC-and										
21	MD-shall	shall/MD	MD-shall										H
22	VBV-not_disclosed	not_be_disclosed/VBV	VBV-not_disclosed								v		V
23	TO	to/TO	TO-PERSON				p				i		I
24	JJ-any	any/JJ	+			m	:				i		I
25	NP-PERSON	other/JJ person_,_firm_or_corporation/NN	+			:	:				i		I
26	PP-by	by/IN	BY-PERSON				p				m		M
27	NP-PERSON	Licensee/NN	+				:				m		M
28	PP-without	without/IN	PPNP-without-consent				p	H					
29	JJ-a	a/JJ	+			m	:	H					
30	NP-consent	prior/JJ written/JJ consent/NN	+			:	:	H					
31	PP-of	of/IN	PPNP-of-PERSON				p	:					
32	NP-PERSON	Company/NNP	+				:	:					
33	SYM	,/SYM	SYM										
34	CC-and	and/CC	CC-and										
35	NP-PERSON	Licensee/NN	NP-PERSON										S
36	MD-shall	shall/MD	MD-shall										H
37	VB-require	require/VB	VB-require								v		V
38	PRP$	its/PRP$	NP-PERSON			m					i		I
39	NP-PERSON	employees/NNS	+			:					i		I
40	TO	to/TO	TOVB-treat								Id		D
41	VB-treat	treat/VB	+								vd		D
42	NP-matters	such/JJ matters/NNS	NP-matters								dd		D
43	PP-as	as/IN	PP-as								cd		D
44	JJ-confidential	confidential/JJ	JJ-confidential								cd		D
45	RB-not_only	not_only/RB	RB-not_only						c				
46	PP-during	during/IN	PPNP-during-life				p		:				
47	JJ-the	the/JJ	+			m	:		:				
48	NP-life	life/NN	+			:	:		:				
49	CC-but_also	but_also/CC	CC-but_also						c				
50	PP-after	after/IN	PPNP-after-termination				p	H	:				
51	JJ-the	the/JJ	+			m	:	H	:				
52	NP-termination	termination/NN	+			:	:	H	:				
53	PP-of	of/IN	PPNP-of-agreemant				p	:	:				
54	NP-agreemant	this/DT Agreement/NNP	+				:	:	:				
55	SYM	./SYM	SYM										

＜例文 2 (25 語) ＞

This Agreement shall also apply to permits of every kind granted by the state of California or a political subdivision or local authority thereof[14].

表 6.5 処理 5 と処理 6 の解析結果 2

No.	処理 5： 構成語句分割		処理 6： 構成語句関連付け									
	分類タグ	構成語句	分類タグ	1	2	3	4	5	6	7	8	9	10
1	NP-agreemant	This/DT Agreement/NNP	NP-agreemant										
2	MD-shall	shall/MD	MD-shall										
3	VB-apply	also/RB apply/VB	VB-apply										
4	TO	to/TO	TONP-permits				p	H					
5	NP-permits	permits/NNS	+				:	H					
6	PP-of	of/IN	PPNP-of-kind				p	:					
7	JJ-every	every/JJ	+			m	:	:					
8	NP-kind	kind/NN	+			:	:	:					
9	VBN-granted	granted/VBN	VBN-M									vm	
10	PP-by	by/IN	PPNP-by-state				p	H				m	
11	JJ-the	the/JJ	+			m	:	H				m	
12	NP-state	state_of_California/NNP	+			:	:	H				m	
13	CC-or	or/CC	+		c		:	:				m	
14	JJ-a	a/JJ	+		:	m	:	:				m	
15	NP-subdivision	political/JJ subdivision/NN	+	:	:	:	:	:				m	
16	CC-or	or/CC	+	c	:	:	:	:				m	
17	NP-authority	local/JJ authority/NN	+	:	:	:	:	:				m	
18	SYM	./SYM	SYM										

(d) 相関接続の関係

(e) 前置詞を含む構成語句と被修飾構成語句の関係

(f) 述部を構成する構成語句の関係

(g) 単文の抽出

等位関係の対象には，単語，句，節がある．すなわち，(a) は他の関係，例えば (b) と相互関係がある．したがって，これら 7 種の関係を求めるプログラムが常時巡回し，各プログラム間で同期を取りながら該当箇所を見付けて処理を行うというのが，理想の運用方法であるが，現実に実装することは困難である．

そこで，優先順位の高い関係から順に構成語句間の関連付けを行う．具体的には,「単語＝構成語句」である場合の関係付けを最優先させる．すなわち，まず (a) 等位関係を調べ，次に (b) 形容詞と構成語句の関係を調べる．最も優先順位の低いものは (g) の単文の抽出であり，(f) 動詞を主辞とする構成語句を含む述部が求まった場合に実行される．また，(b) の関連付けを行った後に，再度 (a) の関連付けを行うなど，人為的に 7 種の関連付けを組み合わせている．

各関係の最適な組み合わせについては，実験を行い検討している段階である．最適な組み合わせが見付かった時点で，構成語句間の状態を調べ，各関係を求めるプログラムを起動させる巡回プログラムを作成する．

以下に各関係について詳述する．

(a) 等位関係

文法上対等の関係にある語と語，句と句，節と節を求める．具体的には，“and”，“or”，“and/or” で接続される対象を 2 種類の手法を用いて求める．

すなわち，最初に (1)「等位関係データ」を参照し，該当データがない場合，(2) 構成語句に付加された分類タグの品詞情報から等位関係を求める.「等位関係データ」とは，本研究で対象としている技術取引に関する書式集 [7] から作成した 71 種の等位関係にある語句のリストをいう (付録 A.3 参照のこと)．表 6.4処理 6 の欄 1 の番号 4 ～ 12 は，(a) の実行結果である．番号 10 の品詞は動詞の現在分詞 “VBG” となっている (品詞タグ付けの誤り) ため，分類タグの品詞情報では，接続対象にはならない．このようなケースを回避するために，本研究では「等位関

係データ」を作成し，分類タグの主辞部と「等位関係データ」の照合結果を優先して，接続対象データを求めることで関連付けの誤りの低減を図っている．

表6.5の欄1の番号15～17は，(a)の実行結果である．欄2の番号13～17は，(b)の直後に実行した(a)の結果である．

等位関係の対象は，単語，構成語句，句，節と範囲が広い．したがって，他の関連付けを行った後にそれらの結果を参照しながら，等位関係を求める必要がある．

表6.4の番号20は述部の等位接続であり，番号34は節の等位接続である．これらの等位接続の前後にある分類タグのパターンが異なることから，現在作成している等位関係を求めるプログラムでは処理がなされない．すなわち，現在のところは「等位関係データ」に登録されているか，または分類タグの品詞情報が同じである構成語句を対象に等位関係を求めている．

(b) 形容詞と構成語句の関係

修飾語が修飾する被修飾語の範囲を決定する．修飾語として，冠詞，代名詞，および形容詞を，被修飾語として名詞を対象としている．関係を求める際，(a)より付加された注釈を参照する．すなわち，修飾語に後続する語句に，等位関係にある語句がある場合，これら等位関係にある語句の全てを被修飾語とする．表6.4および表6.5処理6の欄3において注釈“m”は修飾語を，以下の“:”は被修飾語の範囲を示している．

(c) 前置詞と構成語句の関係

前置詞と後続する名詞句とを連結する．表6.4，表6.5処理6の欄4が実行結果である．

関係を求める際，(a)(b)の処理で付加された注釈を参照する．

(d) 前置詞を含む構成語句と被修飾構成語句の関係

(c)で求めた前置詞を含む構成語句が修飾する範囲を決定する．

表6.4，表6.5処理6の欄5が実行結果である．“H”は，主辞を表す．

関係を求める際，(a)(b)(c)の実行により付加された注釈を参照する．

(e) 相関接続の関係

相関接続詞で接続される対象を求める．相関接続詞として，以下のものを対象としている．

(1) not only **A** but also **B**

(2) both **A** and **B**

(3) either **A** or **B**

(4) neither **A** nor **B**

(5) not **A** but **B**

A および **B** には，単語，句，節の全てが対象となるが，本手法では **A** および **B** が単語または句である場合を対象にしている．

表 6.4処理 6 の欄 6，番号 45 ～ 54 が本プログラムの実行結果である．

(f) 述部を構成する構成語句の関係

本手法では，以下の 2 種類の方法を使って述部を求める．

(1) 動詞と後続する語句の情報を記述したテンプレートを参照する方法

テンプレートについては，付録 A.4 を参照のこと．

(2) 文法規則に文法機能的制約および経験則を併用する方法

具体的には，まずテンプレートとの照合を行い，該当する箇所がない場合は，文法規則を用いる．

表 6.4処理 6 の欄 8 はテンプレートを用いた結果であり，欄 9 は文法規則を用いた結果である．なお，“v” は動詞，“vm” は動詞の形容詞的用法，“c” は補語，“d” は直接目的語，“i” は間接目的語，“m” は Modifier を表す．

本手法では，一般に学校文法とよばれる基本 5 文型を用いている．したがって，テンプレートに登録されていない群動詞については処理がなされない．述部を求める処理は，動詞が見つかるまで入力テキストを right_to_left 方向に走査し，動詞を発見すると，その動詞から left_to_right に後戻りをしながら処理を行う．従って，表 6.4の番号 41 は，番号 37 より先に処理され，番号 37 が処理される際に，述部番号 41 ～ 44 は番号 37 の直接目的語に決定される．

(g) 単文の抽出

表 6.6 処理の推移

	例文 1(72 語) 出力結果	例文 2(25 語) 出力結果	契約文書† 出力結果	平均 比率
(B)	72 語	25	3,496	1.00
(C)	63 語	23	3,061	0.89
(D)	55 構成語句	19	2,693	0.76
(E)	32 構成語句	11	1,872	0.47
time	1.2 秒	0.9 秒	95.6 秒	

注：契約文書とは，文献 [7] に含まれる「ノウハウライセンス契約書」26 条項 124 条文 3,496 語

(f) の結果得られた注釈と分類から，主節および従属節を求める．

表 6.4の欄 10 は，本プログラムの実行結果である．具体的には，欄 8 と欄 9 を参照し述部を決定する．さらに述語動詞の前に出現し，欄 8 と欄 9 に注釈のない名詞句の主辞を主語とする．

なお，“S” は主語，“H” は helping verb を表す．

6.6 評価

表 6.6は，平均より長い条文 (72 語) と短い条文 (25 語) および，書式集に含まれる「ノウハウライセンス契約書」26 条項 124 条文 (3,496 語) を対象に，表 6.1 の処理 (B) 品詞タグの修正，(C) 語句の連結，(D) 構成語句への分割，(E) 構成語句間の関連付け，の各処理における語数および構成語句の数の推移を示したものである．この結果 (C) の語句の連結で，解析処理の対象を約 89 %に縮小できることがわかる．

なお使用機種は，IBM ThinkPad 535E，150MHz，72MB で，プログラム言語は Perl5.00404．測定時間は，各々の処理を 5 回実行した平均の値である．

表 6.7は，テンプレートに 5 種類のパターンが登録されている動詞 “disclose” を含む 39 種の条文 (2,294 語) について構成語句間の関連付けを行った結果である．

(a) 等位関係を求める処理において，節間および分類タグのパターンが異なる複雑な句間の等位関係については処理されないため，処理件数が 40 %と低くなっ

表 6.7 処理結果の評価

処理	総数	処理件数	内正解数	正解率
(a) 等位関係	172	68(40 %)	61(90 %)	35 %
(b) 修飾-被修飾関係	261	215(82 %)	209(97 %)	80 %
(c) 前置詞＋名詞句	352	261(74 %)	247(95 %)	70 %
(d) 前置詞句の修飾範囲	64	48(75 %)	42(87 %)	66 %
(e) 相関関係	3	3(100 %)	2(67 %)	67 %
(f) 述部の抽出	182	115(63 %)	89(77 %)	49 %
(g) 単文の抽出	101	70(69 %)	28(40 %)	28 %

注：動詞 "disclose" を含む 39 条文 (2,294 語).
総数：人手により求めた正解の数.
処理件数：注釈が付加された件数. () 内は処理件数 / 総数.
内正解数：注釈が付された件数の中の正解の数. () 内は内正解数 / 処理件数= 適合率.
正解率：内正解数 / 総数 = 再現率.

ている．正しく処理されない例としては，前置詞を含む句間の等位関係，異なる機能を持つ等位接続詞が近接している場合，動詞の形容詞法と形容詞との等位関係などがある．これらについては，等位関係データを充実することで精度を上げることができる．その他にプログラム上の問題として，単語間の等位関係において接続詞の直後に "other ＋名詞" がある 6 種類について処理がなされなかった．

(b) 修飾と被修飾の関係を求める処理においては，名詞の後ろから修飾する場合と動詞の形容詞法で品詞が形容詞になっていない場合は処理がなされない．また，等位関係が正しく求まっていない場合は正しい注釈が付加されない．

(c) 前置詞と構成語句の関係を求める処理において，(a)(b) の注釈が正しく付加されていれば，前置詞と名詞句については注釈が正しく付加される．一方，群前置詞または二重前置詞と名詞句については処理がなされない．これらは，群前置詞などを連結処理することで改善することができる．その他に前置詞と名詞節，形容詞，副詞，動名詞，および不定詞については処理をしない．

(e) 相関接続の関係を求める処理においては，接続の対象となる名詞句の構造が異なる場合に注釈が正しく付加されない．

(d) 前置詞を含む構成語句と被修飾構成語句の関係，(f) 述部を構成する構成語句の関係，(g) 単文の抽出においては，(a)(b)(c)(e) における処理結果が反映されるため，正解率が低くなっている．(f) の述部を求める処理において，テンプレートに記載されていない場合，Modifier のかかり先が決定されないことがある (表6.4処理 6 の番号 28 ～ 32，45 ～ 55).「群動詞」については学校文法では対処できないため，これらを網羅するテンプレートを作成する必要がある．

6.7 まとめ

本章では，長文で複雑な契約文書の条文を解析する手法について説明した．本解析手法は，契約文書の構造化を支援するものであり，(1) 品詞タグ付け，(2) 慣用句などの連結，(3) 構成語句への分類，(4) 構成語句間の関連付けの四つの処理より解析を行う．(4) の関連付けについては，7 種類の関係を部分的な解析プログラムを用いて，該当する統語構造が見つかった場合に処理をし，見つからなければ何もしないという方法で処理を行う．本手法は，規則と参照データが揃っていれば，条文の長さに因われることなく解析結果を得ることができる．ただし，現在のところ，規則と参照データは，独自で作成しているため充実していない．

今後の課題として，これらの規則および参照データの充実とともに，7 種類の処理の相互関係を明らかにし，インタラクティブに処理操作および注釈・分類タグの変更ができるエディタの作成がある．

第7章
本論文のまとめ

本論文では，契約文書の特徴を説明し，契約文書の電子文書化に関する以下の研究について述べた．

(1) 情報検索・内容抽出の効率向上に関する研究

(a) 「語の基本形」の提案

語から語形を変化させる屈折接辞を取り除いた「語幹」から，語の品詞を変える派生接辞を取り除いたものを「語の基本形」と定義した．「基の語」,「語幹」,「語の基本形」の3種の重要語を使った条項名の推定実験の結果,「語の基本形」は,「基の語」の53％のデータ量で1.06倍の正解率を得た．すなわち，少ないデータ量で同義の多様なテキストデータに柔軟に対応できることを示した．

(b) 重要語の定義と重要語の抽出

契約文書は条項ごとに内容がまとめられているという特徴を活かして，技術取引に関する契約書式集に含まれる34種類の条項における重要語の一覧を作成し，テスト用の条文に含まれる重要語の重要度の累計から，条項名の推定実験を行った．その結果,「基の語」を用いた重要語一覧では80％,「語の基本形」を用いた場合は85％の正しさで推定できることがわかった．

また，正しく推定できなかったテストデータを検討した結果，テストデータの内容が複数の条項に関連する内容であったり，関係のない内容である場合に正しく推定されないことがわかった．すなわち，書式集など模範となるデータから抽出した重要語とその重要度は，実際の国際取引において作成される契約文書の特徴抽出に利用できることがわかった．

(c) 語間の類似度計算とファジィ関係行列を用いたクラスタリングの提案

類義語辞書 (thesaurus) は情報検索を行う際に重要なデータとなる．そこで，類義語辞書を作成する際の人的作業量を減らすために，関連度を用いた類似度計算の提案を行った．

例えば名詞間の類似度は，(名詞，動詞)，(動詞，名詞)，(形容詞，名詞)，(前置詞，名詞) の二つ組の各々において，共通の組の関連度をベクトルとした内積の余弦を求め，全体の組に対する共通の組の比率を掛け合わせた後，重み付けを行い，求めた 4 種の値を加算したものとなる．

このようにして求めた 894 種の名詞の類似度を基に 894 × 894 相似関係行列を作成し，ファジイ同値関係を導き，全体集合 894 を同値類に分解する手法の提案を行った．本手法により 99 個の同値類を求め，人手により選別を行い，34 種の類義クラスを作成した．

(2) 契約文書の電子文書化 (構造化) に関する研究

(a) 共起する語の二つ組の関連度

文法の上で複数の係り受けが考えられる場合に妥当な係り受けを決定するために関連度を導入した．関連度とは，語の二つ組の同時出現回数に，各単語の汎用性の相違を重みとして掛けたものである．

(b) 契約文書における統語構造の抽出

構造化された文書のタグ付けに関しては，SGML(Standard Generalized Markup Language)， XML(eXtensible Markup Language) がある．また文書に含まれる構成語句の統語構造，語義，参照関係など構造化に必要なタグを XML 形式に従って定義した GDA(Global Document Annotation)[49] なども提案されている．

しかしながら，これらのタグを付加するためには，何らかの統語解析および意味解析が必要であり，現在のところ統語解析および意味解析の研究は発展途上の段階にある．

そこで本研究では，契約文書の構造化を目的とした解析手法の提案を行った．

本手法は，(1) 文法規則と動詞を中心とした抽出パターンとを併用して処理を行う，(2) 以下に示す7種類の統語構造の抽出をそれぞれ独立したプログラムが担い，各プログラムが自己の担当部分において該当箇所を見つけた場合に処理を行うという二つの特徴を持つ．

i. 等位関係

ii. 形容詞と構成語句の関係

iii. 前置詞と構成語句の関係

iv. 相関接続の関係

v. 前置詞を含む構成語句と被修飾構成語句の関係

vi. 述部を構成する構成語句の関係

vii. 単文の抽出

現在，試作の段階であり，文法規則も完全ではなく，また，抽出パターンも秘密保持条項に出現する 57 種の動詞に関する 143 パターンのみであるが，(1) 条文の長さに関係なく，結果を出力すること，(2) 出力した結果はテキストデータであり，人手による見直し修正が可能であること，(3) 出力された結果から，解析に用いる緒規則・データを抽出することで解析の精度を上げることができるなどの利点を持つ．

近年，英語を母国語としない人達が，異なる文化や慣習を背景に持ちながら，英語を媒介として意思の疎通を図る機会が増えてきている．当事者双方が英語を母国語としない場合においても，「公平にリスクを負う」という意味で，英文契約書を交わすことも頻繁に生じている．その結果，英語を母国語とする人達には，思いも付かない解釈が生まれ，争いとなるケースもある．

筆者は，英語が母国語ではない日本人を対象にした契約文書の電子文書化を考える際，これらの問題となりそうな解釈を提示する必要性を感じ，本研究に取り組んできた．

その結果今回，文法規則，単文パターン，語句の等位関係，連語を表すパターンなど，異なる複数の諸規則・データを用いた，繁雑な解析手法を提案することとなった．しかしながら，多様な規則をそれぞれ独立したプログラムが担うことで，意図的に多様な解釈の提示が可能となった．

今後，英語を母国語とする研究者らによる統語解析および意味解析に関する研究成果を積極的に流用しながら，多様な規則を併用した本手法を発展させ，日本人を対象にした契約文書のテキストデータベースの構成および運用に関する研究へと進化させて行きたいと考えている．

謝辞

主指導教官の渡邉勝正教授には，博士課程在学中の6年間にわたりご指導を賜りました．研究内容や進め方を親身になってご指導いただいただけではなく，研究者としてのあり方を示していただきました．心より感謝致します．

松本裕治教授には，お忙しい中，副指導教官となっていただき，また，Penn Treebank WSJ.ATIS.Brown Corpus をはじめ，自然言語処理講座所有の貴重な資料をお貸しいただき，更に審査委員もおひきうけ下さいました．植村俊亮教授もまた，ご多忙の中，審査委員に加わって下さいました．両教授には深く感謝致します．

吉川正俊助教授には，研究の要所要所で貴重な助言をいただき，また貴重な資料を貸していただきましたことを，感謝致します．

流通科学大学の向 高男教授には，国際取引の基礎知識ならびに英文契約書の作成方法，解釈の仕方など，多くのことをご教示いただきました．ここに感謝の意を表します．

最後に研究分野が異なるにも関わらず，研究の場を提供して下さった木村晋二助教授，ならびに名古屋大学工学部の高木一義助手と渡邉研究室の皆様に感謝します．

参考文献

[1] 通商産業省：企業活動のグローバライゼーション,
http://www.miti.go.jp/report-j/g82-1j.html

[2] 通商産業省：海外事業活動基本 (動向) 調査,
http://www.miti.go.jp/stat-j/h2c4topj.html

[3] 吉原英樹：国際経営, 有斐閣アルマ (1997).

[4] 高柳一男：国際プロジェクト契約ハンドブック, 有斐閣 (1987).

[5] コンピュータ・エージ社：情報サービス産業白書,
http://www.jisa.or.jp/activity/index-whitepaper-j.html

[6] 山口尚夫:情報化時代の文書管理 ー経営事務と文書情報ー, 嵯峨野書院 (1993).

[7] 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).

[8] 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).

[9] Eric Brill: Some Advances in Transformation-Based Part of Speech Tagging, AAAI-94(1994). http://www.cs.jhu.edu/～ brill/acadpubs.html

[10] 長谷川俊明：法律英語のプロ - 契約と裁判, 東京布井出版 (1992).

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

[12] 小中信幸監修, 仲谷栄一郎：契約の英語, 若竹出版 (1995).

[13] 山本孝夫 ：英文契約書の書き方, 日本経済新聞社 (1995).

[14] 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).

[15] 片山善行：国際ライセンス契約の実務, 中央経済社 (1996).

[16] 中島憲三：英文ライセンス契約書の書き方, 民事法研究会 (1993).

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

[18] 岩崎一生：英文契約書 - 作成実務と法理 -, 同文舘 (1999).

[19] 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).

[20] 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).

[21] 長谷川俊明：法律英語のカギ - 契約・文書・術語 - 法律英語・パート 1 , 東京布井出版 (1985).

[22] 相良かおる, 渡邉勝正 ：英文契約書における内容の抽出 – 条文における権利・義務記載部分の内容を表す句構成セットの作成 –, 情報処理学会研究報告, 99-NL-130, pp.129-136(1999).

[23] 宮川幸久, 綿貫 陽, 須貝猛敏, 高松尚弘：徹底例解 ロイヤル英文法 - Royal English Grammar with Complete Examples of Usage - , 旺文社 (1999).

[24] 安井稔編：コンサイス英文法辞典, 三省堂 (1996).

[25] 大石強：現代の英語学シリーズ 4 形態論, 開拓社 (1994).

[26] William B.Frakes,Ricardo Baeza-Yates:Information Retrieval − Data Structures & Algorithms −,pp.131-160, Prentice Hall PTR(1992).

[27] The ACL NLP/CL Universe(1999),
http://www.cs.columbia.edu/ radev/u/db/acl/html/AREA/IR/

[28] 野村浩郷：自然言語処理の基礎技術,pp.5, 電子情報通信学会 (1988).

[29] 酒井玲子：英語語形成ルール・ブック, 国際語学社 (1999).

[30] Kenneth Ward Church, Patrick Hanks : Word Association Norms, Mutual Information, Computational Linguistics Vol.16, No.1, pp.22-29(1990).

[31] Smadja, McKeown, and Hatzivassiloglou : Translating Collocations for Bilingual Lexicons, Computational Linguistics Vol.22, No.1, pp.1-38(1996).

[32] 北村美穂子, 松本裕治: 対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌, Vol. 38, No. 4, pp.727-736(1997).

[33] 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, 情報処理学会研究報告, 97-NL-128, pp.53-58(1997).

[34] Ido Dagan, Lillian Lee, Fernando Pereira : Similarity-Based Method for Word Sense Disambiguation, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, pp.56-63(1997).

[35] Donald Hindle : Noun Classification from Predicate-Argument Structures, 28th Annual Meeting of the Association for Computational Linguistics, pp.268-275(1990).

[36] 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, pp.16-24(1997).

[37] Yael Karov, Shimon Edelman : Similarity-based Word Sense Disambiguation, Computational Linguistics Vol.24, No.1, pp.41-59(1998).

[38] Jen Nan Chen, Jason S. Chang : Topical Clustering of MRD Senses Based on Information Retrieval Techniques, Computational Linguistics, Vol.24, No.1, pp.61-95(1998).

[39] Anderberg, M.R.: Cluster Analysis for Applications, Academic Press(1973). 邦訳： 西田英郎監訳：クラスター分析とその応用, 内田老鶴圃 (1995).

[40] 河口至商: 多変量解析入門 II, 森北出版株式会社 (1978).

[41] 田中豊, 脇本和昌: 多変量統計解析法, 現代数学社 (1983).

[42] 水本雅晴: ファジィ理論とその応用, サイエンス社 (1992).

[43] 田中英夫: ファジィモデリングとその応用, 朝倉書店 (1992).

[44] 宮本定明：ファジィグラフによる階層クラスタリングについて, 日本ファジィ学会誌, Vol.5, No.6, pp.1354-1371(1993).

[45] Walter Daelemans, Sabine Buchholz, Jorn Veenstra：Memory-Based Shallow Parsing(1999). http://ilk.kub.nl/cgi-bin/chunkdemo/demo.pl

[46] Satoshi Sekine：Apple Pie Parser(1996).
http://cs.nyu.edu/cs/projects/proteus/sekine

[47] Dragomir R. Radev：The ACL NLP/CL Universe(1999).
http://www.cs.columbia.edu/～ radev/u/db/ acl/html/AREA/SYNTAX

[48] Ivan A.sag：stanford HPSG projects(1999).
http://hpsg.stanford.edu/hpsg/links.html

[49] 橋田 浩一：The GDA Tag Set(1999).
http://www.etl.go.jp/etl/nl/GDA/tagset.html

付録

A. 参照データ

A.1 注釈変換データ：681 種類

語句	:	語義
Company_and_Licensee	:	PERSON
business=hours	:	PERIOD
Engineer	:	PERSON
infringement	:	breach
the_terms_and_conditions	:	condition
with_reference_to	:	about

A.2 動詞と前置詞の関連データ：809 種類

	動詞	＋前置詞：	関連度:	頻度:	(動詞,y):	(x, 前置詞):	距離
VP:	agrees,	to:	537.11:	443:	385:	12:	2
VP:	referred,	to:	254.00:	190:	385:	2:	1
VP:	grants,	to:	198.92:	152:	385:	3:	2

A.3 等位関係データ (等位関係にある語のリスト)：70 種類

information assistance
information data
information equipment
information evidence
information know-how
information know-how knowledge data technique material patent

A.4 述部のパターン：146 種類

動詞 ; 目的格・補語・Modifier のリスト

disclose; i * TO-PERSON TO-others, d *＜ NP ＞, m * TO-EXTENT;

divulge ; i * TO-PERSON TO-others, m *＜ PPNP-during ＞;

require ; i * NP-PERSON, d *＜ TOVB ＞

A.5 用語辞書：5,806 語

語	品詞	語幹
assign	: n.vt.	: assign
assignable	: a.	: assignable
assignation	: n.	: assignation
assigned	: pt.pp.	: assign
assignee	: n.	: assignee

A.6 類義語辞書：34 種類

advertisement, promotion

advice, guidance

airmail, cable, air letter, air mail, post, postage, telegram, telex

A.7 重要語一覧：34 条項 637 語

条項名識別子	重要語	重要度	出現頻度	条項頻度
sokusin	: advertise	: 4.8	: 18	: 14
sokusin	: sale	: 4.7	: 48	: 25
sokusin	: distributorship	: 3.8	: 3	: 1
sokusin	: inquire	: 3.1	: 2	: 1
sokusin	: query	: 3.1	: 2	: 1

B. 規則

本研究では，プログラミング言語としてPerlを用いている．そこで，本研究で用いる規則は，Perlの正規表現で表記している．

B.1 品詞タグ修正：183種類

```
s/\b(further)\/VB (pay) \/NN/$1\/RB $2\/VB/g;
s/\b(\S+\/NN\S*) (\S+)\/(VBG|JJ) (said)\/\S+\b/$1 $2\/JJ $3\/JJ/gi;
s/\b(the)\/(\S+) (following)\/S+ (\;|\:)\/\S+/$1\/$2 $3\/NN $4\/SYM/gi;
```

B.2 語句の連結：403種類

```
s/\b(calendar)\/\S+ (quarter)\/\S+/$1\=$2\/NN/gi;
s/\b(trade)\/\S+ (discount)\/\S+/$1\=$2\/NN/gi;
s/\b(trade)\/\S+ (secrets)\/\S+/$1\=$2\/NNS/gi;
```

B.3 構成語句の分割：94種類

```
s/(\S+\/NN\S*)\s+(\S+\/VB\S*)/$1\n$2/g;
s/(\S+\/NN\S*)\s+(\S+\/RB\S*)/$1\n$2/g;
s/(\S+\/NN\S*)\s+(\S+\/JJ)/$1\n$2/g;
```

C. 著者研究業績

本論文に関連する研究業績

論文誌 (査読付き)

(1) 相良かおる, 渡邊勝正 ：英文契約書式集に含まれる名詞間の類似度計算, 情報処理学会論文誌, 1999/12(掲載予定). 本論文第 5 章に関する内容

シンポジウム (査読付き)

(1) 相良かおる, 渡邊勝正 ：英文契約書の構造化のための統語構造の抽出と注釈付け, 「2000 年情報学シンポジウム」, 情報処理学会. 本論文第 6 章に関する内容

研究会等 (査読なし)

(1) 相良かおる, 渡邊勝正 ：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, pp.29-36, 1998/7. 本論文第 4 章に関する内容

(2) 相良かおる, 渡邊勝正 ：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 –, 情報処理学会研究報告, 98-NL-128, pp.159-165, 1998/11. 本論文第 5 章に関する内容

(3) 相良かおる, 渡邊勝正 ：英文契約書における内容の抽出 – 条文における権利・義務記載部分の内容を表す句構成セットの作成 –, 情報処理学会研究報告, 99-NL-130, pp.129-136, 1999/3.

その他の研究業績

研究会等 (査読なし)

(1) 相良かおる, 木村晋二, 渡邊勝正 ：手話単語の日本語による記述とその応用について, 情報処理学会研究会報告, HI62-9, pp.59-66, 1995/9.

(2) 相良かおる, 板倉寛如, 木村晋二, 渡邊勝正 ：手話動作の日本語による記述と3次元表示パラメータの抽出, 情報処理学会研究会報告, HI68-12, pp.87-94, 1996/9.

(3) 渡邊勝正, 朱 強, 相良かおる, 木村晋二 ：能動形プログラミングとスレッドによる実現, 日本ソフトウェア科学会第16回大会論文集, pp.193-196, 1999/9.

研究会等 (査読あり)

(1) 渡邊勝正, 木村晋二, 相良かおる, 高木一義 ：能動形プログラミング－ Active Programming －, 日本ソフトウェア科学会, PPL'99, pp.91-100, 1999/3.