科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２３　日現在

機関番号：　１３９０１
研究種目：　基盤研究(B)
研究期間：　2011～2013
課題番号：　２３３０００９４

研究課題名（和文）法令文作成・英訳支援環境の構築：ターミノロジーと翻訳メモリの利用

研究課題名（英文）Construction of a Support Environment for Drafting and Translating Statutory Sentences: Utilization of a Terminology and a Translation Memory

研究代表者
外山　勝彦（TOYAMA, Katsuhiko）

名古屋大学・情報基盤センター・教授

研究者番号：７０２１７５６１

交付決定額（研究期間全体）：（直接経費）　　13,100,000円、（間接経費）　　3,930,000円

研究成果の概要（和文）：本研究の目的は，法令文作成支援と法令英訳支援のために，法令のターミノロジーおよび翻訳メモリの構築と利用のための手法と環境を確立することである．研究の結果，戦後のすべての日本語法律からなるコーパス（法律10,067本），戦後占領期における文対応付き日英対訳法律コーパス（法律1,624本，日英対訳156,562文）），法令翻訳メモリ（法令259本，日英対訳147,119文）を構築した．また，チャンキングや文書出現頻度を用いて法令用語を抽出する手法，対訳文からの対訳語彙意味カテゴリ自動抽出手法，法令用語とその語義文や法令用語間の関係を抽出する手法などを開発した．

研究成果の概要（英文）：The purpose of this research is to establish technologies and environments for compilation and utilization of a terminology and a translation memory to support drafting and translating Japanese statutory sentences. We have complied a text corpus including all Japanese acts (10,067 acts) enacted after the World War II, a bilingual corpus including 1,624 acts (156,562 bilingual sentences) enacted during the period of U.S. occupation, and a translation memory including 259 statutes, i.e., acts, cabinet orders and regulations (147,119 bilingual sentences). In addition, we have developed methods for extracting legal terms by using shallow parsing (chunking) or document frequency, semantic category from bilingual terms, and definitions of statutory terms as well as semantic relations between them.

研究分野：　総合領域

科研費の分科・細目：　情報学　図書館情報学・人文社会情報学

キーワード：　法律情報　自然言語処理　法令英訳支援　法令ターミノロジー　法令コーパス

１．研究開始当初の背景

(1)　法令英訳支援と問題点

国際社会のグローバル化に伴い，わが国はその法令を英訳することを内外から強く求められている．しかし，従来の法令英訳は人手により個別に行われており，信頼性，品質などの点で問題があった．特に，同一の語彙や表現に対して訳語が統一的に用いられていない場合があり，その意味の正確な理解に対して支障となることがあった．

そこで，本研究の研究代表者・連携研究者は訳語選択の統一を支援するため，対訳表現自動抽出技術に基づく手法やその支援ツール Bilingual KWIC® の開発により，「法令用語日英標準対訳辞書」構築や英訳法令の辞書準拠性検査に従事し，その成果は日本政府の法令英訳プロジェクトで活用された．しかし，単語列が処理対象であったため，連続単語列に対する対訳の抽出・検査までは有効であったが，「～の規定は、～について準用する」，「～した者は、～の～に処する」のような構文パターンに対する対訳パターンの抽出・検査は困難であった．

⑵　法令文作成支援と問題点

一方，国や地方自治体の政策は最終的に法令という形式で文書化されるので，その担当者は法令文の読解だけでなく，作成も必要となる．法令の解釈・運用を統一するために，法令文では慣習に基づく語彙・表現が用いられるので，法令文作成者はこの慣習を修得する必要がある．たとえば，「及び」と「並びに」や，「直ちに」，「速やかに」，「遅滞なく」は，それぞれ日常では類義語であるが，法令文ではその用法や意味が区別される．また，禁止規定に対して「～は、～してはならない」，定義規定に対して「この法律において「～」とは、～をいう」などのように，規定内容の類型に依存して，用いるべき構文パターンは定まっている．そのため，法令文作成は，この慣習を熟知した少数の専門家の人手による作業として実施されてきた．しかし，近年の社会情勢の急速な進展により，政策の適切かつ迅速な策定が求められているため，専門家の負担は増大しており，法令文作成に対する計算機支援が必要である．

⑶　構文情報の利用と新たな問題点

そこで，本研究代表者らは，科学研究費補助金基盤研究（Ｂ）「構文情報付き日英法令対訳コーパスに基づく法令文作成・英訳支援環境の構築」（平成 20～22 年度）において，依存関係など構文情報を付与した日英法令文対訳コーパスから法令用の語彙や法令文の構文パターン・対訳パターンを獲得し，それらを法令文作成者や翻訳者に提供する作業支援環境の構築を目指した．その結果，特有の接続詞を伴う並列構造や頻出する括弧書きなども扱えるような日本語法令文の依存関係の表現手法と半自動的な解析手法を確立し，それに基づき，現在までに 2,052 文（12 法令）に対して高精度の依存関係タグを付与した．また，依存関係の表示ツール KWISC を開発した．一方，依存関係の利用により，語と文脈の共起から類義語を自動獲得する手法の開発・比較や，英訳の統一性評価のための類似法令文への分類手法の開発も行った．それらにより，構文情報の利用による支援に見通しを付けた．

２．研究の目的

本研究は，法令文作成・英訳のための作業支援環境の構築を目的とする．特に，浅い構文情報と文字・語彙情報だけを利用して法令ターミノロジーや法令翻訳メモリを構築し，それらを言語資源として提供する作業支援環境の有効性を明らかにする．具体的には，次の⑴～⑷を行う．

⑴　浅い構文情報付き日英法令文対訳コーパスの構築

浅い構文情報の解析手法を開発し，浅い構文情報付き対訳コーパスを構築する．

⑵　法令翻訳メモリの構築

対訳コーパスから法令翻訳メモリを構築する手法を開発し，翻訳メモリを構築する．

⑶　法令ターミノロジーの構築

ターミノロジーへ蓄積する情報の表現手法の確立，および対訳コーパスや翻訳メモリなどからその情報を抽出する手法の開発を行い，ターミノロジーを構築する．

⑷　言語資源検索・法令文編集 GUI ツールの開発

法令文中の語彙や法令文自体を検索キーとして，対訳コーパス，翻訳メモリ，ターミノロジーに蓄積された情報を提示し，法令文作成・英訳作業を支援する GUI ツールを設計・開発する．

３．研究の方法

各項目について，それぞれ次の方法により研究を推進した．

⑴ 浅い構文情報付き日英法令文対訳コーパスの構築

戦後のすべての法律からなるコーパスを構築する．また，日本語法令文に対して浅い構文解析（チャンキング）を行い，名詞句，動詞句を抽出する手法を確立する．　それを用いて，法務省・法令外国語訳データベースシステム（JLT）が提供する日英対訳法令データ（法令約 200 本，異なり日本語文約 81,000 文）を対象としてチャンキングを行う．

⑵ 法令翻訳メモリの構築

法令文に対する翻訳メモリを構築する．その際, 翻訳メモリは標準化された TMX 形式で記述する．

また，法令翻訳メモリを拡充するための新たな手法を開発する．すなわち，対訳表現自動抽出技術と文字列の近接関係だけを利用して類義語や文脈パターンを抽出する手法 Monaka を結合し，原言語の類義語とその対訳表現を同時に抽出する手法を開発する．

⑶ 法令ターミノロジーの構築

法令ターミノロジーを構築するために，法令用語とその語義，用語間の関係を抽出する手法を開発する．

⑷ 言語資源検索・法令文編集 GUI ツールの開発

言語資源を検索しながら法令文作成・英訳を行える環境を構築するための GUI ツールを設計・開発する．

４．研究成果

本研究の主な成果は次のとおりである．

⑴ 浅い構文情報付き日英法令文対訳コーパスの構築

① 法務省・法令外国語訳データベースシステム（JLT）が提供する日英対訳法令のうち日本語法令 16 本(28,139 文)に対して，SVM に基づいたチャンカ YamCha を用いて，名詞句・動詞句を抽出する浅い構文解析（チャンキング）を施し，それを人手で修正した．さらに，それを学習データとした YamCha を用い，JLT が提供する日英対訳法令 221 本のうち日本語部分（延べ 120,449 文）に対して，チャンキングを施した．その結果，47,782 語（名詞句 38,935 個，動詞句 8,847 個，異なり数）を抽出し，各用語の出現頻度などの特徴量を求めた．

② 戦後のすべての日本語法律を収録するコーパス（昭和 21～平成 24 年，法律 10,067 本）を構築した．それらは，国立印刷局・官報データベースから抽出したもの（昭和 22～平成 24 年，法律 9,915 本）と国立国会図書館が提供している官報画像データベースからテキストを起こしたもの（昭和 21～22 年，法律 152 本）からなる．

③ 上記②で構築した日本語法律コーパスのうち，戦後占領期（昭和 21～27 年）に発行されていた英文官報に英訳が掲載されたもの 1,624 本に対して，①と同じ学習済みのチャンカ YamCha を用いたチャンキングを行ったところ，59,325 語（名詞句 48,360 語，動詞句 9,636 語）を抽出した．一方，出現文書頻度に基づく指標 df2/df を用いた方法により，138,564 個の日本語表現を抽出した．JLT が提供する「法令用語日英標準対訳辞書」に収録されている見出し語のうち，上述の法律 1,624 本に出現する 2,242 語と比較したところ，再現率は，チャンキングによる方法では 98.4%,df2/df を用いた方法では 75.5%であった．また，df2/df を用いた方法では，定型的な表現を抽出できることも確認した．

⑵ 法令翻訳メモリの構築

① JLT 収録の文対応付き日英対訳法令文 276,597 文(法令 259 本，異なり原文 129,120 文，異なり対訳文 147,119 文）を用いて，翻訳メモリを構築した．また，この翻訳メモリに対する検索システムも構築した．この翻訳メモリは標準的な TMX 形式に変換して，出力できるものである．

② ブートストラップに基づく対訳文からの対訳語彙意味カテゴリ自動抽出手法 b-Monaka を開発した．この手法は，語彙意味カテゴリ抽出と対訳表現獲得の 2 段階の処理を統合するものである．JLT 収録の文対応付き日英対訳法令 193 本（90,273 文）を用いて実験したところ，繰り返し 30 回において精度 82.0%で対訳表現を抽出できた．また，従来手法では困難であった複数の語からなる長い専門用語の抽出が可能であった．それにより，対訳文を語彙の文脈情報として使う方法の有効性を示すことができた．

③ 翻訳対象の法令文に類似した文を翻訳メモリから選択する手法を検討するために，地方自治体の条例 1,908 条を用いて，法令文の間の各種距離関数の比較と法令文のクラスタリング手法の比較を行った．その結果，距離関数としてはピアソン相関距離が，また，クラスタリング手法としてはウォード法が有効であることが分かった．

⑶ 法令ターミノロジーの構築

① 法令中の定義規定から正規表現によるマッチングによって，定義語とその語義文，および定義語の上位・下位関係を抽出する手法を開発し，JLT 収録の日本語法令 241 本（10,9380 文）から定義語 1,207 語とその語義文を抽出した．既存のシソーラスと比較したところ，上位･下位関係の判定精度は 64.0%であった．さらに，法令文中で括弧書きにより記述されている定義規定および略称規定について，法令文の文頭の主語にそれらの規定が出現しやすいという性質を利用して，定義語とその語義文を抽出する手法を開発した．前述の法令 241 本から定義語 1,941 語を抽出したところ，抽出精度はトイウ形定義語（1,501 語）が 90.9%，ヲイウ形定義語（440 語）が 77.7%であった．

② ⑴②で構築した日本語法律コーパスのうち，官報データベース収録分（昭和 22～平成 24 年，法律 9,915 本）から正規表現によるマッチングによって，延べ 14,874 個（異なり 9,368 個）の定義語とその語義文を抽出した．その抽出精度は 91.0%であった．しかし，括弧書きによる定義語の抽出再現率は 39.2%に留まった．

③ 上記②において，抽出された語義文には他の法文や法令を引用しているものが多く見られたので，その照応解決を行う手法を開発した．具体的には，JLT 用として既開発の文書型定義（DTD）を用いて，法令文書を構造化（XML 化）するとともに，文書構造を相対アドレスによって指定する引用表記に対して，正規表現を用いて文書構造の絶対アドレスをタグ付けした．そのタグ付け精度は 88.0%であった．また，相対アドレス指定による引用表記のうち 64.3%が絶対アドレス指定に正しく書き換えられた．

⑷ 言語資源検索・法令文編集 GUI ツールの開発

戦後占領期（昭和 21～27 年）の英文官報に掲載された英訳法律（1,624 本）に対して，画像データから英文テキストを起こし，それ

と(1)②で構築した日本語法律コーパスを用いて，文対応付き日英対訳法律コーパス（対訳文 156,562 文）を構築した．さらに．それらに対する対訳表現抽出用 GUI として，Bilingual KWIC® を構築した．

(5) その他

法令の要約文書である「法令のあらまし」の日英統計的機械翻訳手法について基礎的な技術を開発した．法令文日英対訳コーパス（276,597 文）を学習データとし，「法令のあらまし」日英対訳 300 文を開発データとしてデコーダのパラメータ調整を行い，機械翻訳システムの自動評価指標 BLUE，RIBES などを用いて評価したところ，本手法は有効であることを明らかにした．

５．主な発表論文等

〔雑誌論文〕（計１２件）

① D. Inagi, Y. Ogawa, M. Nakamura, T. Ohno, K. Toyama: Statistical Machine Translation for Outlines of Japanese Statutes, Proc. 7th Int. Workshop on Juris-informatics, pp.37-49 (2013) 査読有.

② 小川泰弘，中村誠，外山勝彦: 法律文中における単語出現頻度の変化－法令テキストマイニングの一例－，名古屋大學法政論集，Vol.250, pp.543-556 (2013) 査読無.

③ 角田篤泰: 法令・例規における定義規定の記述方法と理論的背景，名古屋大學法政論集，Vol.250, pp.505-541 (2013) 査読無.

④ 外山勝彦，齋藤大地，関根康弘，小川泰弘，角田篤泰，木村垂穂，松浦好治: 日本法令外国語訳データベースシステムの設計と開発，情報ネットワーク・ローレビュー，Vol.11, pp.33-53 (2012) 査読有.

⑤ Y. Sekine, K. Toyama, Y. Ogawa, Y. Matsuura: The Development of Translation Memory Database System for Law Translation, Proc. 2012 Law via the Internet Conf., 21 pages, http://blog.law.cornell.edu/lvi2012/files/downloads/2012/10/LVI2012_sekine_final.pdf (2012) 査読有.

⑥ M. Nakamura, R. Kobayashi, Y. Ogawa, K. Toyama: A Pattern-Based Approach to Hyponymy Relation Acquisition for the Agricultural Thesaurus, Proc. Joint Int. Symp. on Agricultural Ontology Service 2012, pp.2-9 (2012) 査読有.

⑦Y. Ogawa, M. Mori, K. Toyama: Recall-Oriented Evaluation Metrics for Consistent Translation of Japanese Legal Sentences, New Frontiers in Artificial Intelligence: JSAI 2011 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science, Vol.7258, pp.141-154, Springer (2012) 査読有.

⑧ R. Jin, Y. Ogawa, R. Agro, K. Toyama: Bootstrapping-based Extraction of Bilingual Dictionary Terms from Parallel Corpus, Proc. Joint Int. Symp. on Natural Language Processing and Agricultural Ontology Service 2011, pp.95-99 (2012) 査読有.

⑨ Y. Ogawa, M. Yamada, R. Kato, K. Toyama: Design and Compilation of Syntactically Tagged Corpus of Japanese Statutory Sentences, New Frontiers in Artificial Intelligence: JSAI 2010 Conference and Workshops, Revised Selected Papers, Lecture Notes in Computer Science, Vol.6797, pp.141-152, Springer (2011) 査読有.

⑩Y. Ogawa, M. Mori, K. Toyama: Recall-Oriented Evaluation Metrics for Consistent Translation of Japanese Legal Sentences, Proc. 5th Int. Workshop on Juris-informatics, pp.62-73 (2011) 査読有.

⑪ K. Toyama, D. Saito, Y. Sekine, Y. Ogawa, T. Kakuta, T. Kimiura, Y. Matsuura: Design and Development of Japanese Law Translation Database System, Proc. Law via the Internet Conf. 2011, 12 pages, http://www.hklii.hk/conference/paper/1C2.pdf (2011) 査読有.

⑫ 萩原正人，小川泰弘，外山勝彦: グラフカーネルを用いた非分かち書き文からの漸次的語彙知識獲得，人工知能学会論文誌，Vol.26, No.3, pp.440-450 (2011) 査読有.

〔学会発表〕（計１９件）

① 李寧，小川泰弘，大野誠寛，中村誠，外山勝彦: 中国語専門用語抽出における CRF 法とブートストラップ法の比較，言語処理学会第 20 回年次大会（2014. 3.18）北海道大学（北海道）.

② 福田薫，外山勝彦，野田昭彦: 学内情報翻訳データベースの構築と運用，大学 ICT 推進協議会 2013 年度年次大会（2013.12.18）幕張メッセ国際会議場（千葉県）.

③ 李寧，小川泰弘，大野誠寛，中村誠，外山勝彦: 中国語専門用語の抽出における単語分割の影響，平成 25 年度電気関係学会東海支部連合大会（2013. 9.24）静岡大学浜松キャンパス（静岡県）.

④ 稲木大，小川泰弘，中村誠，大野誠寛，外山勝彦: 統計的機械翻訳に基づく法令のあらましの日英翻訳，平成 25 年度電気関係学会東海支部連合大会（2013. 9.24）静岡大学浜松キャンパス（静岡県）.

⑤ M. Nakamura, Y. Ogawa, K. Toyama: Extraction of Legal Definitions and Their Explanations with Accessible Citations, 5th Workshop on Artificial Intelligence and Complex Legal Systems (2013.12.11). Bologna (Italy).

⑥ Y. Ogawa, D. Inagi, M. Nakamura, K.

Toyama: Translation for Outlines of Japanese Acts, 2013 Law via the Internet Conf., (2013. 9.27) Jersey, Channel Islands (UK).
⑦ M. Nakamura, Y. Ogawa, K. Toyama: Extraction of Defined Legal Terms and their Explanations from a Japanese Legal Corpus - Towards Construction of a Legal Term Ontology, 2013 Law via the Internet Conf. (2013. 9.26) Jersey, Channel Islands (UK).
⑧ **中村誠, 小川泰弘, 外山勝彦: 法令文中において括弧書きで定義されている法令用語とその語釈文の抽出, 言語処理学会第 19 回年次大会 (2013. 3.15) 名古屋大学 (愛知県).**
⑨ **外山勝彦: H24 年度 JaLII 関係研究・開発成果, 法情報学の今後の展開を考える会 (2013. 3. 8) 加賀 (石川県).**
⑩ K. Toyama: Compilation of a Translation Dictionary of Legal Terms in Four Jurisdictions in East Asia, International Symposium on from Legal Assistance to Legal Cooperation - Exploring the New Horizon - **(国際シンポジウム「法整備支援から法協力へ －新たな地平の開拓－」)**, (2012.12. 9) **名古屋大学 (愛知県).**
⑪ A. H Shee, Y. Matsuura, K. Toyama, T. Kakuta, Y. Sekine: Making Legal Information Smart, Friendly and Inspiring, 2012 Law via the Internet Conf. (2012.10.9) Cornell Univ, Ithaca (USA).
⑫ **外山勝彦, 小川泰弘: ブートストラップ法に基づく日英対訳コーパスからの対訳用語自動抽出, 平成 24 年度電気関係学会東海支部連合大会シンポジウム「ここまでできる言語処理技術－音声・言語情報処理の最先端－」(2012. 9.25) 豊橋技術科学大学 (愛知県) 招待講演.**
⑬ R. Agro, **小川泰弘, 外山勝彦: コーパス内の文字出現頻度による言語間漢字変換ツールの作成, 平成 24 年度電気関係学会東海支部連合大会 (2012. 9.24) 豊橋技術科学大学 (愛知県).**
⑭ R. Agro, **小川泰弘, 外山勝彦: コーパス内の文字出現頻度による言語間漢字変換ツールの作成, 言語処理学会 NLP 若手の会第 7 回シンポジウム (2012. 9. 4) 東北大学 (宮城県).**
⑮ **外山勝彦: 名古屋大学法情報研究センターの最近の活動, 法情報学の今後の展開を考える会 (2012. 1.27) 加賀 (石川県).**
⑯Y. Sekine, K. Toyama: Design and Development of Japanese Law Translation Memory Database System, Int. Conf. on Legal Information and East Asian Law: Theories, Practices and Prototypes **(法律資訊與東亞法學國際學術研討會－理論. 實踐. 典範) (2012. 6.15) 中正大学, 嘉義 (台湾).**
⑰ R. Agro, **小川泰弘, 外山勝彦: 構造指向型漢字検索システム, 平成 23 年度電気関係学会東海支部連合大会 (2011. 9.27) 三重大学 (三重県).**
⑱ **森雅紀, 小川泰弘, 外山勝彦: 法令対訳表現に対する再現率指向の統一性評価指標, 平成 23 年度電気関係学会東海支部連合大会講演 (2011. 9.27) 三重大学 (三重県).**
⑲K. Toyama: Brief Introduction to Bilingual KWIC for Taiwan Laws **(台灣法律 Bilingual KWIC 簡要介紹)**, TaiwanLII Opening Ceremony and Round Table Forum **(台灣法律資訊中心開幕茶會與圓卓會議) (2011. 6. 7) 中正大学, 嘉義 (台湾).**

〔その他〕
① ホームページ等
英文官報 Bilingual KWIC®
http://kwic.law.nagoya-u.ac.jp/Official_Gazette_en/
② 受賞
大学 ICT 推進協議会 2013 年度年次大会優秀論文賞 (2013).

６．研究組織
(1)研究代表者
外山　勝彦 (TOYAMA, Katsuhiko)
名古屋大学・情報基盤センター・教授
研究者番号：70217561

(2)連携研究者
小川　泰弘 (OGAWA, Yasuhiro)
名古屋大学・情報基盤センター・准教授
研究者番号：70332707

角田　篤泰 (KAKUTA, Tokuyasu)
名古屋大学・大学院法学研究科・特任准教授
研究者番号：80292001

松浦　好治 (MATSUURA, Yoshiharu)
名古屋大学・大学院法学研究科・特任教授
研究者番号：40104830