専門領域の語彙ネットワークを用いた法律文書の主題類似度判定

A Similarity Measure for the Topic of Legal Texts using a Domain–Specific Lexical Network

松下裕 [1] *　佐藤健 [2]　市瀬龍太郎 [2]　山口高平 [1]
Yu Matsushita[1]　Ken Sato[2]　Ryutaro Ichise[2]　Takahira Yamaguchi[1]

[1] 慶応義塾大学　[2] 国立情報学研究所
[1] Keio University　[2] National Institute of Informatics

Abstract: 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.

1　はじめに

1.1　自然言語と法情報学

計算機によるテキストの深い意味理解の実現には，辞書や構文知識等の言語モデルの他に，本質的に世界知識の参照が必要不可欠であると言われている [1]．さらに述語項構造や照応関係についても，語彙知識に基づく暗黙の関係を知らなければ説明できない場合も多い．つまり，構文論から深い言語理解へと踏み出す場合，直ちに世界知識の問題に直面すると言ってもよいだろう．本研究で扱う法律文書は特に，専門知識に基づく語彙／概念同士の暗黙の関連を頼りに記述されることが多い．従って法律は，知識をテキスト処理に利用することを考える上で興味深い分野である．一方少なくとも日本においては，90 年代の “法律エキスパートシステム” の試み [2] に代表されるように，これまで法的知識の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的知識ベースの構築等といった，かなり高次の意味活動を対象とした研究が盛んであった．本稿では体系的な法的知識や推論を扱わずに，むしろ自然言語と知識を跨ぐ領域に焦点をあてる．法律ドメインの極めて短い文書の類似度判定において，これに世界知識 (法律分野の領域知識) と言語的知識の双方が含まれた語彙ネットワークを用いる手法を提案し，これを通じて法律等専門性の高い分野の言語処理における知識の導入のあり方について考える．

1.2　関連条文検索タスク

手法の適用・評価を行うためのタスクとして，司法試験問題文における関連条文の検索を設定した (図 1)．これは以下のようなものである．

日本の司法試験制度では，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，というのが典型的なパターンである．これらはおおよそ，法的知識を用いた，各選択肢文章の正誤判定に還元できる場合が多い．

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．条文知識とは文字通り，法律の条文内で規定されていることであり，判例知識は条文の適用例として条文と同程度に尊重されている判例のうち，よく出題されるものの知識である．ある問題で問われている知識がある条文／判例に含まれるとき，これをその問題の**関連条文**と呼ぶこととする．関連条文検索タスクは，与えられた試験問題文のテキスト (大問と選択肢の文章を接合する) に対して，番号の付与された条文のテキスト群から関連条文として相応しいものを選び，これを順位付けして出力するものである．

* 連絡先/E-mail: drowse314@gmail.com

平成 18 年度第 15 問の選択肢 2:

- 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。

第 398 条の 3:

- （根抵当権の被担保債権の範囲）
  1　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

図 1　司法試験問題文と関連条文の例

* 太字は法律語彙，その内ハイライトは問題文と条文の重なり語を示している / 条文において，上の問題回答に必要な知識は第 1 段落に全て含まれる．一方，第 2 段落の知識を問う問題も出題される可能性があり，文書内の重要箇所は問いによって変化すると言える．

本研究では，このタスクを民法の分野に限定し，かつ関連条文に判例を含まない問題を対象に実施することを目標とした．民法の短答式試験問題は法務省のウェブサイト[1]，民法条文は総務省の公開データ[2]より全て入手可能である．

本タスクの特徴について，簡単に述べる．関連条文検索は，問題文をクエリ文書，条文を検索文書とし，MeCab[3] および IPA 辞書を用いて文書から語彙を抽出することで，**ベクトル空間モデル**で解く文書検索の問題として捉えることができる．ただちに得られる形態素列は法律概念の粒度とは異なるが，これを名詞の連続等単純なルールでまとめ，またストップワードを設定することで，おおよそ適切な語彙が得られる．

この文書検索の大きな特徴として，問題文，条文ともに非常に短いテキストで構成されているということがある．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．このことは特に条文の記述において顕著で，ルールを定める対象について定義したり議論したりせず，法的な規定のみを簡潔に明記する．

文書検索では，文書に何度も登場する語がその文書の傾向をよく代表しているという仮定に基づき，語彙の文書内頻度がよく用いられる．一方関連条文検索では，上記の特徴によりこれを有効利用するのは困難である．すなわち，普通ひとつの条文にはある法律上のトピックについていくつかのルールが記述されるが，それぞれに割かれる語彙の数やテキストの量は，条文内の重要度や条文を代表しているかどうかでなく，単純に複雑な記述が必要か，あるいは簡潔に済ませることができるかということに依る．加えて法律文書は，比較的語彙の揺れ (同義語の出現) が少ないと言えるが，冗長性の低い記述傾向は，その中でも同義語による検索語の不一致を発生させ易い．

関連条文検索のもう一つの特徴は，検索対象の文書集合 (=条文) に類似したものが多いということである．例えば民法には，“抵当権” に直接関係する条文が 30 本程度含まれており，これらは当然語彙の構成がある程度似通っている．関連条文の検索では，この中から “抵当権” のある特定の側面について述べている条文を探索する，ということが要求されている．

提案手法では，法律概念／語彙の連想関係を保持したネットワークにより法的知識に基づいて語彙出現のコンテクストを補完することで，これら問題への対処を試みた．

[1] http://www.moj.go.jp/jinji/shihoushiken/jinji08_00082.html(平成 25 年度分) など

[2] http://law.e-gov.go.jp/htmldata/M29/M29HO089.html

[3] http://mecab.sourceforge.jp/

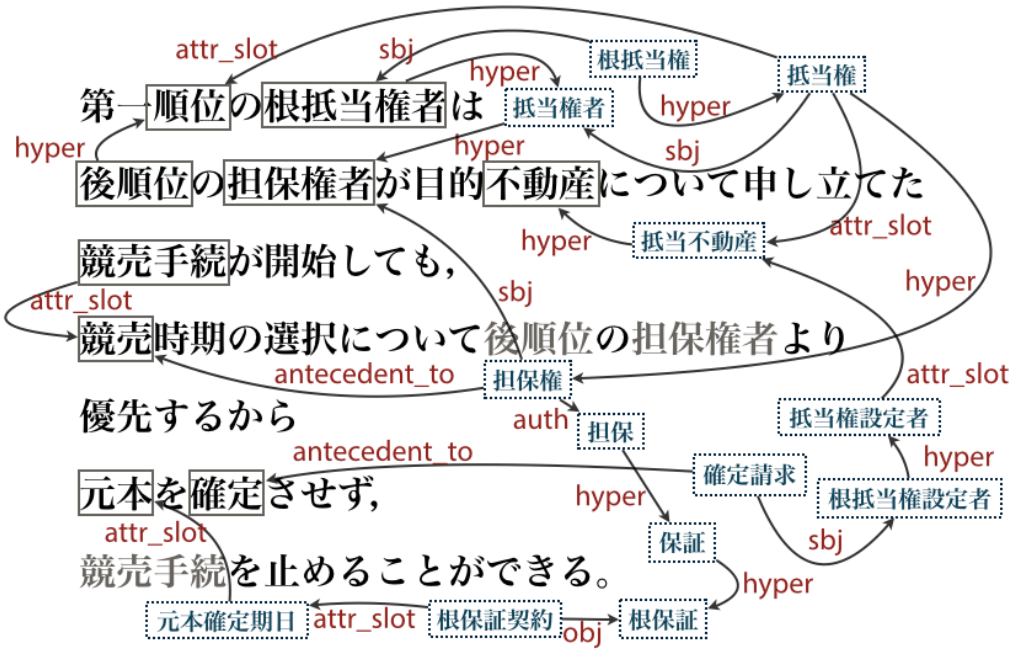

図 2　語彙の潜在的想起関係を辿る

＊ 青緑色の語彙は補完されたもの，灰色の語彙は既出のものである / 最新版のネットワークとは若干設定リンクが異なる

2　語彙ネットワークを用いた文書の主題類似度計算

2.1　語彙ネットワークの構築･利用とその意図

本研究では，問題文や条文に出現した語彙について，これを**潜在的な連想関係／想起関係**を辿って関連付け，更に出現していないが想起される語彙を補完することにより，出現語彙をコンテクストで捉えることを試みた．

私達はまず，実際の問題文および条文をいくつか用いて，出現語彙を整合的に関連付ける試みを，必要な語彙のリンク属性の選出と共に行った (図 2)．法律ドメインに極端に依存しないよう配慮しつつ，本研究では以下の関係属性を選定した．通常のシソーラスにも含まれる上位・下位関係を除けば，法律ドメインの知識を表現しうる関係を数多く含むことが分かるだろう．

- **hyper**: 上位・下位関係
  (例: 地上権 –> 用益物権)
- **sbj/obj**: 主格・目的格
  (例: 被担保債権 –> 担保権者 (sbj)
  競売 –> 財産 (obj))
- **auth**: 法的な根拠付け
  (例: 求償権 –> 償還請求)
- **within**: 場所
  (例: 履行 –> 履行地)
- **attr_slot**: フレームの予約属性
  (例: 管理者 –> 注意義務)
- **antecedent_to**: 因果・前提・順序関係
  (例: 競売 –> 買受け)

これらの関係性を用いて，民法典とその短答式試験問題文に出現しうるあらゆる語彙の関係性を定義したネットワークを構築した．規模は以下の通りである．ただし本研究で構築したのは，民法のトピックの中でも “根抵当権” という概念に関わる部分であり，語彙はこれを表現するのに必要最低限の範囲に制限されている．これは民法全体に対して，およそ 10% 弱程度のカバレッジであると予想される．

表 1　構築した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	7

法律ドメインにおいては，これまでもその体系的知識を表現するオントロジー [3] や，シソーラスの補助的な利用法 [4] が提案されて来た．本稿で提案するのは，法的知識を一意で明確な体系に落とし込むオントロジーでも領域知識を含まないシソーラスでもなく，語彙表記同士の関係性のレベルでこれらをゆるく組み合わせたネットワークのあり方である．

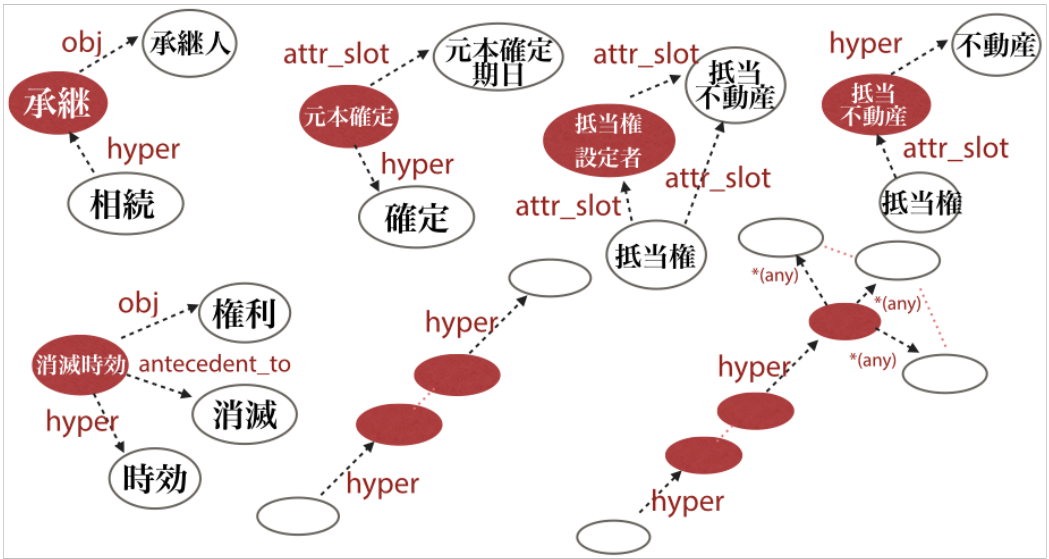

図 3　語彙補完のためのパターン

* 白抜きの語彙が出現した場合，赤色の語彙を補完する / 記した語彙は全て例で，リンク属性のパターンさえ合致すれば良い

なお，ここで作成した語彙ネットワークは，GitHub 上のリポジトリとして公開中である[4]．

続いてこの語彙ネットワークを用い，問題文および条文から抽出した語彙の関連付けを自動的に行うための手法を考えた．まず，図 2 のネットワークに示されるような，各文書に固有な語彙のつながりを，作成した語彙ネットワークから切り出す形で (人の手により) 作成する．これを注意深く観察し，頻出の接続／語彙補完パターンのうち妥当な説明が付与できるものを選出した (図 3)．そうして得たパターンを語彙ネットワークのグラフ探索処理として実装し，語彙の自動補完処理とした．

ここでは，補完された語彙集合により，結果的に問題文および条文のトピックがなるべく一意に示唆されることを期待している．一方，一般的に文書検索において語彙補完を始めとしたクエリ拡張を行う場合，再現率が向上する一方で適合率が低下する傾向，つまりノイズの増加が問題になりがちである (自動クエリ拡張に関する最近の議論は [5] など)．語彙ネットワークのグラフ探索による語彙補完は，あくまで基本的に出現語彙表記同士を整合的に関連付けることを目標としており，ノイズの混入を防ぎながらより明確に文書のトピックを示唆しようとするものである．

[4] https://github.com/drowse314-dev-ymat/lexical-knowledge-base-for-japanese-civil-law/tree/for_thesis_reference

2.2　類似度指標の計算

以上の語彙ネットワークによるコンテクスト補完を含めた，短答式試験問題文と条文の類似度指標の計算方法を示す．以下を関連条文検索タスクに適用する場合，与えられた問題文に対して，それぞれの条文との間で以下で述べる操作により問題文 – 条文間の量化された類似度を得，これを高いものから並べ，順位付けして出力する．大枠として，idf 重みを用いたベクトル空間モデルの所々に手を加えた形とした．従って予め，民法の各条文を検索対象の文書と考え，文書集合として民法全体をとって idf 重みを計算しておく (情報検索の基本的な計算操作に関しては全て [6] に準ずるものとする)．加えて，語彙ネットワークには民法の原文に含まれない補助的な語彙が多く含まれている．これらに idf 重みは設定されていないため，語彙ネットワークのグラフ上で線形にスムージングを行う形で補っておく．

まず先に述べたように，各文書の形態素解析結果を多少処理し，語彙を抽出した．抽出されたそれぞれの語彙集合について，先に述べた語彙ネットワークによる語彙の補完を行った．この際，補完された語彙のうち，以下のいずれにも該当しないものを切り捨てた．ただし，ここで “相手の文書” とは問題文にとっては条文，条文に対しては問題文のことを指す．

- 相手の文書の補完前の語彙集号に含まれている
- 相手の文書の補完候補の語彙に含まれている

これによってここでは，相手の文書のコンテクストを考慮した形で，よりミニマムな語彙の補完を目指した．

続いて，語彙補完の結果として上位・下位関係にある語彙の対が現れた場合，つまり，語彙の連想関係を辿る際，抽象化や具体化といったパスがあった場合に，以下の修正を行った．すなわち，補完された語彙のうち，下のいずれにも該当しないものを削除した．

- この概念 (語彙) の下位概念がその文書の語彙集合に存在しない
- この概念 (語彙) が比較対象の文書の補完前の語彙集合に含まれている

これは例えば，法律知識のある者が “抵当権” を話題にしているとき，抵当権が担保する債権であるところの “被担保債権” の概念を参照する際，「“抵当権” は “担保物権” の一種であり，“担保物権” のフレームにはそれが担保する “被担保債権” という属性が予約されており...」という抽象化を含んだ迂遠な思考過程は辿らず，直接「“抵当権” には “被担保債権” があるはずだ」という関係を得るだろうという直感に基づく修正操作である．

また，条文では，それぞれのルール記述に割かれるテキストの量が重要度ではなく，記述の複雑さに依っているということを既に述べた．これに対し試験問題文では，おおよそ全て，一文につきただ一つのルールに関する知識が問われることが，受験者へのインタビューでわかっている．このとき問題文に対する関連条文検索では，このただひとつの規定をその内に**含む**条文が取得されるべきであり，条文全体と類似しているか？という比較によって探索が行われるのは非合理的である．従って語彙に対する最後の操作として，問題文の語彙による条文の語彙のフィルタリング操作を行った．具体的には，条文の出現語彙 $T_{article}$ を，問題文の出現語彙 $T_{question}$ によって事前に $T_{article} = T_{article} \setminus T_{question}$ とした．このときそれぞれの “出現語彙” とは，語彙ネットワークによって補完済のものを指している．

以上の操作で問題文および条文の “出現語彙” に関する操作を終え，これをベクトル空間モデル下の類似度計算に用いた．先に述べた法律文書の非冗長な性質より文書内の語彙頻度は一切考慮せず，それぞれの “出現語彙” の idf 重みをその語彙の軸の成分とするベクトルを作成した．これをコサイン類似度に用い，最終的な類似度指標とした．

3 実験

本稿では，人手で整備を行った語彙ネットワークの構築規模の範囲内で，表 2 に示した 12 問を用い，提案手法の評価を行った．これ以降 “平成 18 年度 短答式試験 民法分野の大問 15 の選択肢 1” を “Q18/15/1” と短縮して記述する．また，比較のための手法として，以下の 3 項目を用意した．

- **ベースライン手法**: tf–idf 重みを文書ベクトルの成分とする手法
- **提案手法 1 − 知識を用いない改善手法**: 語彙の補完はせず，検索課題の分析に基づき条文の語彙のフィルタリングや文書内頻度の無視などを行う手法
- **提案手法 2 − 語彙ネットワークを用いた手法**: 第 2 章で述べた手法

提案手法 1 は，第 2 章で述べた手法から語彙ネットワークの利用に関する部分を差し引いたもので，法律文書の非冗長性等の考慮を含むものである．

それぞれの問いに対する正しい関連条文[5]に，ベースライン手法による順位付けの結果誤って上位に現れた 10 件程度の条文を加えた 63 件を検索対象とし，3 つの手法を比較した結果を表 2 に示す．各手法の結果として表示した数値は正しい関連条文の検索順位である．関連条文が複数あると判断されたもの，および類似度指標が全く等しいケースの処置等によって一部小数となったものがある．

全体的な傾向を見るに，いずれの提案手法でも，ベースライン手法との比較において問題文 – 関連条文間により適切な類似度判定および順位付けを行うことができたと言える．より細かく分析を行うと，問題文の出現語彙による条文の出現語彙のフィルタリングが最も劇的な改善効果を与えていたことが分かった．また，文書内語彙頻度のファクターの廃止にもかなりの改善効果があった．これらは第 1 章で述べた法律文書の分析の結果導入されたものであり，ここではその有効性を確認することができたと言える．

また，語彙ネットワークを用いた手法においては，“Q19/7/3” における順位の改善効果が際立っている．これは，問題文とその関連条文間に一切の語彙の一致が見られないケースであり，語彙ネットワークを用いたコンテクストの補完処理は，このようなケースの類似度指標を一定の水準まで引き上げるのに有効であると見ることができる．提案手法による語彙の補完につい

[5] 専門家のタグ付けによる

表 2　関連条文検索 – 正しい関連条文の順位

問題番号	ベースライン手法	**提案手法 1** – 知識を用いない改善手法	**提案手法 2** – 語彙ネットワークを用いた手法	正しい関連条文
Q18/15/1	1	1	1	398 の 20
Q18/15/2	7	2	1	398 の 3
Q18/15/3	2	1	3	398 の 2
Q18/15/4	2	1	1.5	398 の 7
Q18/15/5	1	1.875	1.875	398 の 12，398 の 13
Q19/13/エ	1	1	1	375
Q19/13/オ	8	1	1	398 の 3
Q19/16/1	1	1	1	395
Q19/16/2	1	1	1	387
Q19/16/4	8	8.5	9	379
Q19/16/5	1	1	2	371
Q19/7/3	44	44	5	575
平均	6.42	5.36	2.36	–

ては，先に述べた通りノイズの増加の可能性が予期されたが，実験結果を語彙補完のない手法と比較した限りでは，“Q19/7/3” 以外のケースについてはほぼ変化が見られないか，あるいは僅かな順位低下が見受けられる程度である．本実験の語彙補完では，問題文に関して見れば常に 4-5 語以上が追加されている．これを鑑みれば，文書のコンテクストを損なわずに語彙の連想関係を辿るという目標に対して，提案手法は非常によく機能しているといえる．

“Q19/7/3” において，語彙ネットワークを用いる手法により抽出された語彙の構成等を詳しく見ると，含まれる語彙の補完パターンよりも更に大胆な (ネットワーク上で迂遠な接続が必要な) 語彙の想起と補いが必要であることがわかる．一方で，今回提案手法 1 による適切な順位付けを維持していたいくつかのケースでは，これ以上の語彙補完を行うとノイズの混入が発生するものもあった．今後これを排除することができれば，より関連条文を上位に引き上げる余地があると言える．あるいはここから見えるのは，語彙の連想関係をどの程度深く辿るべきかは，問題文と関連条文の対に依って異なるということである．以上を考慮すると，本研究で構築した語彙ネットワークによる語彙の関連付け／ネットワークの探索プロセス自体に，より特定の問題文 – 条文対の語彙比較の場面に依存した処理を加えることで，より語彙ネットワークの利用価値を高めることができると考えられる．

“Q19/16/4” では，ネットワークを用いない改善手法，提案手法ともに改善効果が見られないが，この問いはより体系的な法的知識やメタ推論能力が求められるものであり，提案手法の解決目標とする言語と領域知識を跨ぐ問題圏にあるものではないと言える．

以上で述べた 2 つのケースにつき，この問題文 – 条文対を付録に示しておく．

4　おわりに

本稿では，司法試験問題の関連条文検索タスクにおいて，法律語彙の潜在的な連想関係を保持した語彙ネットワークを用いる手法を提案した．これにより，記述の冗長性が低い法律文書の抽出語彙について，コンテクストを維持したまま補完し，これを拡張することができた．

関連条文検索の性能の面では，提案手法がベースライン手法との比較においてより適切な順位付けを行うことを示した．また，語彙ネットワークによる語彙補完は，語彙の一致がない問題文 – 条文対について特に改善効果をもたらすことを示した．今後は，語彙ネットワークの構築・導入の効果を高めるために，より比較対象の文書双方を考慮しながら語彙の連想関係を活用することが考えられる．

また本稿では検討を行わなかったが，本研究で構築した語彙ネットワークは強く語彙表記上，テキスト上の関係を指向したものであり，日本語テキストを対象とした形式概念分析 ([7] など) と相性が良いと考えられる．本研究の成果物をシードとしてネットワークの連想関係をより密にする可能性が考えられる他，単純に手法の実施可能な問いを拡充し，評価の幅を広げることも期待される．

参考文献

[1] 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)

[2] 吉野一 (編): 法律人工知能—法的知識の解明と法律推論の実現, 創成社 (2000)

[3] 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 Vol.19, No.2, pp.144–150 (2004)

[4] 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 Vol.35, No.6, pp.986–996 (1994)

[5] Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval., *ACM Computing Surveys*, Vol.44, No.1, Article 1, pp.1–50 (2012)

[6] Manning, C. D., Raghavan, P., Schütze, H. (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 (2008)

[7] Duc, N.T., Bollegala, D., 石塚満: エンティティペア間類似性を利用した潜在関係検索, 情報処理学会論文誌, Vol.52, No.4, pp.1790–1802 (2011)

付録: 評価に用いた問題文と関連条文 (抜粋)

実験結果の考察の際に参照した 2 問の問題文，およびその関連条文をここに示す．

Q19/7/3

問題文

- 大問指示文: 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。
- 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、利息を支払うことを要しない。

Q19/16/4

問題文

- 大問指示文: 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。
- 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。