

修士論文

2013 年度

専門領域の語彙ネットワークを用いた
文書の主題の類似度判定
-司法試験(短答式試験)問題の回答に向けて-

松下 裕

(学籍番号: 81221118)

指導教員 教授 山口 高平

2014 年 3 月

慶應義塾大学理工学研究科
開放環境科学専攻

論文要旨

近年人工知能分野の研究では、これまで細分化された形で培われてきた様々な分野の成果を統合し、再び実問題への適用可能性を真剣に探る動きが現れ始めている。近年大きな発展を遂げた機械学習を中心とする技術が非常に重要な役割を担う一方で、構造化された知識の利用にも再び注目が集められている。一方、知識工学と本来親和性の高い法律分野における応用例については、近年注目を浴びている諸分野と比較して盛んであるとは言いがたい。

本研究は、法律ドメインの極めて短い文書間の類似度判定法を、特に、与えられた司法試験の問題文から実際の法令に含まれる関連条文を取得するというタスクに即して提案し、これを通じてこの分野における構造化知識の有効性について論ずる。また提案では、構造化知識として領域知識を含んだ語彙ネットワークの構築と利用について述べ、これの利用可能性や限界について評価を与える。加えて、本研究において設定されたタスクを通し、自然言語によって自由に記述された事件記述から関連条文や判例を取得するという究極の目標に向けた課題や展望についても述べる。

Thesis Abstract

A Similarity Measure for the Central Theme of Documents based on a Domain-Specific Lexical Network: Toward Answering Multiple Choice Questions for the Bar Examination of Japan

In recent years, the research on Artificial intelligence is moving toward the integration of variously divided subdomains to again show a serious intension for the application to the real-world problems. While the urgent grown-up Machine learning centralized techniques play crucially important roles in many fields, using Structured or semi-structured knowledges is also gaining attention. Contrarily, artificial intelligence applications in legal domain, which has an intimate nature with knowledge engineering, are not so popular as other domains in the recent boom.

This paper proposes a similarity measure for extremely short texts in legal domain, especially aiming at the task of retrieving related laws given a question text of the bar examination of Japan. Through this task we discuss the effectiveness of utilizing structured knowledge in text processing in the legal domain. We construct a lexical network providing both linguistic and domain-specific relations between legal terminologies, and evaluate the utilities and limitations for the application to this retrieval task. We finish the work with discussion on the future work to extend the method for the more generalized legal information retrieval from freely typed descriptions of legal facts.

目次

論文要旨.....	2
Abstract.....	2
図の索引.....	4
表の索引.....	4
第1章 序論.....	6
1.1 研究の背景と目的.....	6
1.1.1 研究の背景.....	6
1.1.2 研究の目的.....	7
1.2 本論文の構成.....	7
第2章 関連研究.....	7
2.1 法律人工知能.....	7
2.1.1 法律エキスパートシステムと関連研究.....	7
2.1.2 PROLEG.....	8
2.2 情報検索とクエリー拡張.....	9
2.2.1 情報検索.....	9
2.2.2 クエリー拡張.....	11
第3章 司法試験問題文における関連条文検索.....	13
3.1 関連条文検索タスク.....	13
3.1.1 司法試験における短答式試験の概要.....	13
3.1.2 短答式試験問題の回答プロセスと関連条文検索.....	14
3.1.3 民法・短答式試験問題の関連条文検索タスク.....	15
3.2 民法語彙ネットワークを用いた問題文と条文の類似度判定.....	17
3.2.1 問題文、および民法条文における語彙の収集.....	17
3.2.2 語彙のベクトル空間モデルによる関連条文検索、およびベースライン手法.....	18
3.2.3 問題文・条文の特徴・形式分析と語彙選択.....	21
3.2.4 語彙ネットワークを用いた語彙拡張.....	26
3.3 本研究における提案のまとめ.....	43
第4章 評価.....	44
4.1 関連条文検索タスクに対する手法の評価.....	44
4.1.1 問題文・条文の特徴・形式分析と語彙選択.....	44
4.1.2 語彙ネットワークを用いた語彙拡張.....	46
4.1.3 関連条文検索タスクに際しての評価の総括.....	50
4.2 語彙ネットワークの評価.....	50
4.2.1 語彙ネットワークの構築プロセスの評価.....	50
4.2.2 関連条文検索タスクへの貢献.....	51
4.2.3 法律人工知能分野における利用可能性.....	52
4.2.4 語彙ネットワーク評価の総括.....	52
4.3 本研究の他分野への応用可能性.....	53
4.4 総括.....	53
第5章 結論.....	54
5.1 論文の総括.....	54
5.2 今後の課題および展望.....	54
謝辞.....	55

参考文献.....	56
付録 A 手法の分析・評価に用いた問題文とその関連条文.....	58
付録 B 語彙のフィルタリングによるコサイン類似度修正効果.....	66

図の索引

図 1: PROLEG ルールのイメージ.....	8
図 2: SynCha の出力加工のイメージ.....	8
図 3: PubMed におけるクエリー変換例.....	12
図 4: 問題文の単位のイメージ.....	15
図 5: 条文の単位のイメージ.....	16
図 6: 関連条文におけるストップワードリスト.....	18
図 7: 各文書内の語彙の重なり数.....	21
図 8: 法律要件と効果.....	23
図 9: 複数の要件-効果対による条文.....	24
図 10: ベースライン手法による検索順位と語彙の一致割合.....	25
図 11: 関連条文との語彙一致が全く見られない問い合わせ.....	26
図 12: 語彙のローカル・ネットワーク人力構成例.....	33
図 13: ノード補完パターン : hyperframes.....	34
図 14: ノード補完パターン : hypoframes.....	35
図 15: ノード補完パターン : attr_hypers.....	35
図 16: ノード補完パターン : prerequisites.....	36
図 17: ノード補完パターン : ascending_ways.....	36
図 18: ノード補完パターン : ascendedhubs.....	37
図 19: ノード補完パターン : attrs_of_attrs.....	37
図 20: 語彙の想起関係が必要と思われるケース.....	45
図 21: idf 重み補完で改善効果の高いケースの語彙構成.....	47
図 22: 一致語彙のないケースの語彙補完結果.....	49
図 23: メタ推論と背景知識が必要なケース.....	49
図 24: 全文書集合としての条文リスト.....	65
図 25: 語彙のフィルタリング : 問題設定.....	66
図 26: 語彙のフィルタリング : 操作後のイメージ.....	67

表の索引

表 1: 短答式試験の構成科目.....	13
表 2: 短答式試験の大問例(解答欄の指示等を削除済).....	14
表 3: 法律概念の語彙粒度と IPA 辞書による分割例.....	17
表 4: 形態素ペアのフィルタリング条件.....	18
表 5: TF-IDF 重みによる関連条文検索.....	19
表 6: tf-idf 重みを用いた関連条文の検索結果.....	20
表 7: 語彙頻度ファクターの変更と検索性能.....	22
表 8: 語彙のフィルタリング操作による検索性能の改善効果.....	26
表 9: hyper リンクの使用例.....	29
表 10: hyperx リンクの使用例.....	30
表 11: sbj リンクの使用例.....	30
表 12: obj リンクの仕様例.....	31
表 13: auth リンクの使用例.....	31

表 14: within リンクの使用例.....	31
表 15: attr_slot リンクの使用例.....	32
表 16: antecedent_to リンクの使用例.....	32
表 17: 語彙ネットワークの効果を最大化する語彙抽出.....	39
表 18: 語彙の条件付き補完.....	40
表 19: idf 重みの補完.....	41
表 20: 語彙の関連付けに用いた上位概念の扱いの変更.....	43
表 21: 構造化知識以前の手法改善.....	44
表 22: 語彙ネットワーク導入後の手法改善.....	46
表 23: 語彙ネットワークの導入による改善効果.....	48
表 24: 語彙ネットワークの規模.....	50
表 25: リンクの削除されたグラフと関連条文検索.....	51

第1章 序論

1.1 研究の背景と目的

1.1.1 研究の背景

近年人工知能分野の研究では、ビジネスへの応用に対する期待を軸に、実問題への適用可能性が再び真剣に検討され始めている。特に、これまでの人工知能研究で細分化され、培われてきた様々な分野の成果を統合しようとする動きが注目されている。

2011年にはIBMの研究部門が開発した質問応答システムであるWatson^[1]が、米国のクイズ番組Jeopardyで記録保持者に匹敵する正解率で回答した(ゲームとしては優勝)ことで話題となつた。これは、質問文の言語解析、大量のテキスト及びオントロジー等の構造/半構造知識を含む情報源の事前アノテーション、それによる解の候補生成及び解答の根拠情報の探索、過去のクイズ問題の学習も踏まえた確信度の計算、それらの並列化など、100を超える様々な要素技術の高度化と統合技術によって成し遂げられたものである。Watsonは現在、医療分野への応用^[1]、企業の顧客サポートへの応用^[2]、その他にも様々な事業化に向けた展開が進行^[3]しており、さらに実問題への挑戦に乗り出す構えである。また日本では、"ロボットは東大に入れるか?"と題して大学入試問題を解くための計算機プログラムの開発^[2]が2011年に国立情報学研究所で開始し、現在最も大規模な人工知能プロジェクト("東ロボ"プロジェクトと称されている)として目下注目されている。こちらも、深い意味理解を含む自然言語処理と質問応答の技術研究が進められていることに加え、オントロジーや科目固有のモデルリング等、知の構造化の研究も盛んに行われている。

いずれの試みでも、直近の20年で大きな発展を遂げた機械学習を中心とする統計的な技術は、自然言語処理技術の基礎部分も含めて非常に重要な役割を担っている。一方で、構造化された知識の利用も同時に実施/検討されていることは特筆に値する(Watsonに関しては[1]、具体的な貢献度に関しては明らかでないが、確かに利用されている。"東ロボ"プロジェクトに関しては[3]等)。構造知はいわゆる知識獲得のボトルネック、すなわち知識自体の記述・整備に大変な労力を要することが問題である一方、深い意味理解が必要であると思われる問題に、少なくとも人手で、曖昧性がなく有効な分析を与えることができ、あるいは機械学習等で解決方法が見出されていない類の知識情報に容易なアクセスを可能にする場合があるなど、人工知能の実問題への適用を考える際にはその導入の是非と効果に関して考慮することは一つの課題である。

ところで、先に述べた以外にも様々な分野において人工知能の実問題への適用の取組みがなされている一方、日本の法律分野における応用例については、実務家の支援等を含めた実問題に向けたものに関しては、近年注目を浴びている諸分野と比較して盛んであるとは言いがたい。この分野では、90年代に法律エキスパートシステムの開発研究^[4]として体系的な議論がなされているが、実問題への対処にあたって避けることのできない自然言語処理の高度な技術が発展途上であったこと、エキスパートシステム研究が下火となると同時にその適用領域としての法律人工知能研究も取組みが減ってしまったこと等、恵まれた状況ではないと言える。

-
- 1 [IBM News room - 2012-10-30 Cleveland Clinic and IBM Work to Advance Watson's Use in the Medical Training Field - United States(2014年1月28日参照) : <http://www-03.ibm.com/press/us/en/pressrelease/39243.wss>]
 - 2 [IBM News room - 2013-05-21 IBM Watson At Your Service: New Watson Breakthrough Transforms How Brands Engage Today's Connected Consumers - United States(2014年1月28日参照) : <http://www-03.ibm.com/press/us/en/pressrelease/41122.wss>]
 - 3 [IBM Watson(2014年1月29日参照) : <http://www.ibm.com/smarterplanet/us/en/ibmwatson/>]
-

1.1.2 研究の目的

本研究では、法律ドメインの極めて短い文書間の類似度判定法を、特に、与えられた司法試験の問題文から実際の法令に含まれる関連条文を取得するというタスクに即して提案し、これを通じてこの分野における構造化知識の有効性について論ずることを目的とする。また提案では、構造化知識として領域知識を含んだ語彙ネットワークの構築と利用について述べ、これの利用可能性や限界についても評価を与える。加えて、本研究において設定されたタスクは、人工知能による司法試験問題の回答へ向けた取組みの一つであると同時に、究極的には常用の自然言語によって自由に記述された事件記述から関連条文や判例を取得するという目標へのステップでもあり、これに向けた課題や展望についても述べる。

1.2 本論文の構成

第1章では、本研究に至る背景、およびその流れの中での位置付けとして、本研究の目的を明らかにする。第2章では、本研究の関連研究として法律分野の人工知能の取組み、および設定課題の技術的な基礎である情報検索について述べながら、本研究の技術的な位置付けをより明確化する。

第3章では、設定課題である関連条文検索タスクについて詳しく述べたのち、本研究において実施した手法を、簡単な実験と結果の表示を交えながら示す。構造化知識としての法律分野の語彙ネットワークの構築と利用に関しても、この章で詳しく述べる。第4章では、設定課題に対する本研究の手法の評価、および語彙ネットワークの構築と利用に対する評価を行う。

第5章では、本研究全体を通しての総括を与え、今後の課題と展望に関して言及する。

第2章 関連研究

2.1 法律人工知能

2.1.1 法律エキスパートシステムと関連研究

日本の、法律分野における人工知能研究で最も大きなプロジェクトは、90年代の「法律エキスパートシステム」の試み[4]であろう。このプロジェクトでは、1980年代に盛んに開発が行われたエキスパートシステムを法律分野で実現しようという試みを中心に、法的知識の主に論理学的観点からの解明、法的推論のための暗黙知を含む法律知識ベースの構築、推論エンジン等のソフトウェアの構築等様々なテーマが研究されたが、ここで少なくとも自然言語の領域に触れているのは、推論システムにおける言語生成[4](p.277)の部分だけで、この研究の総括でも自然言語処理の問題はひとつの研究の将来的な課題であったことが述べられている[4](p.381)。ここでは[5]等、法的推論システムの構成要素として概念階層辞書が用いられたものもあったが、役割はあくまで補助的なものであった。本研究では、体系的な法的知識や推論を扱うことはしないが、むしろ今まであまり議論されなかった、自然言語と知識を跨ぐ領域に焦点をあてたものである。

2.1.2 PROLEG

PROLEG[6]は、民法分野の訴訟における証明責任についての議論(要件事実論)を Prolog 言語上に再現したものであり、本研究における関連条文検索タスクは PROLEG のプロジェクトにおける司法試験問題回答チャレンジの一貫である。[6]を引用すると、「要件事実論とは、民事裁判において情報が不完全な状況下でも裁判官が判決を出せるように、民法の条文から得られる各要件に証明責任を付加させたものである。」とあり、法学の非常に論理学的側面に注目したプロジェクトであることがわかる。

PROLEG はコアの推論エンジンに加えて、以下の様なルールで構成されている。

```
条文(126,(取消権消滅時効(X,Y,Cause,T_recision)<=
    追認ができる日(T1),
    call(n年後(T1, T2, 5)),
    先立つ(T2,T_recision),
    意思表示の効力(時効援用,Y,X,Cause,T_recision),
    proleg条件(先立つ日(T2,T_recision))).
取消権消滅時効(X,Y,Cause,T_recision)<=
    追認ができる日(T1),
    call(n年後(T1, T2, 5)),
    先立つ(T2,T_recision),
    意思表示の効力(時効援用,Y,X,Cause,T_recision),
    proleg条件(先立つ日(T2,T_recision))).
```

図 1: PROLEG ルールのイメージ

司法試験問題回答チャレンジでは、試験問題文の SynCha⁴[7]による述語項構造解析結果を利用し、以下の様な中間表現を生成してルールに橋渡しを行う試みも行われている。しかしながら、これらと生文から得られた表記列によって構成される述語様・変数様のテキストは、ルール中の実際の述語や変数の表記と比べた場合、概念粒度にも抽象度にも差があり、さながらひとつの自然言語処理課題のような困難が伴い、苦戦を強いられている。

```
2 場合(前条).
3 期間(相当).
4 定めて(期間).
5 期間内(その).
6 するかどうかを(追認,期間内,定めて).
7 確答すべき(するかどうかを).
8 旨(確答すべき).
9 催告(旨).
10 する(催告).
11 こと(する).
12 できる(こと,場合,本人に対し,相手方).
```

図 2: SynCha の出力加工のイメージ

4 [SynCha: 日本語述語項構造解析器(2014年2月2日参照): <http://www.cl.cs.titech.ac.jp/~ryu-i/syncha/>]

2.2 情報検索とクエリー拡張

以下、大部分を[8]に基づいて述べる。

2.2.1 情報検索

伝統的な研究分野としての情報検索は、[8](p.1)によれば以下のように定義できる。

“情報検索(information retrieval, IR)は(通常、コンピュータに格納されている)大規模なコレクションから、必要な情報を含む非構造的な(通常、テキスト(text))資料(通常、文書(documents))を見つけることである。”

同書ではさらに、WWWを中心とした半構造化情報、より大規模で雑多なコレクションの登場、分類課題やその他様々な周辺領域の課題に波及し、情報検索の課題領域は上記の定義を超えて広がっていることを示唆している。

本研究における設定課題、すなわち(民法分野における)司法試験問題の文章から関連条文を取得する試みは、法律ドメインへの特化が重要な特徴である一方、どちらかと言えば伝統的な定義の範疇に近しいものである。

ベクトル空間モデル

現在非常に多くの情報検索技術の基礎として用いられているアイディアとして、ベクトル空間モデルがある。これは、文書内に出現した語彙とその文書における相対的重要性を、想定しうる全ての語彙を軸とした多次元空間の成分に見立て、文書ベクトルとして表現するものである。この相対的な重要度はしばしば重みと称され、様々な計算手法が提案されている。予期される語彙集合を $t_1, t_2, \dots, t_{N_t} \in T$ 、文書 d に対する語彙 t の重みを $\text{weight}(d, t)$ とすると、文書ベクトルは

$$\vec{V}(q) = (\text{weight}(d, t_1), \text{weight}(d, t_2), \dots, \text{weight}(d, t_{N_t})) \quad (1)$$

と表現される。

重みの計算方法として非常に標準的に用いられている方法に、**tf-idf 重み**がある。これは単純な文書中の語彙の出現回数 $tf_{t,d}$ に対し、検索対象の文書集合 d_1, d_2, \dots, d_{N_d} が与えられ、文書集合中で語彙 t が出現する文書数 df_t が判明したときに、逆文書頻度：

$$\text{idf}(t) = \log\left(\frac{N_d}{df_t}\right) \quad (2)$$

を乗算するものである。すなわち、

$$\text{weight}(d, t) = tf_{t,d} \times \text{idf}(t) \quad (3)$$

である。逆文書頻度は稀な語彙の重みを大きく、頻出語彙の重みを低くすることで、文書識別に有用な情報を与えようとするものである。

語彙頻度のファクターに関しては、語彙の出現回数の小さな変化(特に小さい領域：1回→2回等)が重みに大きな影響を与えるのは実態に合わないとして、修正を行う場合がある。例えば、

$$wf_{t,d} = \begin{cases} 1 + \log(tf_{t,d}) & (\text{if } tf_{t,d} > 0) \\ 0 & (\text{otherwise}) \end{cases} \quad (4)$$

等である。

コサイン類似度

こうして得られた文書表現の類似度を量化する手法として、コサイン類似度がある。すなわちこれは、文書ベクトル $\vec{V}(d)$ をそのノルム $\|V(d)\| = \sqrt{\sum_{i=1}^{N_t} \vec{V}_i^2(d)}$ で正規化したもの同士の内積をとり、

$$\text{similarity}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\|\vec{V}(d_1)\| \|\vec{V}(d_2)\|} \quad (5)$$

とするものである。因みにこれは、2文書のベクトル間の角度 θ に対して、 $\cos\theta$ をとった値と等しくなる。これによってベクトル空間における文書ベクトル同士の方向の近さを量化することができる。

ある文書をクエリーとしてみなし、検索対象の文書集合に属する全ての文書との間でこの類似度を計算して値が大きな順で並べると、クエリー文書との類似度ランキングを作成することができる。ここから、上位の数件、あるいは1件を得ることで、求める文書の検索手法の一つとして用いることができる。

再現率と適合率

情報検索の評価指標として最もよく利用されるものに、再現率(recall)と適合率(precision)がある。これは、情報検索システムの入力を一つのクエリー、出力を文書集合とした場合、それぞれ

$$\text{再現率} = \frac{\text{出力された正解文書数}}{\text{出力文書数}} \quad (6)$$

$$\text{適合率} = \frac{\text{出力された正解文書数}}{\text{クエリーに対応する全正解文書数}} \quad (7)$$

という式で表される。これらは一般的にトレードオフの関係にあり、取得文書数を増加させる、すなわちランクイングの下位まで出力に加えることで再現率は上昇するが、順位づけにおける正確さが向上しない限り、適合率は上がらない。一方で、厳格な順位付けを行うシステムでは、幅を持った正解文書を取得することが難しくなるのが普通である。

これら2つの指標を同等に評価し、一つの指標として量化したものにF値がある。これは再現率と適合率の調和平均で表され、

$$F = \frac{1}{\frac{1}{\text{再現率}} + \frac{1}{\text{適合率}}} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \quad (8)$$

と計算される。

2.2.2 クエリー拡張

以下は[9]の調査研究をもとに述べる。

検索対象の文書集合において、特に本研究のように各文書長が極めて短い場合、単純な語彙の一致のみでは検索能力が不十分な場合がある。すなわち、クエリー文書に現れる語彙や語彙の組み合わせと関係がある文書において、同義で表記の異なる語彙、同表記で(本質的にあるいはコンテキストによって)意味の異なる語彙の存在や言い換えによる不一致によって、類似度が上がらないことが問題となる。本研究の対象は言わば、法律分野の技術文書であり、用語表記の曖昧性はオープンドメインの文書よりも低いと言える。しかしながら、ひとつひとつが固有な役割を担う条文には(第3章で詳しく述べるが)、内容の冗長性が低く、それに伴って、文書内の語彙の出現頻度や逆文書頻度の情報が単純に信頼できないという特徴もあり、この問題が当てはまるのである。

こういった問題に対処するための方法として、主に入力されたクエリーの語彙等の拡張、検索結果の段階的な評価により最終的な結果集合を改善する適合フィードバックの手法、検索結果のクラスタリングなどが研究されている。このうち本研究の設定タスクでは、クエリー(文書)語彙の拡張、特に情報検索システムを利用するユーザーとのインタラクションなしにこれを行う、automatic query expansion (AQE)と呼ばれるものが利用可能であろう。最新の調査[9]によれば、AQEは40年程度の研究を経た今、実用段階に到達し始めているという。

クエリー拡張、特に語彙の補いを用いる手法の大きな特徴は、入力された語彙の同義語等が補完されることによる再現率の向上である。一般的にこれは、補完自体の誤り、あるいは同じ語彙表記でも、コンテキストを考慮しないことによる非関連文書の類似度上昇等によって、適合率とのトレードオフとなる。ここで、専門家・研究者向けの法律、医療、特許、その他一般に学術文献の調査(特に先行研究の調査等)を対象にした情報検査システム／タスクの場合、適合率よりも再現率が圧倒的に重視される傾向にあるのは特筆すべきである。

またこういった領域の文書では、オープンドメインの文書集合と比較して専門用語等の表記のゆれが少ない。このため、本研究でも実施するオントロジーあるいは語彙知識ベース(シソーラス)を利用したクエリー拡張が効果的である場合が多いといえる。よく知られた例として生物学・医学分野における文献検索システムであるPubMed⁵では、"blood pressure"というクエリーで検索を行うと、シソーラスに基づいて自動的に以下のようなクエリーに変換され、同義のより専門的な用語を用いた文献等も取得することができる。

```
"blood pressure"[MeSH Terms] OR ("blood"[All Fields] AND "pressure"[All Fields]) OR "blood pressure"[All Fields] OR "blood pressure determination"[MeSH Terms] OR ("blood"[All Fields] AND "pressure"[All Fields] AND "determination"[All Fields]) OR "blood pressure determination"[All Fields] OR ("blood"[All Fields] AND "pressure"[All Fields]) OR "blood pressure"[All Fields] OR "arterial pressure"[MeSH Terms] OR ("arterial"[All Fields] AND "pressure"[All Fields]) OR "arterial pressure"[All Fields] OR ("blood"[All Fields] AND "pressure"[All Fields])
```

図 3: PubMed におけるクエリー変換例

5 [Home - PubMed - NCBI(2014年1月30日参照) : view-source:<http://www.ncbi.nlm.nih.gov/pubmed>]

一方最近の研究では、オープンドメインの語義曖昧性解消も含めた研究として、WordNet⁶[10]の利用に焦点を当てたものが多い。加えて本研究で実施するリンク解析等も行っているものとして、[11]等が挙げられる。

本研究の設定課題である法令条文の検索については、ドメインは法律、さらに日本民法の分野に限られるため用語表記のゆれが狭い一方、適合率は再現率と同じ程度に重要であると言える。また問題文(=法令適用事例の記述、および知識の確認のための記述)から法令を検索するという、決して研究が盛んであるとは言えない特徴、あるいは致命的に短い文書長によって、語彙の単純な言い換えでは不十分な場合が多く、工夫が必要である。提案では、語彙の組み合わせや小さな集合から、語彙ネットワークのヒューリスティックなリンク解析を通じてクエリー(文書)の拡張を行う。

6 [About WordNet - WordNet - About WordNet(2014年1月30日参照) : <http://wordnet.princeton.edu/>]

第3章 司法試験問題文における関連条文検索

本研究では、対象である法律分野の自然言語理解に向けた取り組みの一環として、司法試験問題文における関連条文検索タスクを設定している。本章では、このタスクの内実を明らかにするとともに、本研究で実施した手法の詳細と実施について述べる。

この章では、法令の具体的な条文に言及して説明を行うことがある。この際法律分野での条文番号の表記は漢数字を用いるのが規則となっているが、視認性に配慮して、アラビア数字を適宜用いることとする。

3.1 関連条文検索タスク

3.1.1 司法試験における短答式試験の概要

日本の(新)司法試験制度について、本研究に関わる範囲で簡単に述べる。法務省の掲載する情報⁷によれば、司法試験は「裁判官、検察官又は弁護士となろうとする者に必要な学識及びその応用能力を有するかどうかを判定する試験」である。平成18年度より開始した現行の制度では、4日間にわたり、短答式試験と論文式試験の二種が実施され、それぞれ評価される。

このうち短答式試験は、以下の三科目で構成され、そのそれぞれで正答率40%以上を得点しなければ論文式試験が採点されることではなく、さらに実際、合格者の最低得点は60-67%を推移し[12](pp.11-12)、司法試験の第一の閑門と呼ぶべきものとなっている。

表1: 短答式試験の構成科目

科目	内容
公法系科目	憲法及び行政法に関する分野の科目
民事系科目	民法、商法及び民事訴訟法に関する分野の科目
刑事系科目	刑法及び刑事訴訟法に関する分野の科目

短答式試験はマークシート方式で回答を行うもので、おおよそ似た形式の大問が続く形式となっている。すなわち、幾つかの選択肢を示し、その中の正しいもの/誤っているもの、大問の条件に合うものを選択するもので、おおよそが各選択肢の文章の正誤問題へと還元できる場合が多い。つぎに典型的な大問の例を示す⁸。なお、表中[平成19年度 第37問]のような、単純な法的知識以上のものが問われる大問は非常に稀である。本研究でこれより取り扱う民法に関する問題に限れば、年度およそ35問程度のうち1-2間にとどまる。ところで、問題番号についても略号を用い、これ以降[平成[yy]年度 第[qq]問 選択肢[cc]]を”q[yy]/[qq]/[cc]”と記すこととする。例えば[平成19年度第37問選択肢A]では、”q19/37/A”となる。

7 [法務省：司法試験(2014年1月27日参照)：
[http://www.moj.go.jp/jinji/shihoushiken/shikenqa.html](http://www.moj.go.jp/jinji/shihoushiken/shiken_shinshihou_shikenqa.html)]

8 過去に実施された試験問題は、全て[法務省：司法試験の実施について(2014年1月27日参照)：
http://www.moj.go.jp/jinji/shihoushiken/jinji08_00025.html]以下で入手できる。

表2: 短答式試験の大問例(解答欄の指示等を削除済)

問題番号	大問の問題文
平成18年度第1問	売主の担保責任に関する次の1から5までの記述のうち、誤っているものはどれか。
平成18年度第8問	親子関係をめぐる訴訟に関する次のアからオまでの記述のうち、正しいものを組み合わせたものは、後記1から5までのうちどれか。
平成18年度第13問	Aは、その所有する甲土地をBに売却したが、その後に重ねて甲土地をCに売却し、さらにCは直ちにDに転売した。甲土地の登記名義は、A・C・Dの合意に基づき、Aから直接にDに移転された。この事例に関する次の1から4までの記述のうち、誤っているものはどれか。
平成19年度第37問	次のアからオまでの各記述のうち、株式会社は定款所定の目的の範囲内でのみ権利能力を有するという考え方に対する批判として、ふさわしくないものを組み合わせたものは、後記1から5までのうちどれか。

3.1.2 短答式試験問題の回答プロセスと関連条文検索

こういった問題で問われているのは、条文知識や判例知識と呼ばれるものである。まずもって条文知識とは、すなわち科目の対象となっている法律の条文内で規定されていることである。

一方、これら条文は、ある程度抽象的な語彙・表現を用いて記述されているのが普通である[13](pp.9-11)。法に基づく紛争の解決に際しては、様々な個別の事例が公平な観点から裁判されることが重要視されている。条文、つまり制定法ではこれを紛争のパターンや事実関係を抽象化し、個々の事例にはこの立法趣旨を解釈して適用することで、判断基準の整合性を確保しようとしている。ここで重要なのが判例である。裁判所、特に最高裁判所による制定法の解釈、事例への適用は、単に条文の解釈例というのではない。以後の裁判における判断はその判例自体との整合性を保つよう拘束され、判例は条文と同程度に尊重されることとなる[13](p.90)。司法試験の受験者は、条文の知識に加え、重要な判例に関する知識をもって回答することも求められる。

短答式試験の各問いは、おおよそこれら条文や判例の規定をどれか一つ用いることで、回答できるものとなっている。その規定が含まれる条文を、これ以後、その問い合わせの関連条文と呼ぶこととしよう。

さて、短答式試験の受験者は、これらの知識が必要とされる問題にどのようなプロセスをもって回答するだろうか。まず直ちに思いつくことといえば、あらゆる条文や判例をそれぞれ記憶し、必要に応じて関連条文を想起できるようにしておくことである。短答式試験には資料や文献を持ち込むことができない。従って条文や判例を、その共通するトピックや構造の類似性を背景に有機的に関連させて覚えておくことが必要だと考えられるのである。

しかしながら実際の司法試験受験者へのインタビューによれば、まず法律知識の記憶は、およそ条文や判例を単位としていない。さらに言えば、条文や判例個々の内容も、全て仔細に記憶する必要性があるわけではないという。前述のように、条文や判例の本質は、抽象化された条文のさらに背景にある立法趣旨とそれらの整合性にあると言つてよい。これらはまさに共通するトピックや構造の類似性そのものであり、これらを体系的知識として直接頭に容れることで、問題に回答することが出来るのだという。

第2章に詳細を述べたように、本研究の背景となっているPROLEGプロジェクトでは、民法分野の司法試験問題の回答に向けた取り組みが行われている。ここでは、以上に述べたような試験受験者のプロセスを再現するための法律人工知能を構成するべきだろうか？

詳しくは第2章を参照されたいが、汎用の言語処理技術を用いて問い合わせの自然言語文を直接PROLEGの記述言語へと割り当てを行うのには、相当な困難が伴うことがわかっている。ここで用いることができるものは、PROLEG(のルール)を実質的に構成する論理構造と、ルールの記述者が述語や変数に付与したテキストである。PROLEGの論理構造には問題文に出現する動詞等の格情報と少なからず関係があるが、現在、深層格の同定精度は十分とは言えない上、述語引数の順序等の論理構造と深層格は本質的に無関係である。また、その困難を解消できたとしても、ルールに含まれるテキストと問題文の自然言語との間には、表現と語彙表記の曖昧性によるギャップが残る。

本研究では、PROLEG言語による民法分野の短答式試験問題の回答に向けた取り組みの一環として、与えられた短答式試験問題の自然言語文から、これを回答するために必要な規定を含む関連条文を検索／探索するタスクを設定した。

これは実際の試験の受験者のプロセスには見られないものと言えるが、以下の効果を期待できる。一つは、試験問題文よりもPROLEGルールの含むテキストとの親和性が高いと考えられる条文のテキストを用いることで、前述のギャップを縮小できるのではないかということである。また、あくまでどちらも自然言語文である試験問題文と条文を分析・比較する中で、ルールと問題文を橋渡しするのに有用な知見や情報が得られるのではないかということである。

以上を受け本研究では、PROLEGの対象とする民法分野において、与えられた短答式試験問題からその関連条文を得るために手法について提案を行った。

3.1.3 民法・短答式試験問題の関連条文検索タスク

本研究におけるタスクの形式的な定義を行う。

本タスクは民法に関わる短答式試験の問題文が与えられた上で、番号の付与(法文に元々記載されているもの)された民法条文のテキスト群⁹から最も関連条文として相応しいものを選び、その番号を出力するものである。関連条文については、[12]およびこれと同系の参考書とともに専門家の協力のもと選出されたものである。

⁹ 全文を[民法(2014年1月28日参照)：<http://law.e-gov.go.jp/htmldata/M29/M29HO089.html>]で入手できる。

まず、試験問題文の単位を示す。本研究では関連条文を検索する単位として、各大問に付属しているそれぞれの選択肢の文章をとる。その上で、大問の文章と合わせることで一つの問題文とした。

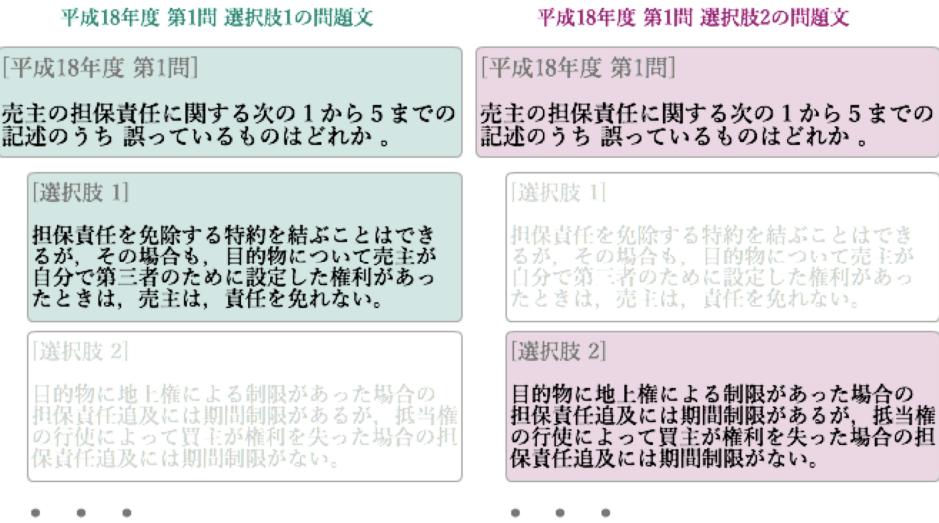


図4: 問題文の単位のイメージ

一方、条文の単位は基本的に、条文番号毎とする。民法条文にはタイトルを含むものが多くあるが、そのテキストに関しても条文の一部として含める一方、章や節のタイトルは含めないこととする。号(要件等の列挙に用いる)や但し書き(条文の規定に対する例外事由を述べる文)に関しては、特に分割したりせずそのまま含めることとする。現行の民法には第一条から第千四十四条まで、数にして1048の条文が含まれている。番号とサイズが一致しないのは、法令では条文の削除や追加があっても条文番号の変更を行わず、番号のスキップや”三百九十八の二条”といった枝番の追加を行うためである。

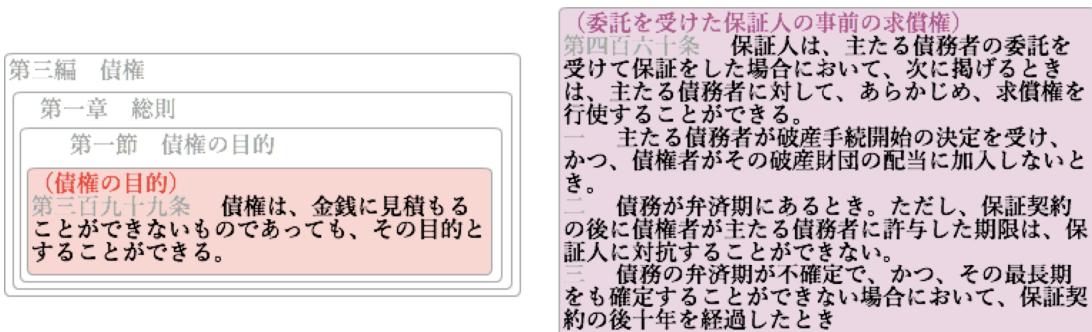


図5: 条文の単位のイメージ

ところで、前述のように短答式試験の問題文の正誤判定には、条文知識に加えて判例知識が必要な場合がある。従って関連条文には判例が当てはまる場合があるが、本研究における実験時にはデータの収集、すなわち試験で問われる可能性のある重要な判例のテキストの集取¹⁰が完了していなかったため、判例が回答に必要であるような問題に関しては対象から外すこととした。

10 法解釈として重要な意味のある判例に関しては、全て裁判所のウェブサイト[裁判所 | 裁判例情報 (2014年1月28日参照) : http://www.courts.go.jp/search/jhsp0010?action_id=first&hanreiSrchKbn=01]で全文や要旨入手することができ、本研究と同時に収集が進められている。

タスクの一次的な評価方法に関して述べておく。このタスクの実施において典型的な方法論として、与えられた問題文と各条文の関連度を計算し、そのスコアが高いものから並べて順位を付与するというものがある。本研究でも同様の方法で評価を行っているが、このとき、関連度の値が等しい状態が発生する場合があろう。本研究ではこの場合、それらの順位がランダムに変化する可能性があるという意味合いで期待値をとり、そのケースの順位として扱っている。すなわち実際には、当該ケースよりも上位(すなわち若い順位)の条文群の順位集合を $R_{superior}$ 、当該ケースと等しいスコアのケース数を $N_{equiv.}$ と書いて、

$$\text{順位の期待値} = \max(R_{superior}) + \frac{N_{equiv.}}{2} \quad (9)$$

によって計算される。こうして得られた順位について、後に研究目的に沿った方法で評価していく。

ある問題文における個々の関連条文の順位は上記によって求められるが、関連条文はひとつでない場合が有り得る。このとき、関連条文間に関連度の差異はないとした上で、問題文 q_i の a_i 本の関連条文について、その順位を若いものから並べたものが r_1, r_2, \dots, r_{a_i} であった場合、以下の式で一つの数値として順位を与える。これにより、関連条文数の異なる問題について同様に評価を行うことができる。

$$\text{順位} = \frac{1}{a_i} \times \sum_{j=1}^{a_i} \frac{r_j}{j} \quad (10)$$

ところでこの関連条文数はもちろん、問題文の内容に依り個々のケースに固有のものであるが、今のところ選出されたものについて、3本を越えるものは現れていない。従って検索課題の評価にあたり、特定の基準を決め、それより上位の文書集合を調査する必要がある場合、3を基準に考えることとする。より具体的には、再現率および適合率、あるいはF値の計算に際してのことである。

3.2 民法語彙ネットワークを用いた問題文と条文の類似度判定

以降、3.1.3で定義したタスクに対する、本研究における取組みと提案手法に関して述べていく。

3.2.1 問題文、および民法条文における語彙の収集

本タスクは3.1.3で定義されたように、民法分野の司法試験問題文から民法の条文を取得するものである。ここで問題文をクエリー文書、民法の全ての条文を検索対象の文書として考えれば、2.2.1で紹介したベクトル空間モデルを直接適用できることが分かる。

ここで問題となるのは、語彙をどのような単位で決定、あるいは抽出するかという点である。単語の区切りが明示されない日本語のテキストにおいて、近年はオープンソースの形態素解析エンジン MeCab¹¹[14]に特定の辞書を組み合わせて、非常に安定した結果を得る環境が整っている。一方、法律分野で概念として取得すべき語彙は非常に大きな粒度を持っていることが多い。本研究では、比較的分割の粒度が大きい IPA 辞書(MeCab と併せて公開されている)を用いて解析を行ったが、以下の表の通り(一部を抜粋したものである)必要な法律概念を直接取得するのは難しい。

表 3: 法律概念の語彙粒度と IPA 辞書による分割例

法律概念として重要な語彙	わかつ書きの結果
所有権	所有+権
被補助人	被+補助+人
極度額	極度+額
法定果実	法定+果実
事務管理	事務+管理
代襲相続	代+襲+相続
自働債権	自+働債+権

この問題を放置した場合、必要な概念粒度で語彙出現を捉えられないばかりか、前／後方にて部分一致する語彙同士が条文の識別に不利な影響を及ぼす場合が数多く存在する。例えば、表中にも見られる”事務管理”という概念は、義務なく他人のために事務を管理(立替え払いや隣家の塀の修理、火災の消火等、生活に利益をもたらすことを広く指す)することをいう[15](p350)。これは民法の第 697-702 条にのみ集中して記述されている。一方、民法において”管理”という語彙は、不在者の財産の管理、未成年に関する財産の管理、相続財産の管理、その他一般的な物の管理等に散発的に現れ、もしわかつ書きの結果を直接用いるのであれば、これらの条文とのスコアの上下が条文識別に影響を与える。また法律文書の文脈で事務管理について述べる場合、”事務”と”管理”が分離した表現しか出ないという状況はあり得ず、少なくとも第 697-702 条を参照すべきであるから、これは実際条文の識別力を落としていると言えるのである。

11 [MeCab: Yet Another Part-of-Speech and Morphological Analyzer(2014/01/28 参照)：
<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>]

本研究ではこの問題に対処すべく、上記のような民法ドメインに現れる粒度の大きな語彙を網羅するため以下の操作を行った。まず、形態素解析結果からバイグラム、すなわち形態素列 $t_1, t_2, t_3, t_4, \dots$ が与えられたとき $(t_1, t_2), (t_2, t_3), (t_3, t_4), \dots$ という連続するペアの列を探る。ここで、ペアの列を以下の条件で絞り込む。

表 4: 形態素ペアのフィルタリング条件

左側の形態素の条件	右側の形態素の条件
品詞が名詞である ¹²	品詞が名詞であり、表記がブラックリスト ¹³ にない
品詞が接頭辞である	品詞が名詞であり、表記がブラックリストにない

続いて、絞り込んだペアによって連語を作成するすなわち、絞り込んだペアとして $(t_a, t_b), (t_b, t_c), (t_d, t_e)$ が与えられたとき、 $t_a, t_a, t_b, t_c, t_d, t_e, t_b, \dots$ を形態素列として得た場合は、 $\langle t_a \rangle, \langle t_a t_b t_c \rangle, \langle t_d t_e \rangle, \langle t_b \rangle, \dots$ を得る。

こうして得た複数の形態素の連続を含む語彙を、問題文、条文双方の文書ベクトル作成に用いる。真に厳密な処理を行うには、これを MeCab のユーザ辞書に登録するか、これらの語彙を含めて何らかの法律ドメインのコーパスを用いて形態素接続コスト等の再学習を行う必要があるが、本研究においてはより簡便な方法で十分な語彙獲得を行うことができた。すなわち文書を先頭から走査し、前方一致で拾うことのできる語彙を逐次的に取得し、その中で表記が最長のものを語彙として抽出するというものである。また、人手で整備したトップワード(除外語彙)をここから除いた。これを以下に示す。

甲，乙，丙，人，等，次，号，項，
年，月，日，週，法，節，条，章，款，
旨，者，前項，一種，前条，後段，当該，次項，前章，一節，
前半，適合，基準，行為，事項，関係，さ，判定，
同等，一致，共通，推知，状態，意見，程度，情，確認，
更，帰，存，適，みそ

図 6: 関連条文におけるトップワードリスト

3.2.2 語彙のベクトル空間モデルによる関連条文検索、およびベースライン手法

以上の方法を用い必要な概念粒度で語彙を取得することによって、2.2.1 で示したベクトル空間モデル、およびコサイン類似度を用いた情報検索の形式で関連条文検索を行うことができた。

12 品詞は MeCab の出力をもとに、IPA 辞書のソースに同梱されている "pos-id.def" を参照して判定している。

13 ここでは {"等", "内", "間"} のみをリストに入れている。ちなみに接尾辞は、IPA 辞書において名詞のサブカテゴリとして分類されている。

以下は、平成18-20年度の(新)司法試験について、3.1.3で定義したタスクを実施したものである。語彙取得は上記通り、重みとしてtf-idf重みを用いた。また、表中の”正答率指標”というのは、「取得条文の上位3件に関連条文が一つでも入っているか?」という基準で判断を行ったもので、あくまで直感的におおよその成績を判断するために便宜上算出したものである。

表5: TF-IDF重みによる関連条文検索

年度	対象問題数	正答率指標(%)
18	65	64.6
19	44	93.2
20	55	80.0
18-20(集計)	164	77.4

上記を見るに、この正答率指標は意外に高いものであると感じられるだろう。すなわち、少なくとも民法分野において、試験問題文による関連条文の検索は、語彙の完全一致によってかなりの部分が達成できるということが判断できる。民法の条文では、あらゆる概念に用語が割り振られているが、司法試験問題においても、これらの用語がそれなりに一貫して用いられていることが一見すれば判り、その結果としてのことだと言って良いだろう。

分析のための問い合わせの限定

さて、この状態から検索手法の精度を向上するには、具体的に問題文からどの語彙が抽出されるか、それを関連条文、あるいは上記の手法で誤って上位に検索された条文からの抽出語彙と比較した場合何が言えるか、そしてそれぞれの語彙にはどのような重みが付与されているか、ということをケース毎に吟味する必要があった。また、これは後に詳しく述べるが、本研究の最終的な提案手法である語彙ネットワークを用いたクエリー拡張においては、語彙ネットワークの人の手による構築規模の関係で、評価に含めることのできる問題文にかなり制限が生じる(手法の限界ではなく、単純に登録語彙のカバレッジが不足していることによる)という事情があった。

上記を踏まえ、分析・評価に用いる問題文を次頁の11件に絞ることとした。文章等の詳細な内容に関しては、付録Aに掲載する。ここでは、tf-idf重みを用い、3.1.3に正確に従って関連条文の順位を求めたものである。

表 6: *tf-idf* 重みを用いた関連条文の検索結果

問題文	関連条文の順位	問題のおおよそのトピック
q18/15/1	1	根抵当権の順位と競売に関する優先権
q18/15/2	7	根抵当権の被担保債権の範囲
q18/15/3	2	根抵当権の制度趣旨
q18/15/4	2	根抵当権の被担保債権の譲渡の効果
q18/15/5	1	根抵当権の譲渡／一部譲渡の要件
q19/13/エ	1	根抵当権の被担保債権の範囲
q19/13/オ	8	根抵当権の被担保債権の範囲
q19/16/1	1	抵当建物使用者の引渡しの猶予
q19/16/2	1	抵当権者の同意と賃貸借の対抗力
q19/16/4	8	抵当権消滅請求の要件
q19/16/5	1	被担保債権債務不履行後の 抵当不動産の果実の帰属
q19/7/3	44	売買の目的物の引渡しと果実の帰属
平均	6.42	—
F 値平均 (3 位基準)	0.358	—

表中の右端の列の内容は、おおよそ問題内で何か問われているかということを簡単にまとめたものだが、かなり“抵当権”，および“根抵当権”に偏っていることが分かるだろう。これは前述の通り、語彙ネットワークを用いた手法への適用可能性の制限による。実際のところは、平成18-20年度の試験においてまとまった数の問題文が採れるトピックを事前に選択し、語彙ネットワークを構築したという事情がある。

順位を見たところ、一件かなり下位に出力されているケースがあるため平均値は低めになっている。一方3位以内が出力されたものの数を数えると、約67%である。先の平成18-20年度の全問の正答率指標は平均が77.4%であったが、これと比較すれば、関連条文の取得の難しさに関しておおよそこれらと同程度の偏りでサンプルできたと言ってよいだろう。

本研究では、これらの問題文の集合をもって、また以上のtf-idf重みと語彙取得法による検索結果をベースラインと定めて、手法の比較提案を行っていくこととする。この研究では、ベースラインからの改善の積み重ねという形で種々の手法を導入している。従ってこれ以降の提案に関しても段階的に行い、また各々一つ前の結果の分析を含めながら、手法の仮説立てと検証という形で詳細を述べていくこととする。

ところで、語彙ネットワークによる実行可能なケースの制限は、当然のことながら条文からの語彙取得にも制限を与える。従って語彙の可能な限りの対応をもって、最低限の条文文書集合から検索課題を実施できるよう整備を行った。すなわち、本研究のこれ以降の検索課題の実施にあたっては、この11問の関連条文に、ベースライン手法、およびいくつかの予備実験において、誤って上位10件に順位付けされた条文を加えた63件の条文を全文書集合として設定した。これを付録Aの後半に追記する。これは見方によっては検索の成功率を上昇させているようにも思えるが、ベースライン手法で条文の絞り込みを行った後に適用する手法として考えれば、特に人手を介さずとも検索課題を実施できることから問題視しないこととする。

3.2.3 問題文・条文の特徴・形式分析と語彙選択

法律文書、特に制定法の条文には、このドメイン独自の特徴が幾つかみられる。以下では、2つの特徴をもとに実施した手法について述べる。

条文：文書としての非冗長性

民法典は、あるトピックに共通する一般的・抽象的な話題をまとめ、より具体的な個別の規定に先立って配置するという方法論、そしてこの原則が、大きな区分の編・章から節・各々の条文の並べ方に至るまで、フラクタルに貫かれているということが知られている¹⁴。民法に限らず一般的に制定法は、規定の重複を可能な限り避けるように構成されている[13](p.49)が、この非冗長性への執着は、各々の条文を文書、さらに言えば語彙列として見た場合にもある程度観察されることが分かる。

法律文書にはその性質上、意味内容を過不足なく表現することが求められるだろう。特に条文に関しては、議論等を含まずに何らかのルールを簡潔に明記するものである。これによって、いかに重要な(すなわち、条文の主な内容を代表していると考えられるような)概念の語彙でも、テキスト上に占める割合(すなわち語彙の出現回数 $tf_{t,d}$)が上がり難い場合があると考えられる。一方で条文によっては、極めて特殊な状況に関する規定、あるいは然程重要でない内容についても、これに複雑な記述が必要であれば、それらの語彙の登場回数は増えて然るべきである。まとめれば、文書中の語彙の出現回数は語彙の重要性でなく、その語彙によって記述される内容の複雑性に依る。

¹⁴ パンデクテン方式という。[16]等に他の方式との比較を含めた議論がある。

これにより、そのような複雑な記述の必要な規定を含まない限り、それぞれの文書中の語彙の重なりが極めて小さくなることが予想される。以下はこれを平成18-20年度の問題文と条文に対して計測したものである¹⁵。ここでは各条文および問題文について、観測されるそれぞれの語彙が同一文書内で何度重複して現れるかを数え、全文書で集計を行った。グラフでは横軸にこの語彙の重複数、縦軸に重複数毎の観測数をとった。文書集合および条文内の語彙出現には、非常に重複が少ないという特徴があることが分かる。

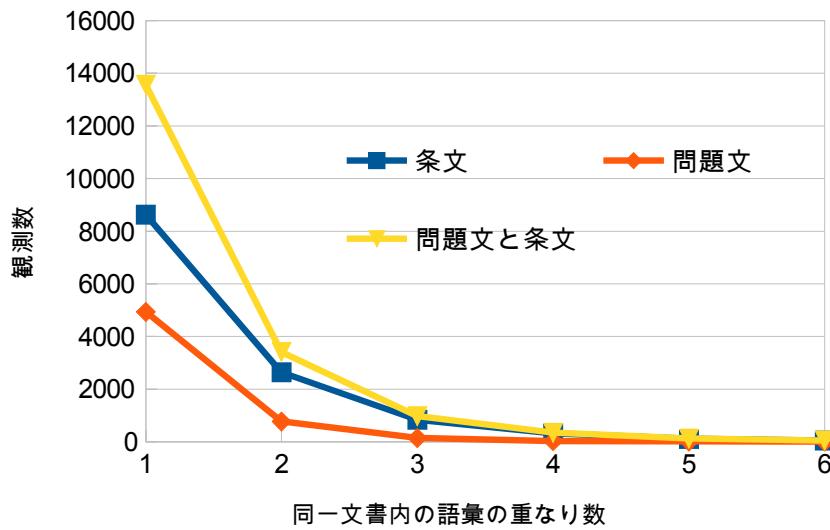


図7: 各文書内の語彙の重なり数

以上より、語彙出現を示すファクターを $tf_{t,d}$ から以下の値へ変更することとした。これを $\delta_{t,d}$ と表示することとする。

$$\delta_{t,d} = \begin{cases} 1 & (\text{if } tf_{t,d} > 0) \\ 0 & (\text{otherwise}) \end{cases} \quad (11)$$

¹⁵ これは視覚的に語彙の頻度と語彙の出現回数順位の相関関係[8](p.80:ジップの法則)に類似しているが、全くの別物であるため注意されたい。

これを適用すると、以下のように確かに改善効果をみることができる。表中の $wf_{t,d}$ については、第2章の式(12)によるもので、こちらも $tf_{t,d}$ からの改善効果を見ることができる。一方こちらも $tf_{t,d}$ における文書中の語彙出現回数の効果を単に和らげる効果もあるから、ここでは $\delta_{t,d}$ の改善効果と同じ理由であって $\delta_{t,d}$ を直接採用すべきだと結論づけて良いだろう。ちなみに逆文書頻度のファクター、および他の条件に関しては、ベースラインに特に変更を与えていない。

表 7: 語彙頻度ファクターの変更と検索性能

問題文	$tf_{t,d}$	$wf_{t,d}$	$\delta_{t,d}$
q18/15/1	1	1	1
q18/15/2	7	8	6
q18/15/3	2	3	3
q18/15/4	2	1	1
q18/15/5	1	1.25	1
q19/13/エ	1	1	1
q19/13/オ	8	7	4
q19/16/1	1	1	1
q19/16/2	1	1	1
q19/16/4	8	6	5
q19/16/5	1	1	1
q19/7/3	44	44	44
平均	6.42	6.27	5.75
F 値平均(3 位基準)	0.358	0.358	0.358

条文構成の基礎と語彙選択

民法に限らず、条文の文書は一般的に(法律)要件と(法律)効果のそれぞれの部分によって構成されている[13](p.7). すなわち、要件と呼ばれる当該規定の成立条件が充足されたときに、効果と呼ばれる(民法であれば)権利関係の設定や(刑法であれば)刑罰の設定が起こるという形になっている。例えば以下の条文(民法第709条)では、これはかなり単純な例だが、緑色でハイライトした部分が法律要件、青色でハイライトした部分が法律効果である。また、これはテキストの上に限ったことではなく、法ルール文の抽象的な論理構造に関する想定されるものである[4](p.150).

(不法行為による損害賠償)

第七百九条 故意又は過失によって他人の権利又は法律上保護される利益を侵害した者は、これによって生じた損害を賠償する責任を負う。

図8: 法律要件と効果

前段では、条文が過不足のない表現をもって記述されており、語彙毎の重要度は文書内の出現頻度によっては判断し難いという検証がなされた。ところでベクトル空間モデルでは、文書におけるある種の主題が、その全体の傾向として語彙集号の出現に現れるということを前提としている。本研究の設定タスクにおいては、これは問題文および条文に現れる主題全体の傾向を比較するということとなる。

$f_{t,d}$ の代わりにを $\delta_{t,d}$ を採用することは、各語彙について、文書主題への寄与度がその出現回数には関わらず、単に出現したかしないかに依存しているという仮説を採用していることもある。一方で $\delta_{t,d}$ を導入したとしても、問題文と条文に含まれる主題全体の傾向を比較するという前提には変更がない。これは、対象の条文文書集合から、問題文で問われていることと同等の主題についての規定を含むものを探しだす、というタスクの性質にあってはいるだろうか？

専門家へのインタビューによれば、短答式試験において、おおよそ全ての問題文において問われるはある法律における单一の規定や判例の知識であるという。従ってこの設定課題における主題とは、一つ一つの規定、すなわち要件と効果の組み合わせを単位に扱うことができれば都合が良いだろう。

しかしながら制定法の条文は1本につき、ある制度のある側面、ある法律概念のある属性等に限定しつつも、その状況下でのいくつかの規定が盛り込まれているのが普通である。例えば以下の例では、箇条書きを用いて複数の要件に対して一つの効果を設定している。青色のハイライト部分は法律効果部であるが、これに対し、紫色のハイライト部分を共通として、トーンの異なる3つの緑色のハイライト部分を組み合わせることにより、3つの別々の要件を見つけることができる。

(根抵当権の被担保債権の範囲)

第三百九十八条の三(第二項から抜粋)

2 債務者との取引によらないで取得する手形上又は小切手上の請求権を根抵当権の担保すべき債権とした場合において、次に掲げる事由があったときは、その前に取得したものについてのみ、その根抵当権を行使することができる。

- 一 債務者の支払の停止
- 二 債務者についての破産手続開始、再生手続開始、更生手続開始又は特別清算開始の申立て
- 三 抵当不動産に対する競売の申立て又は滞納処分による差押え

図9: 複数の要件-効果対による条文

このような条文に対して、以下の問題を指摘することができる。

ある条文の含む規定群がそれぞれ主題であると考える場合、たとえそれらの差異が僅かなものであったとしても、またこれらの主題にオーバーラップする部分があったとしても、これを混然一体とした条文の主題として混合し、それと問題文(これは単一の規定を主題としている)の類似度比較することは、操作として迂遠である。あるいは仮に思考実験として、問題文で問われているある規定について、それが多くの規定を含む条文にある場合と、その規定のみで構成される条文の中にある場合を比較したとき、その類似度に差が生じるのは不都合であり、タスクの目的に合わない。

従ってここでは、ある条文と問題文とを比較する場合に、予め条文が個々の規定に分解された上で行うことができれば理想的である。条文テキストにおける要件-効果対、あるいは要件・効果それぞれのセグメントを切り分ける研究には[17]等があるが、本研究に適用するにはコストが高い上にコントロールが難しいため、適さないと言える。

ベクトル空間モデルにおいては、主題が語彙集号によって表現されるという前提があることに注目する。すなわち問題文がある主題 $tq_1, tq_2, \dots, tq_{N_q} \in T_q$ に関するものであったとき、問題文は少なくともこの部分集合を含むであろう。一方 $T_q, T_{a1}, T_{a2}, \dots$ によって構成される条文があるとき、問題文と条文の類似度はこれら T_{a1}, T_{a2}, \dots とは無関係に、 T_q の部分集合としての問題文と条文に含まれる T_q そのものを比較できればよい。

実際、問題文とその正しい関連条文の関係を見てみると、少なくともベースライン手法では、検索順位が低いほどその中の語彙が(相手の文書には語彙が出現していないために)計算に用いられない傾向が若干ながら見て取れる。すなわち、言ってしまえば正しい関連条文にも関わらず余分な語彙が多く含まれているケースの検索が上手く行っていない。以下では、これを視覚化している。

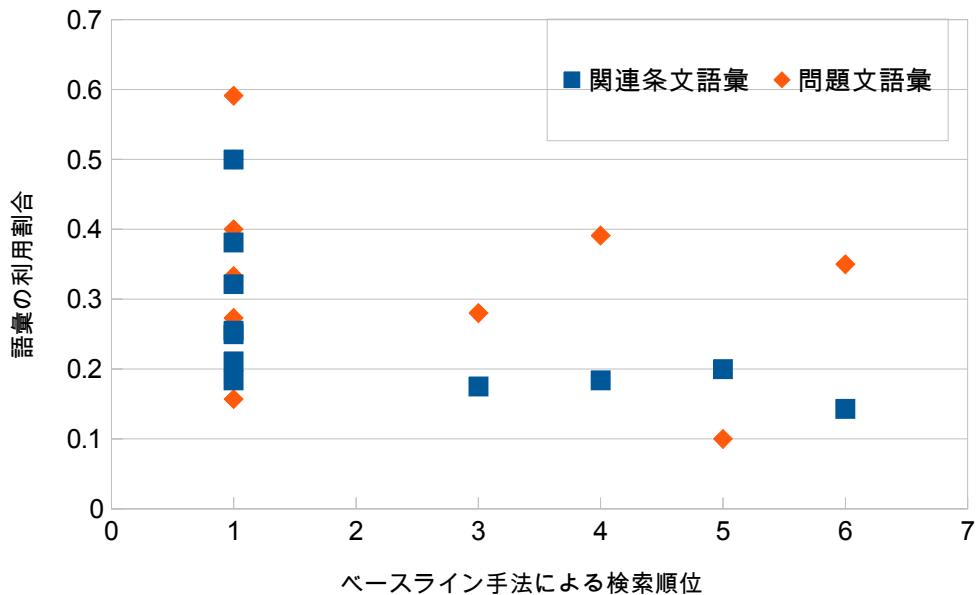


図 10: ベースライン手法による検索順位と語彙の一一致割合

これを、以下のような操作で実現する。すなわちある語彙 t について、条文の語彙出現フアクター $\delta_{t, \text{article}}$ を、類似度計算を行う前に問題文の $\delta_{t, \text{question}}$ を用いて

$$\delta_{t, \text{article}} = \begin{cases} \delta_{t, \text{article}} & (\text{if } \delta_{t, \text{question}} > 0) \\ 0 & (\text{otherwise}) \end{cases} \quad (13)$$

とする。あるいは別の言い方をすれば、条文の出現語彙 T_{article} は、問題文の出現語彙 T_{question} を用いて事前に $T_{\text{article}} = T_{\text{article}} \setminus T_{\text{question}}$ に改変されるということである。

これにより問題文と条文は、問題文に含まれる語彙集号に含まれる語彙が、あくまで条文内にどの程度再現されるかによって比較され、条文内の無関係な語彙の影響を排除できる。この証明を、付録Bに示しておく。これはすなわち、問題文に”余計な語彙”が含まれていないと仮定すれば、タスクの意に沿った操作となる。

この操作を行うことによる改善効果を、つぎに示す。ここでは、語彙出現のファクターとして $\delta_{t,d}$ を用い、idf重み等についても条件を変えず、上記の操作の有無のみを比較したものである。先に複数の要件-効果を含む条文として紹介した条文(398_3)は実のところ、表2行目の **q18/15/2** の問題の関連条文として設定させていたものだが、この問い合わせの改善効果は高く、以上で述べた仮説について、ある程度検証できたと言って良いだろう。ちなみに検索順位の著しく下がっている q19/16/4 については、おおよそ関連条文以外の条文の類似度指標が上昇したことによる効果が現れている。

表 8: 語彙のフィルタリング操作による検索性能の改善効果

問題文	操作なし	操作あり
q18/15/1	1	1
q18/15/2	6	2
q18/15/3	3	1
q18/15/4	1	1
q18/15/5	1	1.875
q19/13/エ	1	1
q19/13/オ	4	1
q19/16/1	1	1
q19/16/2	1	1
q19/16/4	5	8.5
q19/16/5	1	1
q19/7/3	44	44
平均	5.75	5.36
F 値平均(3 位基準)	0.358	0.442

3.2.4 語彙ネットワークを用いた語彙拡張

これまで述べた手法による検索課題の実施結果を見ると、一件どの手法においても 44 位という完全なる圏外に位置づけられる問題文があることがわかる。以下は付録 A より抜粋加工したこの問題文である。緑色のハイライトで取得された語彙を示した。注意深く見ると、問題文と関連条文の間に一致するごいは、一つも存在しないことが分かる。

問題文

次の 1 から 5 までの各記述のうち、正しいものを 2 個選びなさい。

家具の所有者 A が B に賃貸中の当該家具を C に売却した場合、特約がなければ、C は、直ちにその所有権を取得するから、B に対する賃料債権も、C が売買契約時に取得することになる。

関連条文

(果実の帰属及び代金の利息の支払)

第五百七十五条 まだ引き渡されていない売買の目的物が果実を生じたときは、その果実は、売主に帰属する。

2 買主は、引渡しの日から、代金の利息を支払う義務を負う。ただし、代金の支払について期限があるときは、その期限が到来するまでは、利息を支払うことを要しない。

図 11: 関連条文との語彙一致が全く見られない問い合わせ

44 位というのは、類似度が 0 である条文の順位であることが分かる。このケースの検索順位を向上させるには、第 2 章で概説を行った、語彙の補いを中心としたクエリー拡張が必須である。

クエリー拡張のための、語彙知識・法律知識を双方含む語彙ネットワーク

本研究では、筆者が人力で整備した語彙ネットワークを利用し、問題文および条文の出現語彙をもととしたリンク解析によって語彙の拡張を行った。

第2章で述べた通り、法律ドメインのオントロジーやシソーラスはこれまでにも提案されてきた([18]など)。しかしながら現時点では日本語において、特に民法分野の法律概念とそれらの間の関係を、十分かつ適切に含んだ大規模なシソーラスやオントロジーは整備されていない。特に本研究では、自然言語レベルの法律概念の表記に十分に対応できることが求めらるが、そういうたシソーラスのような性質を持つものは補助的な役割に留まりがちである[5]。これらのことと、以下で述べる理由と併せて人手で整備する必要性に行き当たった。

まず、本研究の設定タスクには、検索対象の文書集合があまりに非冗長であるという特徴がある。それぞれの文書、すなわち各条文が、意味内容の冗長な記述を嫌うことは既に述べた。これに加えて、制定法はそれに固有なもので、”ある条文について述べられた条文”というものは数多くあるにせよ、条文自体のサンプルはそれ一つに限られるという制限がある。これにより、民法全体を測定範囲とすればまだしも、条文や章立て毎に区切りを行った場合、語彙同士の関係やその重要性を正しく表現するような語彙の出現情報を採ることが難しくなっているのである。あるいは、判例や参考書等の法律文書を膨大に集めることでこれに対応できると思われるかも知れない。しかしながら少なくとも民法の各条文は、参照される機会が多いものと少ないものの差が大変大きい。”親族間の扶(たす)け合い”に関する民法第730条に法律的な義務がなく、少なくとも訴訟等の場で参照されることは決してない[15](p.421)というのは良く知られた話である。

これは実のところ、機械学習を活用した設定タスクの実施を難しくしている一因である。あるいは単純に、問題文から条文に対する全ての関連付けのモデルが等価であると考えて学習を行えると考えられなくもないが、これは下で述べる、法律の体系的知識の必要性から妥当な取組みだとは言い難い。本研究は設定タスク、および法律ドメインのテキスト処理において、構造化知識の利用可能性と限界を与えることを一つの目的としている。

語彙ネットワークを人手整備するもう一つの理由として、クエリー拡張に必要な語彙の補いには、明らかに法律の体系的知識なくしてモデル化することができないものが含まれているという特徴を挙げることができる。

例えば先に上げた一致語彙の存在しないケースでは、”果実”という語彙に対する理解が非常に大事なファクターとなっている。これは、ある財産を所有等している場合に、その財産から副次的に生じる財産、すなわち土地から生ずるタケノコや、不動産を賃貸することによって得る賃料等を総称する語である。本研究では、自然言語の深い意味理解の技術の制限もあって、あくまで語彙集合とその類似性を単位に検索課題にあたることとしているが、これをこのケースのクエリー拡張で実現するとすれば、問題文中の”賃貸”物である”家具”の”賃料”は”果実”であり、かつ”売却”か”引渡し”の概念を含んでいるといったことを語彙の組み合わせから関連付け、必要語彙を補う必要がある。

以上で述べたことはある意味、このタスクが法律分野の自然言語表現に対して構造化知識の利用を検証することに適しているということを保証してくれるかもしれない。ともあれ、これ以後、問題文と条文の自然言語表記に耐えうる語彙知識、および法律知識に基づいた概念間関係をサポートする語彙ネットワークの開発に関して、述べていく。

以下ではここまで議論を受けて、クエリー拡張のための語彙ネットワークとはどのようなものであるべきか論じ、本研究において構築された語彙ネットワークについて特徴づけを行う。

構造化知識の構築にあたって最も留意すべきは、それが一貫した利用目的に従ってなされるべき[19](p.9など)であるということである。従ってはじめに、この語彙ネットワークの大まかな利用方法を概観する。その後に、語彙ネットワークの特徴およびスキーマについて述べていく。その後、この利用に関する工学的な操作について詳説を行い、手法の提案とする。

語彙の潜在的関係とクエリー拡張

まずは本研究で、クエリー拡張としての試験問題文および条文の語彙の補いにおいて、語彙の関係性のネットワークが具体的にどのように利用されるべきであると考えたかを述べる。

まず、先の述べた”果実”と”賃料”的の関連付けの例に関して、本研究ではこれを、語彙の潜在的な想起関係の、与えられた文書コンテキストにおける生起の現象であると分析した。

語彙の潜在的な想起関係とは、すなわち以下のようなものである。ある語彙、例えば”賃料”が与えられたとき、これを民法の知識をもつ者がみれば、賃貸借契約において、貸主が借主に対して定期的に請求することのできる対価であるということはただちに知るところであろう。これは”賃料”という語彙それ自体が含んでいる定義的な知識が想起されることによるものである。一方、もしこの者が”賃料”という語を、賃貸することによって”賃料”を発生せしめている土地等がみずから財産であるという観点から見た場合、あるいは少なくともそれを促すようなヒントを与えられた場合、”賃料”はその財産にとって、それを損なうことなく副次的に産出される価値であるところの”(法定)果実”であると知る所となる。このように、所与の視点・観点の導入によって想起、関連付けされるような語彙の関係を、ここでは語彙の潜在的な想起関係と呼んでみる。

語彙ネットワークの活用にあたっては、このような潜在的な想起関係をもつ語彙同士を関連付けられることが、一つの条件となるだろう。ここで上の例は、その”視点”こそが肝要であって、単純な語彙の上位・下位関係ではないことに注意されたい。法律知識は無論の事ながら、上の例のように、コンテキスト依存で流動的な視点の導入が無ければ明確な関係として捉えられないものについて、ここでは論じているのである。ただし、これは通常の上位・下位関係を内包するものであるとも言える。

またここで概念同士の関係ではなく語彙同士の関係とした理由は、自由に記述された自然言語レベルの意味理解において、こういった概念の関連付けが、語彙の意味内容=概念の確定、概念間の関係の洗い出しと探索、可能な関連付けからの推論による関係の確定、というような内的なプロセスで全て行われるとは、少なくとも筆者には思えないためである。従って、語彙レベルで潜在的な関係を関連付けられるというのが、ここで述べた語彙ネットワークの利用条件である。

また、クエリー拡張としての語彙の補いにおいては、以下の典型的な問題に対処する必要がある。すなわち第2章でも述べたように、補完の誤りや語彙の出現コンテキストが考慮されないことによる適合率の低下を防ぐことである。具体的には、これまでの手法で関連条文に適切な順位付けがなされたケースについて、この検索性能の妨げとなるような操作を行わないことである。

以上を考慮して、本研究における語彙ネットワークが貢献すべきプロセスに、2つの選択肢を想定してみる。すなわち、いずれも上記の潜在的な想起関係を考慮しながら、

A 与えられた語彙をもとに、語彙の補いを許容しながら、これらが整合的かつ密に関連づいたローカル・ネットワークをそれぞれの文書で構成し、その結果(語彙の集合である)同士を比較する

B それぞれの文書の語彙集合間で、互いの想起関係の可能性を探索しながら、文書間に密な関連付けを行い、その結果として語彙の補いを行う

本研究では、ベースライン手法の最も重要な基礎をなすベクトル空間モデルと語彙の出現ファクターとの相性がよいこと、加えて語彙ネットワークがある種のグラフであることを鑑みて、語彙の想起関係の探索には相当の計算コストを要するであろうことを慮り、Aのアプローチを選択することとした。

ところで、まずどちらの選択肢に関しても、語彙の潜在的な想起関係が語彙ネットワークより提供されるべきであるということは共通した要請であることが分かる。以下ではまず、語彙ネットワークに”潜在的な想起関係”として相応しいと考えられるリンク属性について提案を行う。これらは、基礎的なオントロジーあるいはシソーラスに用いられているアイディアを意識しながら、より法律知識の形態に特化して述べるところの[20]を参照しつつ、より自然言語レベルの語彙表記に対応するべく独自に定義したものである。

語彙ネットワークのリンク属性

以下より、語彙ネットワークの構築の際に定義・導入した各リンク属性について詳しく述べていく。

hyper：上位・下位リンク

これは、一般的なオントロジーに見られる概念間の上位・下位関係と同等である。すなわち、語が示す概念の具体例や具体的な状況、すなわちインスタンスについて、下位の概念の全てのインスタンスを集めた集合は、上位概念のそれの部分集合となっているという外延的定義によって見つけることの出来る関係である。

また、この語彙ネットワークはシソーラスの役割も期待されているため、WordNet等の語彙知識ベースにおける上位語・下位語(hypernym/hyponym)の一部をカバーする。一部というのは、こと語彙の関係における上位・下位関係では上記の外延的定義に正確に当てはまるとは言えないものがある為である。これについてはhyperxリンクの解説で詳しく述べる。

以下、実際の語彙の関連付けの例である。

表 9: *hyper* リンクの使用例

リンク元語彙	リンク先語彙
家具	動産
地上権	用益物権
被保佐人	制限行為能力者
取消し	単独行為

hyperx：役割継承リンク

これは上位・下位リンクと非常に似ているが、より自然言語的な性質の強いものである。この語彙関係は、下位の語彙の表す概念が上位の語彙の表す概念に対して、何らかの付加的な役割を伴って現れた状態を示している。

このような概念関係について詳しく述べているものとして、[21]がある。ここでも挙げられている解りやすい例として、”番犬”と”犬”的関係がある。”番犬”はもちろん”犬”的意味内容を含むが、「”番犬”は”犬”である」という文に違和感がない一方、「”番犬”は”犬”的一種である」という説明は馴染まないことが分かるだろう。

これには2種類の説明を付加することができる。まず一つは”番犬”という概念が、なんらかの”犬”的一時的に”番”という役割を担っている状態に対する單なる呼称であるというものである。外延的定義の面から考えても、すなわち”番犬”とは、何らかのインスタンス集合を示す概念としてはあまりに不安定であり、これを独立した概念クラスとして切り取るのは相応しくないということが言える。

では、”番犬”という呼称、およびそれが示す確かに概念であるはずの何かと、整理された概念階層との関係性は、一体何であろうか。

もう一つの説明を行えば，“番犬”を“犬”的個体からではなく、その語彙が示す概念自体を起点として見ると、今まさに家屋等の“番”という役割を“犬”に対して強いる／強いようとしている特定の人間の視点にとって，“番”に使えそうな“犬”はおおよそ、いつでもおしなべて“番犬”に見えるといって良いだろう。一方このとき，“犬”が“哺乳類”的一種であるとか“食肉目”的一種であるといった上位・下位関係はほぼ無視され、家屋の守護という役割に応じた働きに応じた価値が前面に現れる。加えて，“番”を行うことのできる“人”（“番人”）との関係は兄弟クラスのように近くなる一方，“猫”（“番猫”？）との関係は遠くなるといった特殊な分類階層の再編が行われる。

すなわち、この視点から見れば、逆に通常の上位・下位関係の外延的定義が揺らいでしまうのである。ここでむしろ通常の上位・下位関係は、“実体としての概念の分類”という観点から見た意味役割の階層であるということもできる。すなわちこの関係は、逆に自然言語的な語彙関係においてオントロジー等に見られる上位・下位関係を明らかにすることで理解を得ることができ、より一般的な概念の汎化関係に対して視野を啓くものである。

hyper リンクとの分類をこのように厳密に行つたのは、自然言語の多様な振舞いの一部である語彙間の関係と、純粋な思考実験的、あるいは明確な目的に従った仕様化の側面を持つ概念間の関係には相応の隔たりがあると考えられる為である。ただし本研究においては、これらのリンクの相違によって語彙の補いにおける扱いを有効に切り替える方法論は見つけられていない。

ところで全く別の例えをすると、このリンクのリンク元語彙とリンク先語彙の関係は、Javaなどの体系的なオブジェクト指向表現力をもったプログラミング言語において、ある普通のクラス C、およびそれとは無関係のインターフェース I を継承した C のサブルー S の関係の類似するものである。これは工学的な観点からこの関係を区別することは、決して無意味ではないということを示唆していると言えないだろうか。

以下に、実際の語彙の関連付けの例を示す。

表 10: hyperx リンクの使用例

リンク元語彙	リンク先語彙
瑕疵	状態
後順位	順位
永小作人	人
売却	譲渡

sbj/obj：主格および目的格リンク

これは、フレーム意味論[22]の文脈で用いられる主格および目的格の概念を、さらに自由に拡大解釈したものである。例えば sbj リンクによって、名詞である“担保権”と“担保権者”が関連付けられている。

以下に関連付けの例を示す。

表 11: sbj リンクの使用例

リンク元語彙	リンク先語彙
相続	相続人
利益	受益者
抵当権	抵当権者
被担保債権	担保権者

表 12: *obj* リンクの仕様例

リンク元語彙	リンク先語彙
所有	所有物
物権	目的物
賠償	損害
競売	財産

auth : 法的な根拠づけのリンク

これは、法律の制度的な側面や権利を表す語彙に用い、これが実質的に根拠づけている内容を示す語彙との関連を示すものである。便宜的に auth_by という逆向きのリンクを用いる場合がある。

以下に関連付けの例を示す。

表 13: *auth* リンクの使用例

リンク元語彙	リンク先語彙
不動産登記法	登記
請求権	請求
求償権	償還請求
対抗力	対抗

within : 場所リンク

場所を示す語彙について、かなり限定的な場面で用いる。以下に例を示す。

表 14: *within* リンクの使用例

リンク元語彙	リンク先語彙
日本人	日本
生活	居所
履行	履行地
住所	住所地

attr_slot : フレームの予約属性へのリンク

これは、ある概念にとって、それをフレームの中心として見た場合に想定しうる属性、あるいは付随する概念等にかなり広く用いるものである。以下に実際の関連付けの例を示す。

表 15: attr_slot リンクの使用例

リンク元語彙	リンク先語彙
登記	順位
物	用法
管理者	注意義務
利息	利率

antecedent_to : 前提事項のリンク

このリンクは、何らかの概念が他の概念を順当に導くと考えられる場合、あるいはより具体的に、何らかの出来事や手順が他の出来事の前に必ず起きる、あるいは必要であるという場合に、これらに対応する語彙を関連付けるための、比較的法律知識の現れやすいリンク属性である。これには、因果関係のようなものも含まれうる。ここではリンクの方向の有用性を考慮して、前提条件という意味づけのリンクを導入している。

以下にリンクの使用例を示す。

表 16: antecedent_to リンクの使用例

リンク元語彙	リンク先語彙
要件	効果
担保権	競売
収益	果実
競売	買受け

語彙ネットワークの構築方法

ここで、語彙ネットワークの構築の方法に関して軽く触れておく。

まず登録すべき関係の洗い出しへは、先に述べた通り、分析のために選択した問い合わせ、その関連条文、そしてベースライン手法において誤って上位に配置された条文をもととした。3.2.1の手法で収集した語彙リストをもとに、これらに登場する語彙それぞれが文書内でどのように関連付けられるか考慮しながら、ゼロベースで関係を定義していった。またこれに先立ち、あらゆる民法の概念が依って立つ所の基礎の提供を期待し、民法の第一編：総則、第二編：物権の第一章：物権総則等から抜粋した条文群より、同様の関係収集を行った。

一方、構築作業については、視覚化に NetworkX¹⁶を用いながら、簡単なリンクの検証等を含む、独自に開発したプログラムを用いて行った。構造化知識、特にオントロジーの構築支援を行うソフトウェアには、特に法律オントロジーの構築に際して開発された DODDLE¹⁷[23]等高機能なものが利用可能であるが、今回の語彙ネットワークは通常のオントロジーと異なり、語彙表記への対応をその機能に含む特殊なものであり、半自動構築等の恩恵をあまり受けられないことから、これを見送った。

この方法においては、上に述べたもの以外にも視覚化を含むあらゆる支援を一切受けることができないが、ネットワークの定義を全て YAML¹⁸による、簡易なテキスト情報で記述することができる。これは構築段階における定義の差分を抽出するのに役立つ他、構築者にとって自由な書式の整形を許している。プログラムに関しては、語彙ネットワーク本体の定義とともに筆者の GitHub アカウント¹⁹にて近日公開予定である。

16 [Overview — NetworkX(2014年2月1日参照)：<http://networkx.github.io/>]

17 [DODDLE プロジェクト - ホーム(2014年2月1日参照)：<http://doddle-owl.sourceforge.net/ja/>]

18 [The Official YAML Web Site(2014年2月1日参照)：<http://www.yaml.org/>]

19 [<https://github.com/drowse314-dev-ymat>]

語彙ネットワークによる語彙の補完

以上で述べた方法により構築した語彙ネットワークを用いて、語彙の補完を行う方法について提案を行う。

ここで当初の方針を思い出せば、これは各文書から取得された語彙の「整合的かつ密に関連づいたローカル・ネットワーク」の構成の結果として行われるものである。本研究では、これを語彙ネットワークのリンク解析／リンク探索によって、潜在的な関連付けのうち必要なものを効率的にしていくことで達成することとし、そのアルゴリズムについて以下の方法で検討を行った。

すなわち、問題文の数件(具体的にはq18/15/1とq18/15/2)およびその関連条文、そしてベースライン手法において誤って上位に配置された条文群について、ひとまず人の手と知識判断力によって、語彙ネットワークに基づき理想的なローカル・ネットワークを構成してみる。これは以下の図のようなものである。続いて、その成果物を通して語彙の関連付けや補いのパターンに頻繁に見られ、かつ理論的に妥当な説明を行うことができるものを選出し、これをナイーブに実装した。

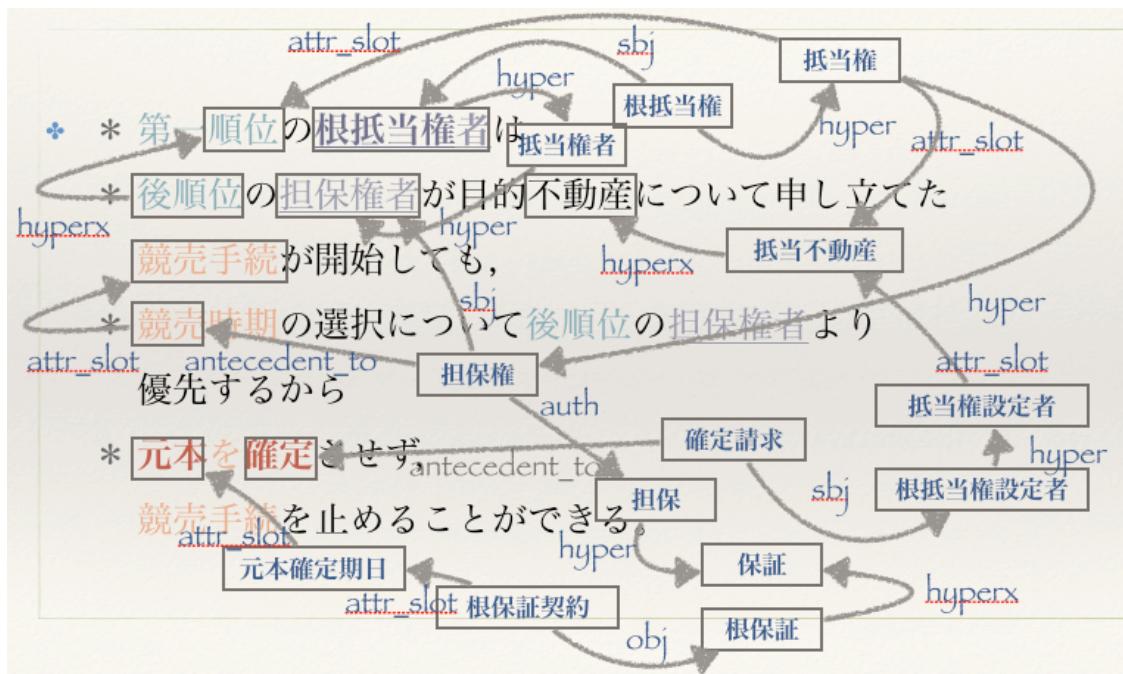


図 12: 語彙のローカル・ネットワーク構成例

リンク解析によるノード補完の形式分析

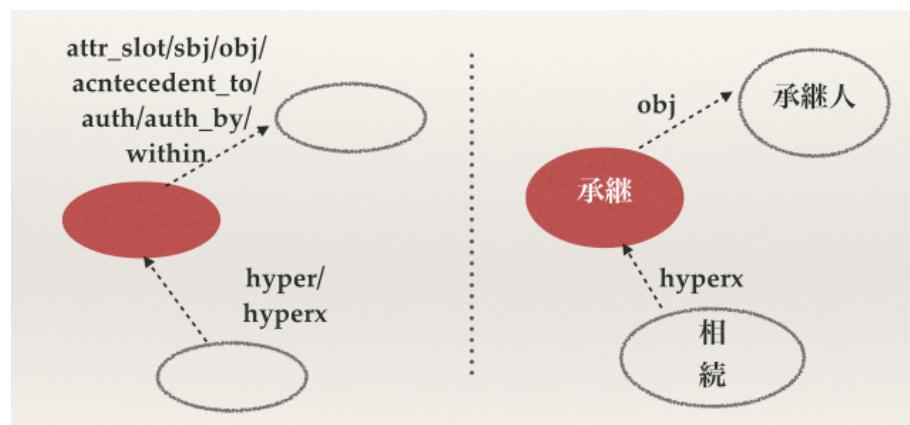
以下より、調査によって見出され、かつ妥当性が確認されたノード補完のパターンについて、それぞれ詳しく見ていく。

hyperframes：上位概念のフレームの参照

これは、ある概念と、その上位概念のフレーム属性のどれかが出現していた場合、中間に位置する上位概念を補完するというものである。

上位・下位リンクは外延的定義からも分かる通り、下位概念のインスタンスがその上位概念のインスタンスでもあるということを示している。hyperx リンクに関してはこの関係が常に成立しないということも先に述べたが、これは自然言語の世界を軸に考えた場合 hyper リンクとて同様であるので、いずれにせよあるコンテキストにおいて、置換可能であるということである。これを受けて、上位概念のフレーム属性の現れを置換すべきコンテキストと捉え、これを結ぶことは理論的に妥当である。

以下に例を示す。赤色で塗りつぶしたノードが補完されるもの、白抜きのノードは実際に出現した語彙を示す。



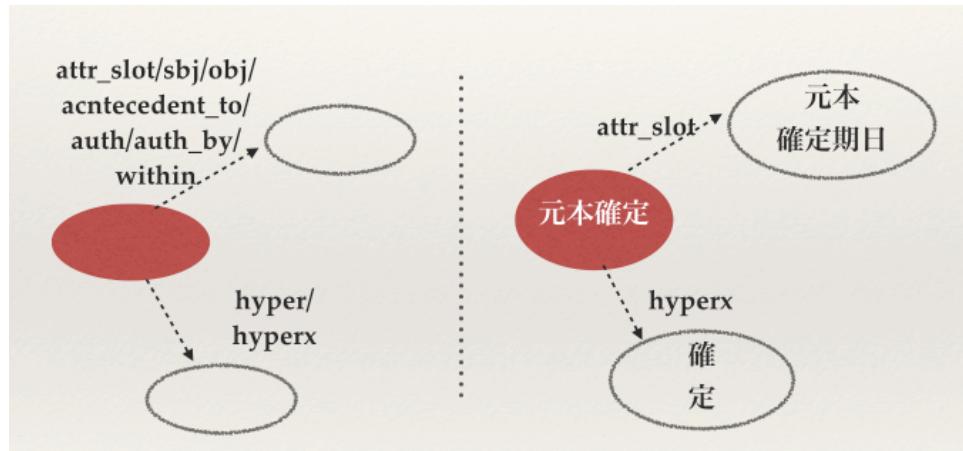
(左：模式図／右：例)

図 13: ノード補完パターン : *hyperframes*

hypoframes：下位概念のフレームの参照

これは、hyperframes とは逆に、ある概念の下位概念のフレーム属性をコンテキストとして、より具体的な語を導くものである。これは、hyperframes で述べた置換が既に起こり、もとの具体的な語彙が省かれたケースと見ることができる。

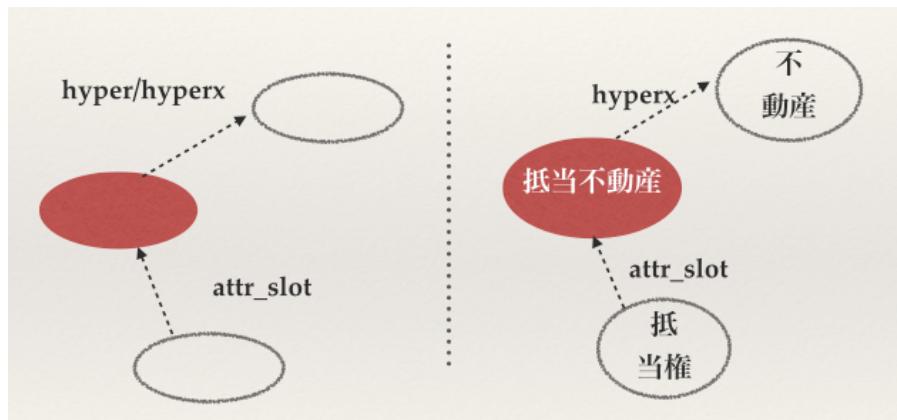
以下に例をしめす。



(左：模式図／右：例)

図 14: ノード補完パターン : *hypoframes***attr_hypers**：フレーム属性の上位概念の参照

これは、ある概念とそのフレーム属性の上位概念が現れた場合に、このフレーム属性を補うものである。hypoframe とほぼ同様の説明を加えることができる。以下に例を示す。



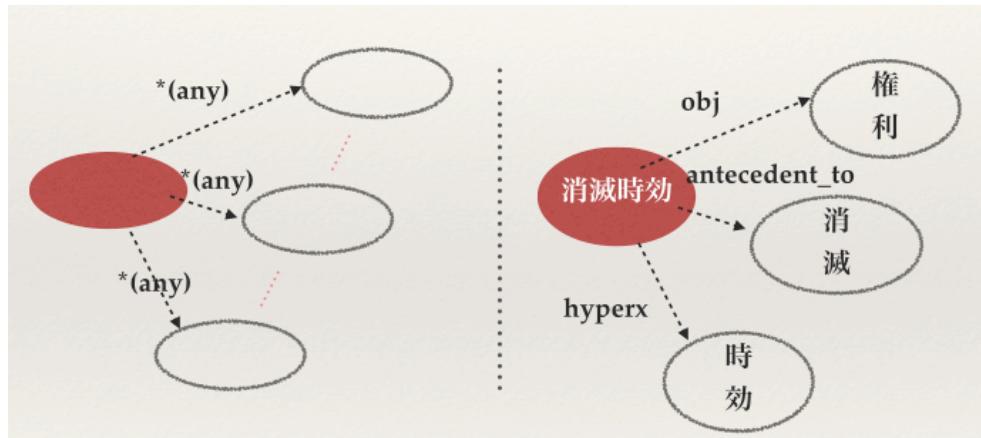
(左：模式図／右：例)

図 15: ノード補完パターン : *attr_hypers*

prerequisites：複数の概念の背後にある概念の発見

これは、リンク属性に関わらず様々な出現語彙のリンク元となっている語彙を補完するものである。これには特に厳密な理論は想定しないが、おおよそ整合的な関連付けが多量に得られるという点で、妥当であると言って良いだろう。

以下に例を示す。

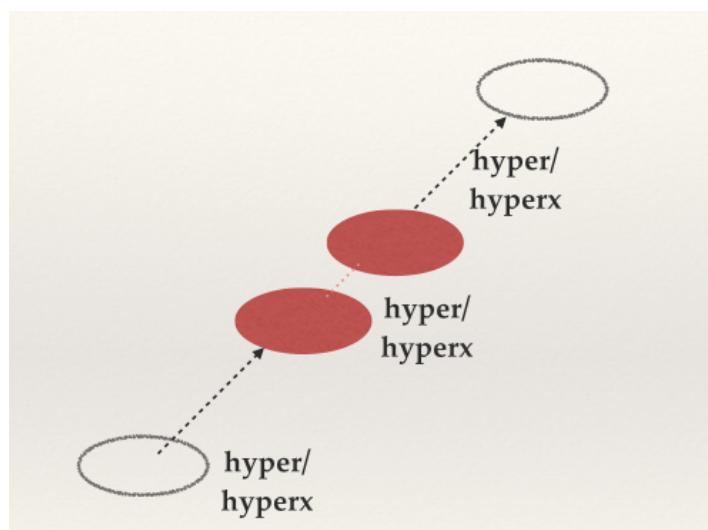


(左：模式図／右：例)

図 16: ノード補完パターン : *prerequisites***ascending_ways**：概念階層の補完

これは、階層の離れた上位・下位概念の間を埋めるための補完である。一般的に、上位・下位関係は階層が離れていても特にその外延的定義に影響を及ぼすものではない。従って、hyperframes と同様の説明で事足りるであろう。

以下に模式図を示す。

図 17: ノード補完パターン : *ascending_ways*

ascendedhubs : 概念階層を跨いだ背後概念の発見

これは、以下の模式図に示す通り、`prerequisites` と `ascending_ways` を組み合わせたような補完の方法である。それぞれの妥当性を鑑みれば、この方法に関しても認めて良いと言えるだろう。

以下に模式図を示す。

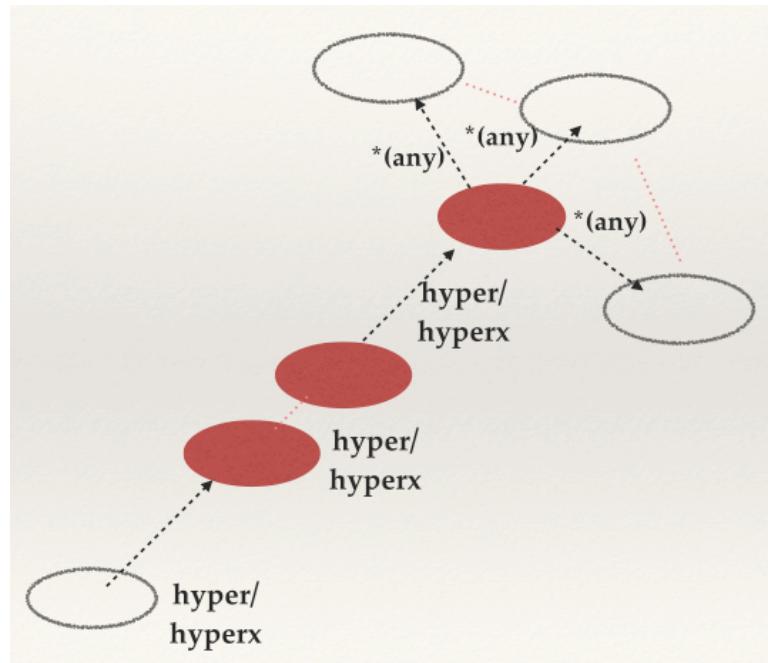
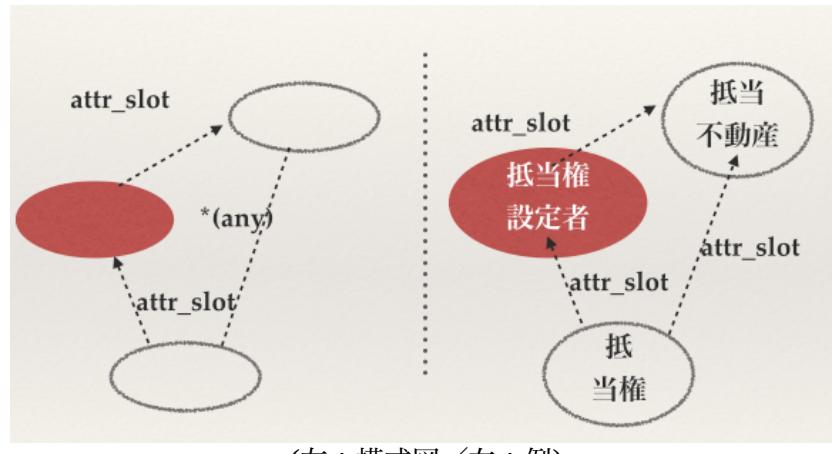


図 18: ノード補完パターン : *ascendedhubs*

attrs_of_attrs : フレーム属性の連鎖

これは、ある語彙が出現語彙によってフレーム属性リンクで関連付けられ、かつ出現語彙間にも何らかの関係があるような場合、この語彙を補完するというものである。これについては、`prerequisites` と同様、整合的な関連付けを一度に3本得られるという点で妥当性があると考えた。

以下に例を示す。



(左：模式図／右：例)

図 19: ノード補完パターン : attrs_of_attrs

補完された語彙を用いた類似度の計算

これまで述べた操作によって語彙を補完した文書対において、元のベクトル空間モデルの枠組みで類似度を計算する方法を与える。

語彙ネットワークによるローカル・ネットワークの構築のうち、ここに補完された全ての語彙を用いてただちにベクトル空間モデルによる計算を行うこともできるが、本研究では尚、語彙ネットワークの導入効果の最大化、あるいはこれまでの仮説との整合性の確保のため、いくつかの改良を行うこととした。以下、語彙抽出、語彙補完と選択、重み付けの順に、クエリ拡張のない手法に対する変更を論じていく。

またその後、最後に語彙の汎化関係における一考察と補完語彙の見直しを行い、語彙ネットワークを用いた手法の提案とした。

語彙ネットワークと語彙抽出

本研究で構築した語彙ネットワークは、その語彙あるいは概念の粒度を、3.2.1で論じた方法で民法全文から抽出した語彙集号に拠っている。一方、これらの語彙集号に含まれていながら、出現語彙との関連付けによって、他の法律語彙に関する有効な”潜在的な想起関係”的の発見に特に寄与することはないとすることでネットワークには登録されなかった語彙、あるいは”潜在的な想起関係”を発見するために出現語彙のネットワークの背後で有効に働くということで、新たに追加された語彙がそれぞれ存在する。

ここでは、以下の方法で語彙ネットワークの導入効果を最大化する試みを述べる。すなわち、語彙ネットワークによる潜在的な関連付けの発見やそれによる語彙の補完は、語彙ネットワークに含まれる語を文書から得られてこそ働くものである。3.2.1の手法において、元の全文からの抽出語彙とネットワークの登録語彙の和集合から語彙を抽出する。このとき、前方一致の比較過程において、ネットワークに登録されている語彙の最長表記を優先し、その中で見つからない場合に元の抽出語彙の最長表記から比較を始める。ストップワードの扱いに変更はない。

以下では語彙抽出に関して、これまでの手法の語彙とネットワーク語彙の和集合を用いながら、先に述べたネットワークの語彙を優先する操作を行うものと行わないものを比較している。出現語彙に関して問題文により条文の語彙集号をフィルタリングするのはこれまで通り、idf重みづけに関してもこれまで通りである。先に述べた工夫は非常に細かいものであるが、確かに僅かな改善効果が見られる試みであることがわかる。因みに、語彙ネットワーク自体の導入効果については後の章に譲ることとする。

表 17: 語彙ネットワークの効果を最大化する語彙抽出

問題文	操作なし	操作あり
q18/15/1	1	1
q18/15/2	1	1
q18/15/3	6	6
q18/15/4	1.5	1.5
q18/15/5	2.5	2.75
q19/13/エ	2	1
q19/13/オ	1	1
q19/16/1	1	1
q19/16/2	1	1
q19/16/4	9	9
q19/16/5	2	2
q19/7/3	4	4
平均	2.67	2.60
F 値平均(3 位基準)	0.367	0.367

語彙補完と語彙のフィルタリング操作

続いて、3.2.2 の中で述べた語彙のフィルタリング操作に関して、この操作の背景にある仮説と語彙ネットワークを用いた語彙の補いが、どのように理論的な整合しうるか論ずる。

3.2.2 の語彙のフィルタリング操作は、文書、特に条文における語彙出現の捉え方を、関心をもって比較の対象とする問題文の語彙集合によって歪めるものである。これは先にも述べた通り、ある語彙の集合が何らかの主題、すなわち法律における規定が話題とするところのものを表現していると仮定してのことと、フィルタリング操作は問題文語彙が表現しようとしている主題のバイアスで、条文を観察することを目指している。

この考え方は、手法として選ばなかった **B** の選択肢、すなわち「それぞれの文書の語彙集合間で、互いの想起関係の可能性を探査」することに少なからず通じている。そしてこれは、想起関係の可能性の探索が、互いに相手の文書のコンテキストに拘束されることを暗に示している。一方で、**A** の各々の文書の語彙集合に対して、語彙ネットワークにより潜在的な想起関係を可能な限り密に繋ぐよう探索することは、その文書主題を最も大きな語彙集合によって表現しようとすることと等価であると言える。これはもし、問題文がある主題(これは単位としては一つの規定であった)の、さらに断片に関して問うているという場合に、余計な情報を追加することに繋がる。

上記のBの選択肢に生きる、語彙のフィルタリングの手法の指向性をさらに発展させた考え方を、都合良く提案手法に取り込めないだろうか？

ここでは、問題文語彙の補完可能性と条文語彙の補完可能性が、協同してこの「互いの想起関係の可能性を探査」する状態を、擬似的に以下のプロセスで表現できると仮定してみる。すなわち、各々の文書の語彙集合について、普通に提案手法で補完を行った後に、その語彙の拡張部分を、以下のいずれかに該当する語彙以外切り捨てる。

1 相手の文書の元々の語彙集号に含まれている

2 相手の文書の語彙の拡張部分に含まれている

これによって、ミニマムかつ相手の文書のコンテキストを考慮した形で、主題の想起関係を調べられるのではないかと考えた。

以下はこの操作を実践したものである。語彙の拡張方法以外、語彙抽出で行った改善後の状態から手法に変更を加えていない。わずかだが、改善効果が現れていることが分かる。

表 18: 語彙の条件付き補完

問題文	無制限に補完	条件を適用
q18/15/1	1	1
q18/15/2	1	1
q18/15/3	6	5
q18/15/4	1.5	1.5
q18/15/5	2.75	2.75
q19/13/エ	1	1
q19/13/オ	1	1
q19/16/1	1	1
q19/16/2	1	1
q19/16/4	9	9
q19/16/5	2	2
q19/7/3	4	4
平均	2.60	2.52
F 値平均(3 位基準)	0.367	0.367

語彙補完と語彙の重み付け

語彙抽出の際に述べた、語彙ネットワークに登録があり、かつ民法全文からの抽出語彙が含まれない語彙について、これには idf 重みが存在しない。

本研究では、語彙ネットワークにおけるそういった語彙の位置付けを利用してこれを推定することとした。すなわち、重みの設定されていない語彙 t の近隣の語彙集合を $n_1, n_2, \dots \in N$ としたとき、 t の idf 重みの初期値を 0 として、

$$\text{idf}(t) = \frac{1}{|N|} \cdot \sum_{n \in N} \text{idf}(n) \quad (14)$$

によって idf 重みの変化するノードがなくなるまで更新を続ける。

以上による結果はこの通りである。比較的大きな改善効果が現れている。

表 19: *idf* 重みの補完

問題文	重み補完なし	重み補完あり
q18/15/1	1	1
q18/15/2	1	1
q18/15/3	5	3
q18/15/4	1.5	1.5
q18/15/5	2.75	2.5
q19/13/エ	1	1
q19/13/オ	1	1
q19/16/1	1	1
q19/16/2	1	1
q19/16/4	9	9
q19/16/5	2	2
q19/7/3	4	5
平均	2.52	2.42
F 値平均(3 位基準)	0.367	0.408

語彙の潜在的な想起関係と汎化関係

語彙ネットワークを用いた手法の最後の検討項目として、語彙の汎化関係、すなわち語彙ネットワークのhyper/hyperxリンクを用いた語彙の補完の意味について考察する。

上位・下位の関係が、自然言語的な側面からみれば”実体としての概念の分類”という観点における意味役割の階層に過ぎないということを既に述べた。視点によって意味役割の階層が揺らぐのは、この概念の汎化という関係の起点がそもそも、概念を眺める場合の視点ではなく、ある状況や実体に固有なフレームの集合から、経験則によって共通した意味を勝ち取るという所にあるためであると考える。それではどのみち現実世界の存在に固有な本質としての“実体”的概念階層は揺らがないのではないかというのは、こと自然言語の自由な世界であるテキストの上では通用しないことである。この言語特有の汎化関係の方向付けは、言語教育の分野で、語彙の意味の学習において文脈依存から脱文脈に至る過程を重要なファクターとする理論(コア理論:[24]など)と共鳴するところがある。

以上を受けて、本研究の語彙の補完方法に話を戻してみる。概念階層の補完を含む語彙の補いは、あくまで抽象化の可能性の先にあるフレームにアクセスする操作である。このとき、もしその過程にあった上位概念自体が対象として論じられていなければ、あくまで起点はそれ自体で固有かつ完結したフレームをもつ下位概念であるから、上位概念自体がその後も主題を表現する要素として残るのは不合理である。例えば、非常に具体的な話をすれば、法律の知識を相当にもつ者が”抵当権”を話題にする際に、これが担保する”被担保債権”的概念を参照するのに”抵当権”は”担保物権”的概念で、”担保物権”的フレームにはそれが担保する所の”被担保債権”という属性が予約されており…』というような迂遠な思考過程は避け、直接具体的な”抵当権”には”被担保債権”があるはずだ”という関係をえるはずだ”というのである。

これまでの議論より、語彙ネットワークのリンク解析によって獲得した上位概念・語彙の内、以下のいずれにも該当しないものを削除することとする。

- 1 この概念の下位概念がその文書の語彙集合に存在しない
- 2 この概念が比較対象の文書の語彙集合に存在する

これを実施したのが以下の表である。左側は idf の補完まで改善を行ったものであるが、そこから更に大きな改善効果が現れていることがわかる。

表 20: 語彙の関連付けに用いた上位概念の扱いの変更

問題文	全て保持	条件により削除
q18/15/1	1	1
q18/15/2	1	1
q18/15/3	3	3
q18/15/4	1.5	1.5
q18/15/5	2.5	1.875
q19/13/エ	1	1
q19/13/オ	1	2
q19/16/1	1	1
q19/16/2	1	1
q19/16/4	9	8
q19/16/5	2	2
q19/7/3	5	5
平均	2.42	2.36
F 値平均(3 位基準)	0.408	0.442

3.3 本研究における提案のまとめ

本章では、本研究の設定課題である、民法分野の司法試験問題の関連条文検索についてその詳細を定義し、情報検索のベクトル空間モデルの枠組みの中で tf-idf 重みを用いた手法により、この課題のベースラインを設定した。

これに対して本研究では、法律文書としての問題文、条文双方の特徴を分析し、語彙出現と語彙の重要性に関する仮説、それぞれの文書の主題とそれを表現する語彙集合に関する仮説について述べた後、その検証という形でベースライン手法の改善を提案・実施し、効果を確認することができた。

さらに語彙の一貫性が得られない問題文-関連条文対に対処するため、クエリー拡張の必要性を述べ、またこれが潜在的な語彙の想起関係によって為されるべきである旨の仮説を立てて、これに即して語彙ネットワークの構築を行った。語彙ネットワークのリンク解析／リンク探索によって、仮説に見合った文書語彙の補完手法を提案し、またこれ以前のそれまでの手法改善における仮説との整合性確保を通じて、設定課題への構造化知識源の利用効果を検証することができた。

第4章 評価

本章では、第3章で定義および実施した、本研究の設定課題である関連条文検索タスクに際して、同じく第3章で定義を行ったベースライン手法の改善の観点から、提案手法の評価を行う。

加えて、提案手法において導入した語彙ネットワークの設定課題への貢献を評価することで、設定課題における構造化知識の有効性を論ずる。

同時に、司法試験問題の回答に向けた取組み、および広く法律分野のテキスト処理に向けた取組みに対する適用可能性について述べ、本研究に対する評価とする。

4.1 関連条文検索タスクに対する手法の評価

本研究では、第3章でも述べた通り、ベースライン手法を明確に設定し、それに逐次改善を加える形での手法提案を行ってきた。従って提案の各段階において、設定課題の実施結果は既に大部分が示されてきたが、ここではこれらを総括し、より俯瞰的な観点、あるいはより個別具体的な観察をもとに考察を加えていくこととする。

4.1.1 問題文・条文の特徴・形式分析と語彙選択

ここでは主に、構造化知識を用いる前段階において提案された手法に関して、評価を行う。以下に、手法改善の各段階における検索課題の実施結果を示す。ただし表中の $\delta\text{-}idf$ とは、語彙出現のファクターを語彙頻度から語彙の出現(1)／非出現(0)に変更したことを示す。

表 21: 構造化知識以前の手法改善

問題文	ベースライン 手法	$\delta\text{-}idf$ の導入	$\delta\text{-}idf +$ 語彙の フィルタリング	語彙のフィル タリングのみ
q18/15/1	1	1	1	1
q18/15/2	7	6	2	1
q18/15/3	2	3	1	3
q18/15/4	2	1	1	2
q18/15/5	1	1	1.875	3.75
q19/13/エ	1	1	1	1
q19/13/オ	8	4	1	4
q19/16/1	1	1	1	1
q19/16/2	1	1	1	1
q19/16/4	8	5	8.5	21.5
q19/16/5	1	1	1	1
q19/7/3	44	44	44	44
平均	6.42	5.75	5.36	7.02
F 値平均(3 位基準)	0.358	0.358	0.442	0.333

総合的な指標に際しての評価

これら2つの手法改善について、検索課題のおおよその順位付けの向上という意味で効果的であることは既に第3章で確認されている。

ここまで触れる機会のなかった、3位基準のF値に関してここで述べておく。これは実際に、提案手法を司法試験回答等に利用するにあたって、有効な取得数を3位までとした場合に関連条文の回収がどの程度可能かという観点を量化する指標となっている。前掲の表では実際に、

δ -idf の導入後から語彙フィルタリングの適用に際して、正しい関連条文を取得できるケースが増えていることが判る。一方でベースラインから δ -idf の導入に際しては変化が見られないが、これは3位圏内外における変化が現れなかつことを反映しており、実際の順位変動をみれば語彙フィルタリングの導入時よりも著しくらいである。

ところで一般的に、情報検索の評価には適合率-再現率曲線やMAP(平均適合率の平均)[8](p.140-)というような評価指標が用いられるが、今回の設定課題のように正解文書(関連条文)がほぼ全てのケースでたった1つであるような場合、あまり有益な検証となり得ず直接順位を比較した方が直感的であるため、今回は以降の評価も含め、割愛する。

個別具体的な観察および評価

さて、表中の特筆すべき個別の順位変化に関して、より処理内容に踏み込んだ分析を行い、手法の評価を深めることとする。

まず δ -idf の導入に際して、q19/13/才およびq19/16/4の両ケースにおいて、比較的大きな順位変化が見られることが判る。q19/13/才について、スペースを節約のためここでは引用を避けるが(付録Aを参照されたい)、この関連条文である第398_3条は、むしろ語彙フィルタリングの仮説立ての際に特徴付けた、条文内に複数の要件-効果対が存在する条文であることがわかる。これは同ケースが語彙フィルタリング導入の際に、さらに著しく順位を改善させていくことからも検証できる。 δ -idf の導入は、語彙フィルタリングの有効なケースを改善させる場合もあるようである。これを見るに、実のところ δ -idf 自体には仮説通りの改善効果がないのではないかという危惧が過るが、表中最右列の語彙フィルタリングのみを施した実施結果をみれば、決してそうではないことが見て取れるだろう。

q19/16/4(以下のケース)を見ると、これはどちらかと言えば、語彙の想起関係の考慮が必要な問題文-関連条文の対であることがわかる。 δ -idf により”抵当権”的影響が緩和されることで語彙の一致率が上がり、順位が向上しているが、これは望ましい振舞いではない。

本文

抵当権の法律関係に関する次の1から5までの各記述のうち 誤っているものはどれか。

抵当権が設定された不動産について、地上権の設定を受けた者は、抵当権消滅請求をすることができない。

関連条文

(**抵当権消滅請求**)

第三百七十九条 抵当不動産の第三取得者は、三百八十三条の定めるところにより、抵当権消滅請求をすることができる。

図20: 語彙の想起関係が必要と思われるケース

このまま語彙フィルタリングにおけるq19/16/4の変化を見る。ここでは大きな順位低下が見られるが、実のところ問題文-条文間の類似度は上がってすらいる。すなわち、他のケースにおける類似度の上昇に打ち負かされているのがこの順位低下の原因であるが、これは単純に語彙フィルタリングの仮説が語彙の補いの必要なケースに役に立たないため、仕方のないことである。

語彙フィルタリング導入における順位改善として、**q18/15/2, q18/15/3, q19/13/オ**が挙げられる。この内q19/13/オは上記で、q18/15/2は第3章で、それぞれ複数規定を含む条文であることが確認されている。q18/15/3に関しては、これは引用を避けるが、条文(根抵当権の定義)における第3項の記述が、根抵当権のより特殊な適用要件に関して定めたものであり、この文の語彙が捨て置かれることで、順位が改善していると考えられる。ただし、このケースでは問題文と条文の実際の意味的な関連の仕方が迂遠であり、あまり必然的に検索が成功したとは言えないと判断できる。

総括

以上を見た限りで、 δ -idf および語彙フィルタリングの導入は非常に事前の仮説立てに即した働きを行っており、これらが解決をサポートしない問い合わせの検索順位が悪化ないしは無反応であることも含めて、関連条文の検索課題に対して有効であると評価することができる。

4.1.2 語彙ネットワークを用いた語彙拡張

ここでは、構造化知識、もとい語彙ネットワークの導入・利用後の提案手法に関して評価を行う。ここではまず語彙ネットワーク導入後の手法改善に関して検証した後、語彙ネットワーク自体の導入を評価する。以下は導入後の段階的な手法を比較したものである。

表 22: 語彙ネットワーク導入後の手法改善

問題文	語彙補完のみ	語彙抽出の最適化	$\leftarrow +$ 条件付語彙補完	$\leftarrow +$ idf重み補完	$\leftarrow +$ 上位概念の条件付削除
q18/15/1	1	1	1	1	1
q18/15/2	1	1	1	1	1
q18/15/3	6	6	5	3	3
q18/15/4	1.5	1.5	1.5	1.5	1.5
q18/15/5	2.5	2.75	2.75	2.5	1.875
q19/13/エ	2	1	1	1	1
q19/13/オ	1	1	1	1	1
q19/16/1	1	1	1	1	1
q19/16/2	1	1	1	1	1
q19/16/4	9	9	9	9	9
q19/16/5	2	2	2	2	2
q19/7/3	4	4	4	5	5
平均	2.67	2.60	2.52	2.42	2.36
F 値平均(3 位基準)	0.367	0.367	0.367	0.408	0.442

総合的な指標に際しての評価

平均順位の順次的な改善については、第3章で既に確認されていることである。

3位基準のF値について見ていこう。ここではidf重み補完、および上位概念の条件付き削除操作の導入時においてのみ、関連条文の回収について改善が見られるに至っている。語彙ネットワーク導入前の仮説との整合性をはかるために努めた割には、あまり報われない結果が示されている。

個別具体的な観察および評価

手法改善段階毎のケースの順位の変化を見ていく。ここでは効果の小さい手法改善が多いいため、idf重みの補完における**q18/15/3**、および上位概念の条件付き削除操作における**q18/15/5**を観察する。

まずq18/15/3について、この問題文-関連文対の補完後の語彙構成を、一致語彙についてはidf重み補完前の重み(語彙の"/"区切りの後の数値)と共に示してみる。太字は表記の一致が見られた語彙で、条文において太字の語彙ばかりが見られるのは、語彙のフィルタリングの操作によるものである。このケースは単純に、"根保証"や"抵当権"等、"抵当権"についての基礎的ではあるが、民法本文には登場しない語彙／概念が多く補完されていることがわかる。これは、idf重みの補完による改善効果としては適正なものであろう。

q18/15/3:

一般, 保護, 債権/2.0, 債権者, 優先, 利害関係, 弁済, 後順位, 抵当権設定者, 担保/2.3, 担保権, 有, 根抵当権/2.4, 極度額/3.2, 欄, 的, 目的, 範囲/2.8, 被担保債権, 解答, 記述, 限定/4.1, 限度/2.9, №, 抵当権, 根抵当/0.0, 抵当/0.0, 根保証/0.0

398_2:

債権/2.0, 抵当権/2.4, 担保/2.3, 根抵当権/2.4, 極度額/3.2, 範囲/2.8, 限定/4.1, 限度/2.9, 根保証/0.0, 根抵当/0.0, 抵当/0.0

図21: idf重み補完で改善効果の高いケースの語彙構成

次にq18/15/5について、この語彙構成とその成り立ちについて調査してみると、この関連条文(398_12および398_13)との比較において、特に前段階(idf重み補完)から、問題文自体にも関連条文にも変化がないことがわかる。即ちこのケースにおいては、細かい調査結果は省くが、前段階で不正に上位に順位付けされてケースについて、第3章で述べた不都合な上位語の補完が打ち消され、正しい関連条文の順位が上昇したということが観察できる。

総括

ここでは、語彙ネットワーク導入前の仮説との整合性を確保するための操作についての改善効果がわずかである一方、idf重みの補完および語彙の汎化関係の仮説に基づいた手法改善においては、おおよそ仮説立てに即した働きを見せ、有効であると結論づけることができた。

続いて、語彙ネットワーク自体の導入について評価を行う。以下に、ベース・ライン手法、語彙ネットワーク導入前の最も改善効果の高かったケース、語彙ネットワーク導入後の最も改善効果の高かった手法(条件付語彙補完+語彙の最適化+idf補完+上位概念の条件付削除)による実施結果を示す。

表 23: 語彙ネットワークの導入による改善効果

問題文	ベースライン 手法	δ -idf +語彙の フィルタリング	語彙ネットワーク を用いた手法
q18/15/1	1	1	1
q18/15/2	7	2	1
q18/15/3	2	1	3
q18/15/4	2	1	1.5
q18/15/5	1	1.875	1.875
q19/13/エ	1	1	1
q19/13/オ	8	1	1
q19/16/1	1	1	1
q19/16/2	1	1	1
q19/16/4	8	8.5	9
q19/16/5	1	1	2
q19/7/3	44	44	5
平均	6.42	5.36	2.36
q19/7/3 を除いた 平均	3.00	1.85	2.13
F 値平均(3 位基準)	0.358	0.442	0.442

総合的な指標に際しての評価

順位平均を見ると、順次改善していると言えなくはないが、q19/7/3 の順位向上を除くと、語彙ネットワークの導入に際しての効果が特に現れているとは言い難い。一方で、これまでの語彙ネットワークによる語彙補完を含む手法の実施結果を見るに、q19/7/3 の関連条文の順位を上げながら、他のケースの順位付けの整合性を保つことがかなり困難であることがわかるだろう。

また、関連条文の回収率を示す所の F 値について見ると、語彙ネットワークを用いた手法において、これを用いない手法の最も良いものと同水準には達していることがわかる。

個別具体的な観察および評価

ここでは、表記一致がなく語彙ネットワークの利用時にのみ上位に取得することのできる q19/7/3、および語彙ネットワークの導入によっても(良くも悪くも)順位の変化がない q19/16/4について個別に観察する。

まずは q19/7/3 について見る。取得語彙とその重みは以下の通りである。ここでは第3章において想定していた程語彙の潜在的な関連づけ(すなわち”賃料”と”果実”的な)が取得できていないことが分かるが、初めの文書内語彙の関連づけ時点でこれ以上柔軟な拡張を行うのは、他のケースに悪影響を及ぼすため難しい。

考えうるのは、より双方の文書コンテキストを重視すること、すなわち、語彙の潜在的な想起関係を初めから文書内ではなく文書間で探索し、依存性の上で全てを解決しようとすることがある。

ただこのケースによって、各文書について独立した形で主題の表現語彙を広げ、これを比較に用いるという手法に限界があることが明確に示されたということもできるだろう。

q19/7/3:

賃貸/3.4, 取得/2.4, 家具/2.6, 賃料/3.2, 売買契約/3.7, 売却/3.3, 売買/2.7, 所有権/2.7, 特約/3.2, 所有者/2.4, 債権/2.0

575:

売買/2.7, 売却/3.3

図 22: 一致語彙のないケースの語彙補完結果

つづいて q19/16/4 について見る。取得語彙は以下の通りであり、”地上権”等の語彙の一一致しない事等が問題のように思える。しかしながら、これについてはいくらか解説が必要である。すなわち、問題では「地上権の設定を受けた者は、抵当権消滅請求をすることができない」ことの正誤が問われているが、実際、民法の条文にそのような規定はない。なぜなら、「抵当権消滅請求権者は抵当不動産につき所有権を取得した第三者に限り、地上権または永小作権を取得した第三者は除外される」[1](p.172)からである。これはもはや語彙の連想関係の問題ですらない。

「規定されていないこと」を判断するためのメタ推論能力、およびこれを根拠付けるための背景的な法律知識が必要である。従って、”地上権”という語彙の出現に対処できないことは、今回の手法の枠組みでは妥当な結果であると言えよう。

q19/16/4:

抵当権/2.4, 法律/3.3, 不動産/2.3, 抵当不動産/2.8, 設定/2.7, 抵当権消滅請求/3.3, 地上権/3.0

379:

抵当不動産/2.8, 抵当権消滅請求/3.3

図 23: メタ推論と背景知識が必要なケース

総括

ここでは、語彙ネットワークを用いた手法によって、表記一致が全く得られない問題文-関連条文対にそれなりの効果を発揮しながら、適合率を下げがちな語彙の補いの操作を含まない手法の中で最も性能を発揮したものとほぼ同程度の関連条文回収率を得ることができた。

一方で、今回のどの手法においても改善効果が見られなかつたケースは、語彙表記の想起関係よりも深い知識や推論能力が必要なもので、これは妥当な結果であったと言える。

4.1.3 関連条文検索タスクに際しての評価の総括

司法試験問題文の関連条文検索タスクを実施するにあたり、本研究ではまず、問題文・条文が法律分野の文書特有の構造・特徴を持つことに注目し、語彙集合の出現を主題と見立てる単純なモデルの中で仮説立て・検証を行い、ベースライン手法からかなりの改善効果をもつ手法提案を行うことができた。

一方この手法改善により、語彙出現の有無のみの枠組みの限界を見出すことができ、構造化知識、特に語彙表記のレベルと領域知識の双方への対応力を持つものの必要性について提案を行うことができた。

これを受け、語彙の潜在的な想起関係を表現する民法分野の語彙ネットワークを構築し、各文書の出現語彙の表現する主題を可能な限り大きな語彙集号へと広げる試みの提案を行った。この枠組において、今回劇的な改善効果をもたらす提案を行うには至らなかったが、より文書間の相互依存性のもとで語彙ネットワークの想起関係を利用するとの必要性に行き当たることができた。加えて、語彙の集合比較の枠組みでは解決し得ない問い合わせの形式を明らかにすることことができた。

4.2 語彙ネットワークの評価

以下より本研究において構築した語彙ネットワークについて、このリンク属性や利用可能性における評価を与える。

4.2.1 語彙ネットワークの構築プロセスの評価

本論に移る前に、本研究における語彙ネットワークの構築プロセスについて振り返り、評価を行う。

本研究で構築した語彙ネットワークの規模は、次のようなものである。

表 24: 語彙ネットワークの規模

ノード／語彙数	リンク／関係数	リンク属性の種類
425	753	9

構築プロセスの技術的な側面に関しては第3章で述べた通りだが、この構築コストとして、これを作り上げるまでの約120時間、および、この予備知識として、民法の学習時間約60時間を挙げることができる。

これは非常に多大な労力であると言えるが、第2章でも述べた通り、テキスト処理という観点から構築された法律分野の語彙知識ベースやオントロジーは、あまり整備されていないのが現状である。本研究では、この種類の語彙ネットワークがテキスト処理に対してどのように貢献できるかという検証が少なからず行えたと言え、法律分野の今後の構造化知識の利用可能性について示唆を行うことができれば、コストに対する収穫があったと言ってよいだろう。

一方で、概念の整然とした関連ではなく、強く語彙表記、テキストへの指向をあらわしたこの語彙ネットワークは、概念間の関係学習([25]等)と相性がよいと考えられ、今回この可能性を評価するには至らなかつたが、そのシードとしても用いることが期待される。

語彙ネットワークの構築に膨大な時間を費やしたのは、これが一般的なシソーラスにおける上位・下位関係以外の、法律知識を含む様々な関係を保持しているからである。こういったリンク属性の有用性について、以下より述べていく。

4.2.2 関連条文検索タスクへの貢献

4.1では、語彙ネットワークの全体を用いて文書の語彙集号の拡張を行い、それにより、条文検索タスクに貢献できる部分があることを明らかにした。一方、本研究における語彙ネットワークには、先に述べた通り9種の関係が含まれている。これらはそれぞれ、本研究の設定タスクにおいて意味のあるものだったであろうか？

以下は、先に述べた語彙ネットワークを用いる手法のうち最も効果を発揮したものについて、これをいくつかのリンク属性を省いた状態で実施して比較したものである。ここで、フレーム属性とは attr_slot, sbj, obj の3種のことである。

表 25: リンクの削除されたグラフと関連条文検索

問題文	ベース ライン 手法	完全な グラフ	上位・ 下位 のみ	フレーム 属性なし	前提／因果 なし	場所なし	auth なし
q18/15/1	1	1	1	1	1	1	1
q18/15/2	7	1	1	1	1	1	1
q18/15/3	2	3	1	2	6	3	3
q18/15/4	2	1.5	1	1	1.5	1.5	1.5
q18/15/5	1	1.875	2.75	2.75	1.875	1.875	1.875
q19/13/エ	1	1	1	1	1	1	1
q19/13/オ	8	1	2	2	2	1	2
q19/16/1	1	1	1	1	1	1	1
q19/16/2	1	1	1	1	1	1	1
q19/16/4	8	9	4	4	9	9	9
q19/16/5	1	2	1	2	2	2	2
q19/7/3	44	5	44.5	44.5	5	5	5
平均	6.42	2.36	5.10	5.27	2.70	2.36	2.45
F 値平均 (3 位基準)	0.358	0.442	0.408	0.408	0.400	0.442	0.442
関係数 の減少	-	753	281	213	29	7	32

これを見ると、あまりに関係定義数の少ないリンク属性は仕方がないものの、どれも少なからず語彙の補完に関わっていることがわかるだろう。特に上位・下位リンクのみを残して実施した順位結果に注目すると、かなりの性能低下が見られることに留意する。本研究においてこの語彙ネットワークでなく、ごく単純な上位・下位関係のみを持って構築したシソーラス状のものであつた場合、ここまで手の改善効果は齎されなかつたということがわかる。

上位・下位リンク以外の属性は、確かに語彙ネットワーク構築の際に気の遠くなるような(主に整合性を保つための)試行錯誤や関係の洗い出しが必要な一方で、同時に本語彙ネットワークの貢献領域が、単に語彙のごく単純な上位・下位の言い換えの解消のみではなく、より広い、すなわち手法提案の際に目指した、文書の主題を可能な限り大きな語彙集号によってあらわすために、"語彙間の潜在的な関連付け"を探索するという本来の目的に触れていると言つてよいのだといふことが分かる。

4.2.3 法律人工知能分野における利用可能性

本研究は、司法試験問題の回答を目指すプロジェクトのPROLEG言語のプロジェクトの一端として開始したものであった。第2章では自然言語からPROLEG言語へと橋渡しをする際の困難に関して述べたが、これは述語項構造解析等の深い言語理解が、本質的に世界知識の参照なくして成し得ない領域であり[26]、かつPROLEGの対象では更に、法律ドメインの知識の参照が必要となることが大きな理由のひとつである。

本研究で提案を行った語彙ネットワーク、および関連条文検索の枠組みでは、この課題への対処に貢献できる可能性がある。すなわち、試験問題文に対して正しい関連条文を検索することができた場合、既にここには、語彙ネットワークの潜在的な想起関係が生じしたものとしての概念間の関係がアノテーションされた状態となっているのである。これを従来の解析器の出力等と組み合わせることで、少なくとも"橋渡し"という枠組みにおける取組みに、より進んだステップを提供できる可能性がある。

以上で述べた例と同様、法律人工知能の分野には常に、法律の知識・学問体系を論理学やそれに準ずる領域で充分に論じたいという要請があるといえる。一方でPROLEGの例同様に、こと自然言語と触れる部位をもつ取組みを行おうとする場合、ただちに法律ドメインの知識と自然言語的な語彙の多様性の境界部分をいかに上手く立ち回るかという課題に直面することとなる。本研究で提案を行った、語彙ネットワークとその構築指針は、こういった課題に最も積極的に取り組むものとして発展させうるものである。

4.2.4 語彙ネットワーク評価の総括

本研究において提案・構築を行った語彙ネットワークは、法律分野において自然言語の領域とドメイン知識の双方をまたぐ構造化知識として、設定課題である関連条文検索や法律人工知能の分野で利用価値のあるものである。

また、この構築コストは無視できないほど大きなものだが、このような語彙ネットワークにとって重要なのは、通常のシソーラスには見ることのできない上位・下位関係以外のドメイン知識を含む関係であるということを確認するに至った。

4.3 本研究の他分野への応用可能性

ここでは、本研究で得られた知見、すなわち本研究の主目的である構造化知識の利用法に限らず、関連条文検索タスクの分析と実施を通しての成果の応用可能性について、早足で論じてみる。

まずは、法律分野のより広い領域への応用可能性について、自然言語によって自由記述された事例のテキスト表現から、これに関連する法令の条文や判例を取得するというものがある。これは例えば本研究に近いものとして、非専門家による法的知識へのアクセスによって素人訴訟を行おうとする者や初学者を支援する技術が考えられる。こういった課題に関しては、本研究において扱った試験問題文と条文との関係性よりも、概念粒度、抽象度に隔たりのあるもの同士を関連付ける必要がある。本研究において得た、語彙の潜在的な連想関係の探索を、直接文書比較に依存した形で実施すべきかもしれないという展望について、これがどの程度迂遠な関係に適用しうるかという検証が必要である。また、より一步領域を広げ、これも法律規定との関わりを持つが、企業活動のコンプライアンスの支援に応用するという方向性もあり得るだろう。

留意すべきは、本研究では司法試験問題文が、その中に主題をひとつだけ持っているという大切な仮定があったということである。提案を行った語彙のフィルタリングは、もし問題文にあたるテキストが複数の主題から成っており、かつそれらが常に同じ文書内に現れないというのであれば、ただちに本来の意味を果たせなくなる。そういう課題に応用をしようという場合、クエリー側の文章の主題を予め分割あるいは抽出するための技術を開発する必要がある。こういったことを特許検索の分野で精緻に行っているものとして、[27]等が挙げられる。本研究では、こうした分析の一端を垣間見つつ語彙ネットワークの活用へと注力していったが、文書検索におけるより工夫された仮説を様々に検証することで、語彙ネットワークの利用方法に関する仮説を改良できる可能性もある。あるいは本研究のように、検索される側の文書を予め全て入手できるのであれば、これら互いの類似度や重なりを分析することで、有効な主題の単位を発見するということも考えられるだろう。このとき、本研究で提案を行った、非常に単純な主題の包含検査のための語彙のフィルタリングが活躍する機会も、また得られる可能性がある。

4.4 総括

本章では、語彙ネットワークの導入前後で、設定課題への貢献が仮説立ての範囲内で整合的に効果を発揮していることを確認し、設定課題に対する構造化知識の有効性の分かれ目を明らかにすることことができた。また、語彙ネットワーク導入後にも対処が困難なケースを分析することで、本研究で提案を行った語彙ネットワークの枠組みあるいは利用法の限界について知ることができた。

一方で語彙ネットワークの語彙拡張への貢献について、通常のシソーラス等がサポートする上位・下位リンク以外のリンク属性の予想以上の寄与度を垣間見、自然言語的な側面とドメイン知識の両輪で課題に対処する本研究の提案方針の有効性を確認することができた。

また、これらの知見を受けて、本研究のおおもとのプロジェクトや、法律ドメインその周辺領域のテキスト処理に対する語彙ネットワークや本研究の手法の利用可能性を論じた。

第5章 結論

5.1 論文の総括

本研究では、民法分野の司法試験問題文からその関連条文を検索するというタスクを通じて、法律分野の極めて短い文書間の類似性をどのように判定するかということについて論じ、またその計算手法の提案を行った。またその課題の上で、構造化知識が必要となるのはいつで、それがどのような特徴をもつべきで、かついかに利用すればこの力を引き出せるかということについて論じ、更に提案を行った。

ここでは、試験問題文と法令の条文の非冗長性の検証から始まり、また語彙の集合として観測されるところの、文書がその内容を依って立つという主題というものの考察を行った。この主題の単位をここでは法的な規定であるとした上で、試験問題文と条文とではそれらの現れ方に相違があるという仮説を立て、これが設定課題に従って評価するに、有効な分析であったと結論づけることができた。

一方で、極めて単純な手法を洗練させていく過程で構造化知識の必要性に行き当たり、即ち表記の一貫しない語彙同士の関連付けが、少なからず法律ドメインの知識を参照しながらなされるべき旨の発見があった。これに対応するために、本研究では民法分野のあらゆる語彙について、これらが互いに連想され得る可能性を網羅した語彙ネットワークを構築した。この語彙ネットワークを用いて、各文書内に現れる語彙を整合的かつ密に結びつけることで、文書の主題を可能な限り豊かな語彙によって表現することを提案し、これによって表記の一貫性を解消とした。

こうして作成した新しい文書の主題表現を用いて関連条文検索の枠組みで評価するに、語彙の表記一致を得られないケースに対しての有効性があるとした上で、この語彙ネットワークを、対象の2文書間比較という限定的なコンテキストにおける語彙の想起関係の探索に用いることで、より対応できるケースの幅が広がる可能性があると結論づけた。

一方で構築された語彙ネットワークは、自然言語的側面とドメイン知識の境界的な性質にもっとも価値があり、本研究のタスクにおいても実際有効に働いていることを確認しつつ、この性質は法律分野のテキスト処理を含む人工知能研究にとって極めて有用なものとなる可能性があると示した。

5.2 今後の課題および展望

本研究において残された課題は、ひとつに現時点の語彙ネットワークを、2文書間比較の際の相互依存的な語彙の関連付けに利用する可能性について検討することである。本研究におけるリンク解析等の成果を再利用して検討できる部分があると考えられる一方で、ある文書における自分の語彙と比較相手の語彙をどのようにそれぞれ扱いながら潜在的関連性を探索していくか考える必要があるだろう。

加えて、本研究の成果は、試験問題文の少なくとも一年分等より広い枠組みの中で評価されるべきだが、現時点における語彙ネットワークの民法全体のカバレッジは、10%にも満たないと考えられる。これを受けて、この語彙ネットワークの構築プロセスをなるべく標準化して構築作業者の一人あたりの負担を減らすこと、または、この語彙ネットワークに定義されている関係をシードとして、関係学習の手法がどの程度有効に働くか検証することが必要である。

謝辞

本研究および本論文は、多くの方々のご指導およびご助力をもって形にすることことができたものです。

研究の全過程を通じて、恵まれた機会と環境の下で活動が続けられるよう暖かく見守って下さり、また常に俯瞰的な視点から適切なご指導を頂いた山口高平先生に心より感謝致します。

また、本研究の技術的・実質的な面に関して、多大な時間を割き一貫してご指導・ご助言を賜り、遅々とした研究進行にも辛抱強くご教示を頂いた、国立情報学研究所の佐藤健先生、市瀬龍太郎先生に深く感謝申し上げます。また佐藤健先生には、国立情報学研究所における活動に不自由がないよう、様々に便宜を計って頂きました。

学士から現在までの研究を通じて、また研究内容に関わらず、研究生活のあらゆる事項について、いつでも様々なご相談に快く乗ってくださった玉川獎氏に、深く感謝の意を表します。

完全なる法学の素人であった筆者に、日本民法の基礎に関して60時間を優に超える講義にてご指導頂き、また国立情報学研究所における研究活動において、様々なご助言と日々の仕事への激励を賜った大森健太郎氏、姜明訓氏、高井亮氏、土屋一貴氏への感謝は筆舌に尽くしがたいものです。

最後に、研究生活全般にわたって共に励まし合い、時に重要なご助言を頂いた山口研究室の同期・後輩の皆様に、心より御礼申し上げます。

平成26年2月4日

参考文献

- [1]: David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty, "Building Watson : An Overview of the DeepQA Project", AI Magazine 31(3), pp.59-79, 2010
- [2]: 新井 紀子, 松崎 拓也, "口ボットは東大に入れるか?", 人工知能学会誌 27(5), pp.463-469, 2012
- [3]: 川添愛、宮尾祐介、松崎拓也、横野光、新井紀子, "史実としてありえない」という判断を可能にする世界史オントロジー", 2013年度人工知能学会全国大会 (JSAI2013) 論文集, 2013
- [4]: 吉野 一(編), "法律人工知能–法的知識の解明と法的推論の実現", 創成社, 2000
- [5]: 大嶽能久, 新田克己, 前田茂, 小野昌之, 大崎宏, 坂根清和, "法的推論システム HELIC-II", 情報処理学会論文誌 35(6), pp.986-996, 1994
- [6]: 佐藤健, et al., "PROLEG: 論理プログラミングをベースとした民事訴訟における要件事実論の実装", 知識ベースシステム研究会 92,
- [7]: Ryu Iida, Massimo Poesio, "A Cross-Lingual ILP Solution to Zero Anaphora Resolution", The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), 2011
- [8]: Christopher D.Manning, Prabhakar Raghavan, Hinrich Schutze 著, 岩野 和生, 黒川 利明, 濱田 誠司, 村上 明子 訳, "情報検索の基礎", 共立出版, 2012
- [9]: Claudio Carpineto, Giovanni Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Computing Surveys 44(1), Article 1, pp.1-50, 2012
- [10]: George A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM 38(11), pp.39-41, 1995
- [11]: Roberto Navigli, Velardi Paola, "An analysis of ontology-based query expansion strategies", Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, 2003
- [12]: 法学セミナー編集部, "司法試験の問題と解説 2012 法学セミナー増刊 (別冊法学セミナー no. 216)", 日本評論社, 2012
- [13]: 道垣内 弘人, "プレッピ法学を学ぶ前に (プレッピシリーズ)", 弘文堂, 2010
- [14]: 工藤拓, 山本薫, 松本裕治, "Conditional Random Fields を用いた日本語形態素解析", 情報処理学会研究報告. 自然言語処理研究会報告 2004(47), 2004
- [15]: 川井 健, "民法入門 第7版", 有斐閣, 2012
- [16]: 松尾弘, "法整備支援における民法典整備の意義と課題", 慶應法学 4, pp.31-62, 2006
- [17]: Le-Minh Nguyen, Ngo Xuan Bach, Akira Shimazu, "Supervised and semi-supervised sequence learning for recognition of requisite part and effectuation part in law sentences", Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing. Association for Computational Linguistics, 2011
- [18]: 樽松理樹, 山口高平, "法律知識の体系的定義としての法律オントロジー(<特集>開発されたオントロジー)", 人工知能学会誌 19(2), pp.144-150, 2004
- [19]: 溝口理一郎 著, 人工知能学会 編, "オントロジーエンジニアリング (知の科学)", オーム社, 2005

- [20]: 山口高平, 横松理樹, "法律オントロジー (<小特集>法律と人工知能)", 人工知能学会誌 13(2), pp.189-196, 1998
- [21]: 黒田航, 井佐原均, "意味役割名と意味型名の区別による新しい概念分類の可能性 : 意味役割の一般理論はシソーラスを救う?(推論・意味)", 情報処理学会研究報告. 自然言語処理研究会報告 2005(73), 2005
- [22]: Charles J. Fillmore, "Frame semantics", In "Linguistics in the morning calm", ed. by The Linguistic Society of Korea, Seoul: Hanshin Publishing Co., 1982
- [23]: 関内律恵子, 小森聰, 横松理樹, 青木千鶴, 山口高平, "DODDLE: 計算機可読型辞書を利用した領域オントロジー構築支援環境 (2) : 概念定義構築支援", 情報処理学会全国大会 55(2), 1997
- [24]: 田中茂範, "基本語の意味のとらえ方-基本動詞におけるコア理論の有効性-", 日本語教育 121, pp.3-13, 2004
- [25]: Peter D. Turney, "A uniform approach to analogies, synonyms, antonyms, and associations", Proceedings of the 22nd International Conference on Computational Linguistics, 2008
- [26]: 乾健太郎, "事態オントロジー: 言語に基づく推論のためのコトに関する基本知識", 言語処理学会第 13 回年次大会ワークショップ 「言語的オントロジーの構築・連携・利用」 論文集, 2007
- [27]: 高木徹, 藤井敦, 石川徹也, "検索質問の主題分析に基づく類似文書検索と特許検索への応用 (情報検索)", 情報処理学会論文誌 46(4), pp.1074-1081, 2005

付録A 手法の分析・評価に用いた問題文とその関連条文

対象の問題文と関連条文

以下に、本研究の設定タスク、すなわち司法試験の短答式試験問題文に対する関連条文検索の分析および評価に用いたケースを、その正しい関連条文とともに掲載する。ただしここでは簡便のため、問題の配点、解答欄の指示を省き、かつ問題番号の表示等を可読性に関する配慮のもと変形している。これらの原文は[http://www.moj.go.jp/jinji/shihoushiken/jinji08_00025.html]にて完全な状態で入手することが出来る。

平成18年度 第15問

大問の指示文

根抵当権に関する次の1から5までの記述のうち、正しいものはどれか。

選択肢1(略号 q18/15/1) :

本文

第一順位の根抵当権者は、後順位の担保権者が目的不動産について申し立てた競売手続が開始しても、競売時期の選択について後順位の担保権者より優先するから、元本を確定させず、競売手続を止めることができる。

関連条文

(根抵当権の元本の確定事由)

第三百九十八条の二十 次に掲げる場合には、根抵当権の担保すべき元本は、確定する。

- 一 根抵当権者が抵当不動産について競売若しくは担保不動産収益執行又は第三百七十二条において準用する第三百四条の規定による差押えを申し立てたとき。ただし、競売手続若しくは担保不動産収益執行手続の開始又は差押えがあったときに限る。
- 二 根抵当権者が抵当不動産に対して滞納処分による差押えをしたとき。
- 三 根抵当権者が抵当不動産に対する競売手続の開始又は滞納処分による差押えがあったことを知った時から二週間を経過したとき。
- 四 債務者又は根抵当権設定者が破産手続開始の決定を受けたとき。

2 前項第三号の競売手続の開始若しくは差押え又は同項第四号の破産手続開始の決定の効力が消滅したときは、担保すべき元本は、確定しなかったものとみなす。ただし、元本が確定したものとしてその根抵当権又はこれを目的とする権利を取得した者があるときは、この限りでない。

選択肢 2(略号 q18/15/2) :

本文

根抵当権も元本が確定すれば普通抵当権と同じに扱われるから、被担保債権の利息や損害金のうち根抵当権によって担保される部分は、最後の2年分に限定される。

関連条文

(根抵当権の被担保債権の範囲)

第三百九十八条の三 根抵当権者は、確定した元本並びに利息その他の定期金及び債務の不履行によって生じた損害の賠償の全部について、極度額を限度として、その根抵当権を行使することができる。

2 債務者との取引によらないで取得する手形上又は小切手上の請求権を根抵当権の担保すべき債権とした場合において、次に掲げる事由があったときは、その前に取得したものについてのみ、その根抵当権を行使することができる。ただし、その後に取得したものであっても、その事由を知らないで取得したものについては、これを行使することを妨げない。

- 一 債務者の支払の停止
- 二 債務者についての破産手続開始、再生手続開始、更生手続開始又は特別清算開始の申立て
- 三 抵当不動産に対する競売の申立て又は滞納処分による差押え

選択肢 3(略号 q18/15/3) :

本文

根抵当権が優先的に弁済を受ける限度は極度額によって定まっており、後順位担保権者や一般債権者は、どのような債権が担保されるのかについては利害関係を有しないから、被担保債権の範囲の限定は、もっぱら抵当権設定者の保護を目的としている。

関連条文

(根抵当権)

第三百九十八条の二 抵当権は、設定行為で定めるところにより、一定の範囲に属する不特定の債権を極度額の限度において担保するためにも設定することができる。

2 前項の規定による抵当権（以下「根抵当権」という。）の担保すべき不特定の債権の範囲は、債務者との特定の継続的取引契約によって生ずるものその他債務者との一定の種類の取引によって生ずるものに限定して、定めなければならない。

3 特定の原因に基づいて債務者との間に継続して生ずる債権又は手形上若しくは小切手上の請求権は、前項の規定にかかわらず、根抵当権の担保すべき債権とすることはできる。

選択肢4(略号 q18/15/4) :

本文

根抵当権の元本の確定前であっても、弁済期が到来した被担保債権をすべて弁済した第三者は、債務者に対する求償権を確実にするため、根抵当権者に代位して、根抵当権を行使することができる。

関連条文

(根抵当権の被担保債権の譲渡等)

第三百九十八条の七 元本の確定前に根抵当権者から債権を取得した者は、その債権について根抵当権を行使することができない。元本の確定前に債務者のために又は債務者に代わって弁済をした者も、同様とする。

2 元本の確定前に債務の引受けがあったときは、根抵当権者は、引受人の債務について、その根抵当権を行使することができない。

3 元本の確定前に債権者又は債務者の交替による更改があったときは、その当事者は、第五百八条の規定にかかわらず、根抵当権を更改後の債務に移すことができない。

選択肢5(略号 q18/15/5) :

本文

元本確定前の根抵当権は、被担保債権とは切り離された極度額の価値支配権であるから、その全部又は一部を譲渡することができるが、債務者や被担保債権も変わり得るから、根抵当権設定者の承諾を得なければならない。

関連条文1

(根抵当権の譲渡)

第三百九十八条の十二 元本の確定前においては、根抵当権者は、根抵当権設定者の承諾を得て、その根抵当権を譲り渡すことができる。

2 根抵当権者は、その根抵当権を二個の根抵当権に分割して、その一方を前項の規定により譲り渡すことができる。この場合において、その根抵当権を目的とする権利は、譲り渡した根抵当権について消滅する。

3 前項の規定による譲渡をするには、その根抵当権を目的とする権利を有する者の承諾を得なければならない。

関連条文2

(根抵当権の一部譲渡)

第三百九十八条の十三 元本の確定前においては、根抵当権者は、根抵当権設定者の承諾を得て、その根抵当権の一部譲渡（譲渡人が譲受人と根抵当権を共有するため、これを分割しないで譲り渡すことをいう。以下この節において同じ。）をすることができる。

平成19年度 第7問

大問の指示文

次の1から5までの各記述のうち、正しいものを2個選びなさい。

選択肢3(略号 q19/7/3) :

本文

家具の所有者AがBに賃貸中の当該家具をCに売却した場合、特約がなければ、Cは、直ちにその所有権を取得するから、Bに対する賃料債権も、Cが売買契約時に取得することになる。

関連条文

(果実の帰属及び代金の利息の支払)

第五百七十五条 まだ引き渡されていない売買の目的物が果実を生じたときは、その果実は、売主に帰属する。

2 買主は、引渡しの日から、代金の利息を支払う義務を負う。ただし、代金の支払について期限があるときは、その期限が到来するまでは、利息を支払うこと不要しない。

平成19年度第13問

大問の指示文

担保物権の効力に関する次のアからオまでの各記述のうち、正しいものを組み合わせたものは、後記1から5までのうちどれか。

選択肢エ(略号 q19/13/エ) :

本文

根抵当権でない抵当権は、担保する債権の元本のほか、利息その他の定期金のうち満期となった最後の2年分に限り、それらを担保する。

関連条文

(抵当権の被担保債権の範囲)

第三百七十五条 抵当権者は、利息その他の定期金を請求する権利を有するときは、その満期となった最後の二年分についてのみ、その抵当権を行使することができる。ただし、それ以前の定期金についても、満期後に特別の登記をしたときは、その登記の時からその抵当権を行使することを妨げない。

2 前項の規定は、抵当権者が債務の不履行によって生じた損害の賠償を請求する権利を有する場合におけるその最後の二年分についても適用する。ただし、利息その他の定期金と通算して二年分を超えることができない。

選択肢オ(略号 q19/13/オ) :

本文

元本の確定した根抵当権は、確定した元本のほか、利息その他の定期金のうち満期となった最後の2年分について、極度額を限度として担保する。

関連条文

(根抵当権の被担保債権の範囲)

三百九十八条の三 根抵当権者は、確定した元本並びに利息その他の定期金及び債務の不履行によって生じた損害の賠償の全部について、極度額を限度として、その根抵当権を行使することができる。

2 債務者との取引によらないで取得する手形上又は小切手上の請求権を根抵当権の担保すべき債権とした場合において、次に掲げる事由があったときは、その前に取得したものについてのみ、その根抵当権を行使することができる。ただし、その後に取得したものであっても、その事由を知らないで取得したものについては、これを行ふことを妨げない。

- 一 債務者の支払の停止
- 二 債務者についての破産手続開始、再生手続開始、更生手続開始又は特別清算開始の申立て
- 三 抵当不動産に対する競売の申立て又は滞納処分による差押え

平成19年度 第16問

大問の指示文

抵当権の法律関係に関する次の1から5までの各記述のうち 誤っているものはどれか。

選択肢1(略号 q19/16/1) :

本文

抵当権が設定された建物を、抵当権者に対抗することができない賃貸借に基づいて使用する者は、競売手続開始前から使用していれば、建物の買受人が買い受けた時から6か月を経過するまでは、その建物の買受人への引渡しを猶予される。

関連条文

(抵当建物使用者の引渡しの猶予)

第三百九十五条 抵当権者に対抗することができない賃貸借により抵当権の目的である建物の使用又は収益をする者であって次に掲げるもの（次項において「抵当建物使用者」という。）は、その建物の競売における買受人の買受けの時から六箇月を経過するまでは、その建物を買受人に引き渡すことを要しない。

- 一 競売手続の開始前から使用又は収益をする者
 - 二 強制管理又は担保不動産収益執行の管理人が競売手続の開始後にした賃貸借により使用又は収益をする者
- 2 前項の規定は、買受人の買受けの時より後に同項の建物の使用をしたことの対価について、買受人が抵当建物使用者に対し相当の期間を定めてその一箇月分以上の支払の催告をし、その相当の期間内に履行がない場合には、適用しない。

選択肢2(略号 q19/16/2) :

本文

登記をした賃貸借は、その登記前に登記をした抵当権を有するすべての者が同意をすれば、その同意をした抵当権者に対抗することができる。

関連条文

(抵当権者の同意の登記がある場合の賃貸借の対抗力)

第三百八十七条 登記をした賃貸借は、その登記前に登記をした抵当権を有するすべての者が同意をし、かつ、その同意の登記があるときは、その同意をした抵当権者に対抗することができる。

2 抵当権者が前項の同意をするには、その抵当権を目的とする権利を有する者その他抵当権者の同意によって不利益を受けるべき者の承諾を得なければならない。

選択肢4(略号 q19/16/4) :

本文

抵当権が設定された不動産について、地上権の設定を受けた者は、抵当権消滅請求をすることができない。

関連条文

(抵当権消滅請求)

第三百七十九条 抵当不動産の第三取得者は、第三百八十三条の定めるところにより、抵当権消滅請求をすることができる。

選択肢5(略号 q19/16/5) :

本文

被担保債権の債務不履行後に、抵当不動産の所有者が、その後に生じた果実を收受しても、不当利得にはならない。

関連条文

第三百七十二条 抵当権は、その担保する債権について不履行があったときは、その後に生じた抵当不動産の果実に及ぶ。

検索課題の全文書集合としての条文リスト

以下では、本研究においてベースライン手法、および各提案手法において仮定した全文書集合としての条文リストを示す。ここでは条文番号のアラビア数字のみを示すので、本文が必要な場合は[\[http://law.e-gov.go.jp/htmldata/M29/M29HO089.html\]](http://law.e-gov.go.jp/htmldata/M29/M29HO089.html)を参照されたい。

86, 167, 175, 180,
239, 240, 241, 249, 325, 339, 341, 343, 344, 358, 361,
371, 372, 373, 375, 379, 380, 382, 387, 388, 390,
391, 393, 395, 397, 398, 398_2, 398_3, 398_4, 398_5,
398_6, 398_7, 398_11, 398_12, 398_13, 398_15,
398_16, 398_17, 398_19, 398_20, 398_21, 399,
404, 405, 475, 479, 483, 489, 500, 520, 558, 575,
601, 602, 604, 616, 640, 876, 876_6

図 24: 全文書集合としての条文リスト

付録 B 語彙のフィルタリングによるコサイン類似度修正効果

第3章で述べた、語彙のフィルタリング効果による検索効果の向上に関して、ここで証明を行う。

文書 D について、この構成語彙を $T(D)$ と記すこととする。またこれ以降登場する語彙集合は、特に断らない限りすべてそれぞれ独立であるとする。加えて集合の要素数を $|T(D)|=d$ のように小文字で記すこととする。

ある問題 Q について、 $T(Q)=M_1 \cup M_2 \cup R$ であり、また条文 A_1 , A_2 について $T(A_1)=M_1 \cup M_2 \cup X_1$, $T(A_2)=M_1 \cup X_2$ であるとする。また X_1 と X_2 それぞれの語彙数について、 $x_1 > x_2$ であるとしておく。

すなわち、条文 A_1 は A_2 よりも多くの問題文との共通部分を含んでいる一方、それ以外の関係のない語彙についても A_2 より多い数が含まれている。これは以下のような状態である。

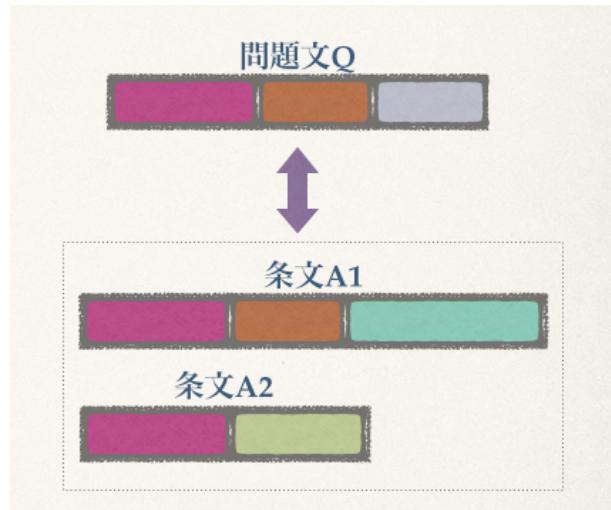


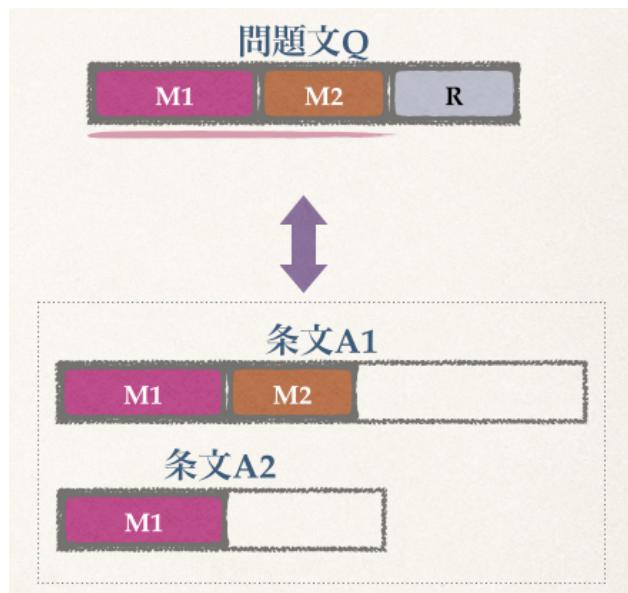
図 25: 語彙のフィルタリング：
問題設定

このとき、簡単のため $\delta_{t,d}$ のみで語彙ベクトルをつくり、コサイン類似度を計算すると以下の通りとなり、この大小は判定することができない。

$$\text{similarity}(\text{T}(Q), \text{T}(A_1)) = \frac{1}{\sqrt{m_1+m_2+r}} \times \frac{m_1+m_2}{\sqrt{m_1+m_2+x_1}} \quad (15)$$

$$\text{similarity}(\text{T}(Q), \text{T}(A_2)) = \frac{1}{\sqrt{m_1+m_2+r}} \times \frac{m_1}{\sqrt{m_1+m_2+x_2}} \quad (16)$$

ここで、事前に問題文語彙による条文語彙のフィルタリングを行った場合、すなわち以下のような状態となる。



26: 語彙のフィルタリング：
操作後のイメージ

このとき、先の式は以下のように変化し、
 $\text{similarity}(\text{T}(Q), \text{T}(A_1)) > \text{similarity}(\text{T}(Q), \text{T}(A_2))$ を得ることができる。

$$\text{similarity}(\text{T}(Q), \text{T}(A_1)) = \sqrt{\frac{m_1+m_2}{m_1+m_2+r}} = \sqrt{\frac{1}{1+\frac{r}{m_1+m_2}}} \quad (17)$$

$$\text{similarity}(\text{T}(Q), \text{T}(A_2)) = \sqrt{\frac{m_1}{m_1+r}} = \sqrt{\frac{1}{1+\frac{r}{m_1}}} \quad (18)$$