平成 27 年度特許庁委託事業

平成２７年度
中国特許文献の機械翻訳の品質評価及び辞書整備に関する調査

報告書

平成 28 年 2 月

株式会社　高電社

目次

1. 調査目的

近年、世界の特許文献において、日本語以外の言語で記載された外国特許文献の割合が急増している。世界で通用する安定した権利の設定を行うためには、こうした外国特許文献の的確かつ効率的な先行技術文献調査が不可欠である。なかでも、増加が著しい中国特許文献を審査官及び外部ユーザーが容易に調査できる環境を実現しなければ、不十分な先行技術文献調査による権利の安定性の低下に加え、日本企業が進出先において現地企業から訴えられる可能性の増大等のリスクが高まるおそれがあることから、中国特許文献への日本語によるアクセス性の向上が必要である。

増加の著しい中国特許文献へ日本語によりアクセスするには、大量の文章を翻訳することができる機械翻訳が効率的である。特許庁でも、機械翻訳等を活用した日本語による中国特許文献の検索システムの開発を重点的に進めており、平成27 年1 月より中韓文献翻訳・検索システム[1]を公開している。当システムでは、特許文献で用いられる技術用語の対訳辞書を用いることで、翻訳精度の向上を図っており、そのために特許庁では、平成24 年度に「中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査」、平成25 年度に「中国特許文献の機械翻訳のための新語に関する調査」、平成26 年度に「中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査」をそれぞれ実施し、中日のパテントファミリーなどから中日対訳辞書データ作成して当システムで活用している。また、これらの調査において、対訳辞書を作成する過程で生成された対訳コーパスについても、特許文献の統計的機械翻訳の研究等で活用している。技術の進歩に伴う新しい技術用語に対応し、中日特許文献の機械翻訳精度を引き続き向上させていくためには、技術分野ごとの中日機械翻訳の精度を的確に把握し、必要な分野において中日対訳辞書データをアップデートしていく必要がある。

また、機械翻訳の精度を適切に評価するためには、評価のための基準が必要となることから、特許庁は、平成25 年度に「特許文献機械翻訳の品質評価手法に関する調査」を実施し、「特許文献機械翻訳の品質評価手順[2]」を作成している。この評価手順は、使用していく中で事例を追加するなど適宜改訂していくことを想定しているものである。

本調査では、①中国特許文献の中日機械翻訳文の人手による評価と自動評価をそれぞれ行うことで、中国特許文献の中日機械翻訳の技術分野ごとの精度を把握するとともに、②機械翻訳の精度が低い技術分野において重点的に中日対訳辞書データを作成し、さらに、③「特許文献機械翻訳の品質評価手順」についての改善すべき点を把握することを目的とする。

---

1 http://www.ckgs.jpo.go.jp/

2 http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

2. 調査概要

2.1 調査の内容

2.1.1 中国特許文の翻訳精度調査・分析

中国特許文の翻訳精度調査・分析について、図2.1.1-1に概要図を示す。

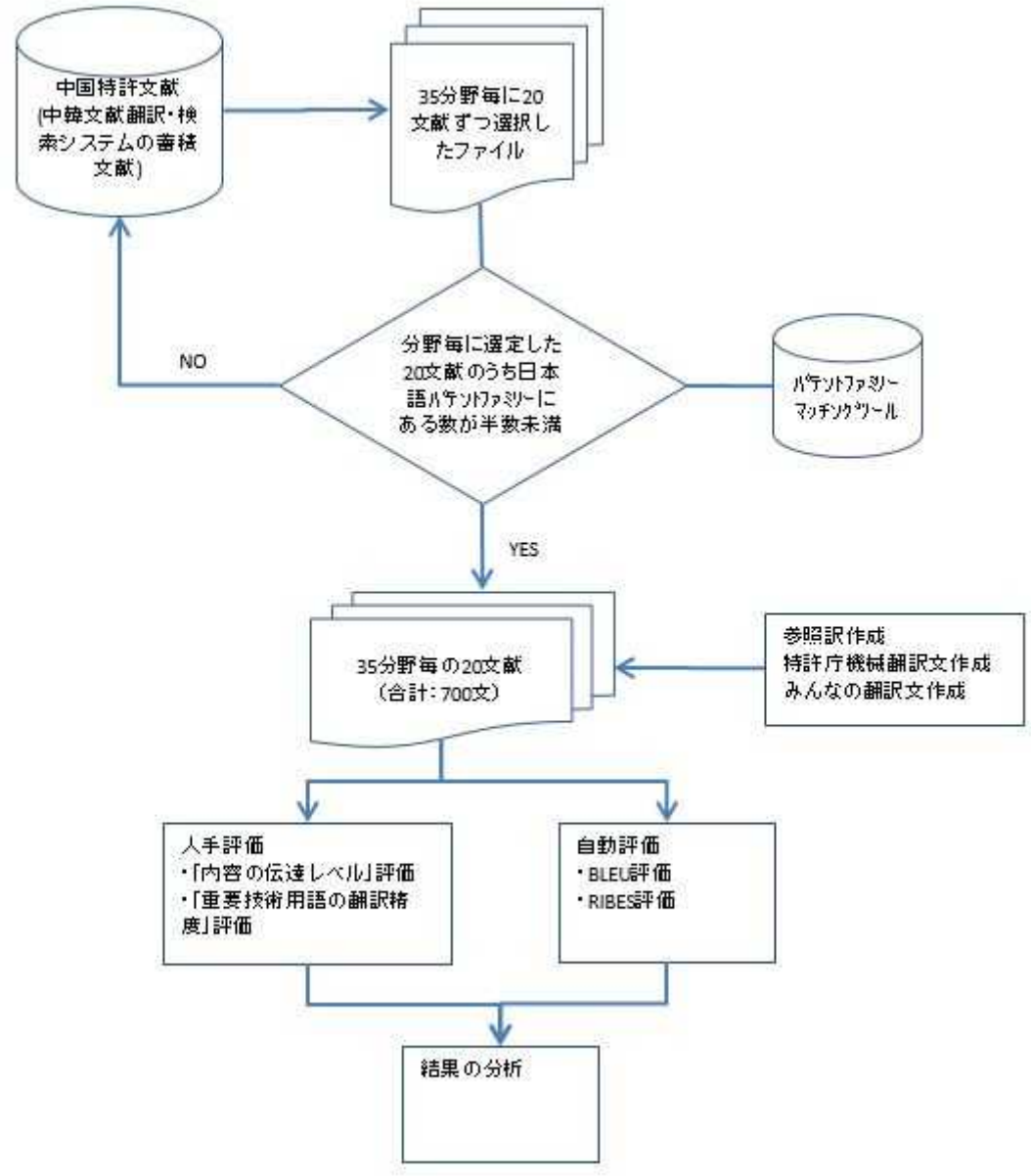


図2.1.1-1中国特許文の翻訳精度調査・分析概要図

①中国特許文選定、基準翻訳文作成、機械翻訳文作成

WIPOが作成した「Technology Concordance Table[3]」に基づいた35技術分野ごとに、中国

[3] http://www.wipo.int/ipstats/en/statistics/technology_concordance.html

特許文献（「中韓文献翻訳・検索システム」の蓄積対象のもの）を20文献ずつ選択し、それらの発明の詳細な説明に相当する部分の中から、1文献につき1文ずつ中国語の文を合計700文抽出した。この際、日本語のパテントファミリー[4]のある文献は技術分野ごとに半数未満にした。また、抽出した中国語の700文について、人手による基準翻訳文を作成した。さらに、中国語の700文に対応する機械翻訳文を、「中韓文献翻訳・検索システム」及び「みんなの自動翻訳[5]」サービスを利用して、２セット用意した。

②人手による評価と自動評価

用意した２セットの機械翻訳文について、「特許文献機械翻訳の品質評価手順」に従って、人手により「内容の伝達レベル」及び「重要技術用語の翻訳精度」の評価を実施した。さらに、用意した２セットの機械翻訳文について、BLEUとRIBES[6]による自動評価を行った。

③翻訳精度の評価の分析

２セットの機械翻訳文について、人手による評価と自動評価の結果を分析し、35技術分野ごとの機械翻訳精度の差を明らかにした。また、人手による評価と自動評価の結果を比較し、両者に差がある場合には、差の要因と考えられる点を分析した。

④「特許文献機械翻訳の品質評価手順」についての改善点の検討

人手評価を行う中で、「特許文献機械翻訳の品質評価手順」について発見された問題点等をまとめた。

---

4 内外国を通じて、共通の優先権を持ち、技術内容が完全又は部分的に一致する関係を有する特許文献群。

5 情報通信研究機構(NICT)が提供する無料のオンライン機械翻訳サービス
https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

6 http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html

2.1.2 対訳辞書データの作成・分析

対訳コーパスから辞書作成処理について、図2.1.2-1に概要図を示す。

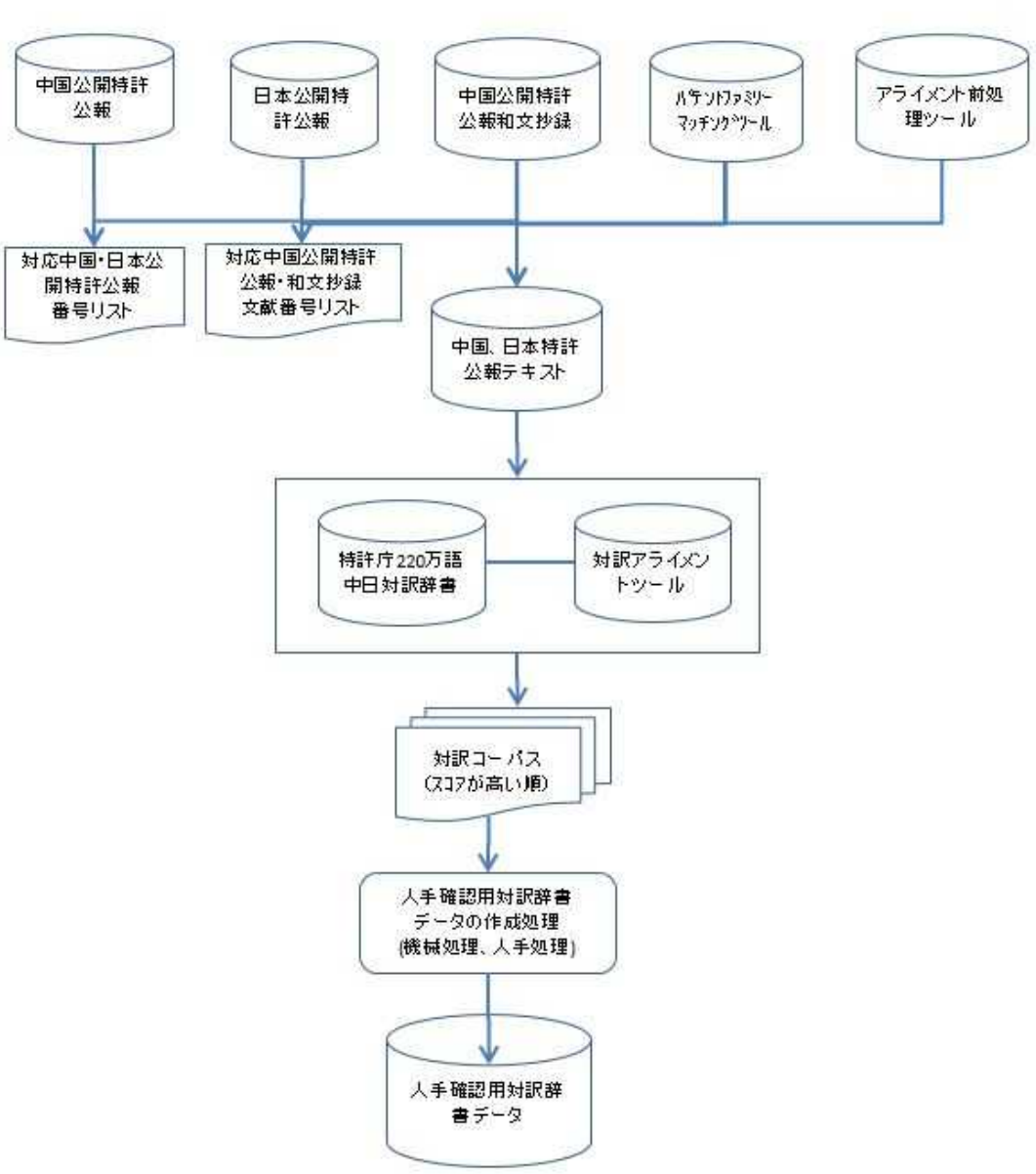


図2.1.2-1 対訳コーパスから辞書作成概要図

①対訳コーパスの作成

DOCDB[7]に蓄積されているパテントファミリー情報を利用し、DOCDB における”family id”が同一の中国公開特許公報と日本公開特許公報を、技術内容が対応するものと判断して、中国の文献と日本の文献の対を抽出した。また、中国公開特許公報和文抄録とそれに

[7] ＥＰＯ（欧州特許庁）が発行している世界７０カ国以上の公開特許公報を収録したデータベース

対応する中国公開特許公報の対を用意した。これらの中国語と日本語の文献の対の全てを用いて、対訳コーパス、対応中国・日本公開特許公報番号リスト、対応中国公開特許公報・和文抄録文献番号リストを作成した。

②対訳コーパスからの対訳辞書データの作成

作成した対訳コーパスを用いて、人手確認用対訳辞書候補データを作成し、人手確認用対訳辞書候補データの採否を訳語確認者が判断することで、５万語対以上の対訳辞書データを作成した。

③「未知語」への訳付けによる対訳辞書データの作成

特許庁が貸与する「未知語」を含むサンプル文（「中韓文献翻訳・検索システム」において未知語と推定された語を含む中国語の文）を複数参照し、②で作成した対訳辞書データ及び特許庁が貸与する対訳辞書のいずれにも含まれず、かつ、辞書に登録する用語としてふさわしい未知語に対し、未知語がどのような文脈で出現するかを理解しつつ、訳付けを行い、7,000語対以上の対訳辞書データを作成した。

未知語サンプル文から対訳辞書データの作成について、図2.1.2-2に概要図を示す。

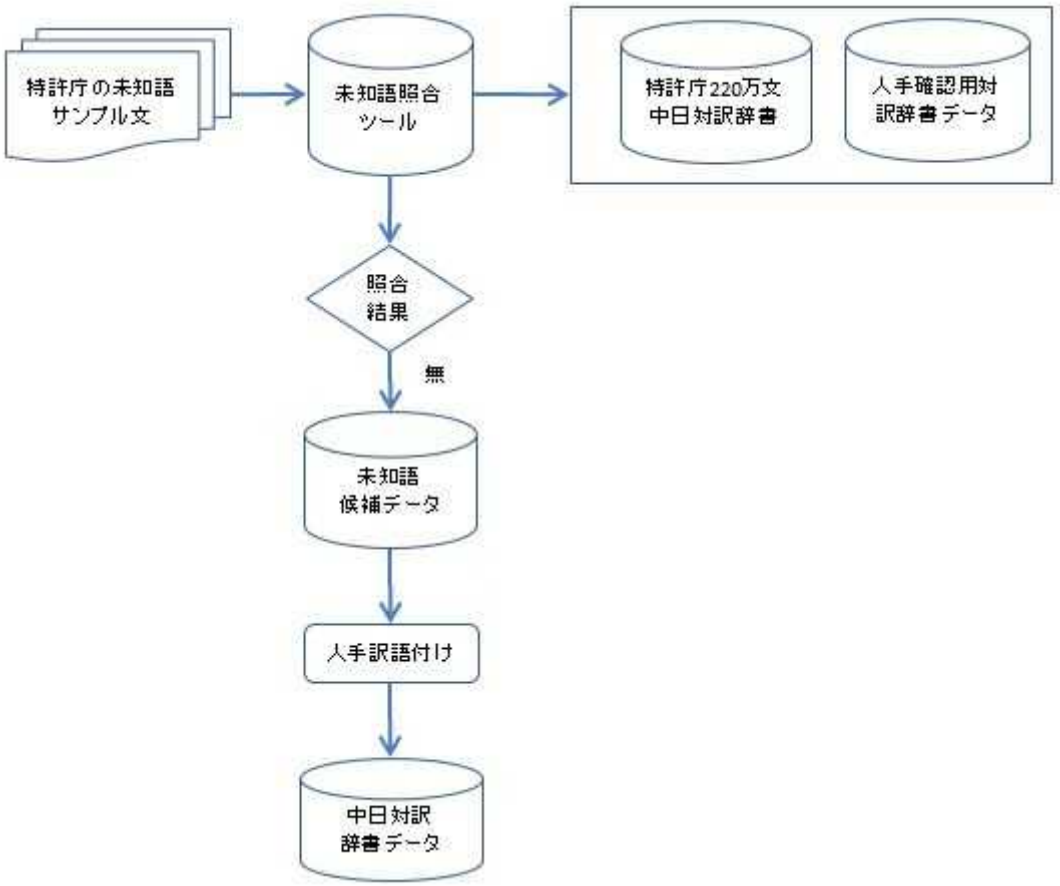


図2.1.2-2 未知語サンプル文から辞書作成概要図

④統合した対訳辞書データの作成

コーパスと未知語から作成した中日対訳辞書と特許庁が貸与する中日対訳辞書データを統合し、「見出し語（中国語）、訳語（日本語）、品詞（見出し語（中国語））、品詞（訳語（日本語））、特許庁が指定する出現頻度情報」が付された形式とした上で、ＵＴＸ形式（UTX1.11） に変換し、中日対訳辞書データを作成した。

2.2 調査実施体制

本調査作業は表2.2-1に示す体制にて実施した。

表 2.2-1 実施体制

チーム名	主な役割	要員数
統括責任者	本調査全体の統括。	1 名
翻訳品質評価チーム	基準翻訳文作成、翻訳品質評価、辞書データの人手確認	5 名
対訳コーパス、辞書作成チーム	対訳コーパス、対訳辞書データの作成	3 名
報告書作成チーム	報告書の作成	4 名

2.3 調査スケジュール

本調査は平成27年8月から平成28年2月の期間に表2.3-1のスケジュールで実施した。

表 2.3-1 調査実施スケジュール

作業項目		開始日	終了日	8月	9月	10月	11月	12月	1月	2月
大分類	小分類									
翻訳精度の調査	中国語原文、基準翻訳文及び評価対象の機械翻訳文の作成	2015/8/3	2015/8/20	●						
	評価対象の機械翻訳文の作成	2015/9/2	2015/9/4		●					
	人手評価と自動評価、評価の分析、品質評価手順	2015/8/21	2015/9/11	●	●					
対訳辞書データの作成分析	作業用ツールの整備	2015/8/3	2015/10/2	●	●	●				
	対訳コーパス作成	2015/9/14	2015/10/9		●	●				
	対訳コーパスから対訳辞書データの作成	2015/10/13	2015/12/18			●	●	●		
	未知語訳付けによる対訳辞書データの作成	2015/12/21	2016/2/5					●	●	●
	統合した対訳辞書データの作成	2016/2/8	2016/2/17							●
	対訳辞書データの分析	2016/2/18	2016/2/23							●
報告書の作成	作業報告書	2015/8/3	2016/2/23	●	●	●	●	●	●	●

2.4 調査環境

2.4.1 ソフトウェア

本調査では大量の特許文献からコーパスの作成、辞書の作成作業が必要となる。対訳コーパスと辞書は表2.4.1-1に示したツールを利用して作成した。

表2.4.1-1　主要ツール

ツール名	用途(ツール情報 URL)
中日対訳文アライメントツール	中日対訳コーパスの作成 http://www.nict.go.jp/
中日対訳用語抽出ツール	対訳コーパスから対訳辞書候補の抽出 http://www.nict.go.jp/
統計的機械翻訳システム Moses	対訳コーパスからフレーズテーブルの抽出 http://www.statmt.org/moses/
中国語単語分割器	中国語文の単語分割
中国語品詞付与ツール	中国語単語の品詞付与
日本語形態素解析器	日本語文の形態素解析

2.4.2 ハードウェア

本調査では大量の特許データを扱い、長時間の機械処理が必要であるため、表2.4.2-1に示したマシンで並列処理等を行い、作業を効率化した。

表2.4.2-1　ハードウェア

処理	スペック(CPU / Memory)
特許文献選定	Core i7-3770 3.40GHz / 32GB
中日自動文アライメント	Xeon x3450 2.67GHz / 32GB
対訳辞書データ作成	Xeon x3430 2.40GHz / 24GB
頻度情報計算	Xeon x3430 2.40GHz / 16GB

3. 中国特許文の翻訳精度調査・分析

3.1 評価方法

3.1.1 評価対象データの作成

（１）評価文選定

①WIPOが作成した「Technology Concordance Table」に基づいた35技術分野(表3.1.1-1)ごとに特許庁の貸与物である「中国公開特許公報テキストデータ（2012年～2014年公開分）」から中国特許文献（「中韓文献翻訳・検索システム」の貯蓄対象のもの）を各技術分野内でIPCが偏らないように20文献ずつ選択する。

表3.1.1-1 35技術分野表

セクタ	番号	技術分野	IPC 番号
電気工学	1	電気機械、電気装置、電気エネルギー	F21#, H01B, H01C, H01F, H01G, H01H, H01J, H01K, H01M, H01R, H01T, H02#, H05B, H05C, H05F, H99Z
	2	音響・映像技術	G09F, G09G, G11B, H04N-003, H04N-005, H04N-009, H04N-013, H04N-015, H04N-017, H04R, H04S, H05K
	3	電気通信	G08C, H01P, H01Q, H04B, H04H, H04J, H04K, H04M, H04N-001, H04N-007, H04N-011, H04Q
	4	デジタル通信	H04L
	5	基本電子素子	H03#
	6	コンピューターテクノロジー	(G06# not G06Q), G11C, G10L
	7	ビジネス方法	G06Q
	8	半導体	H01L
機器	9	光学機器	G02#, G03B, G03C, G03D, G03F, G03G, G03H, H01S
	10	計測	G01B, G01C, G01D, G01F, G01G, G01H, G01J, G01K, G01L, G01M, (G01N not G01N-033), G01P, G01R, G01S; G01V, G01W, G04#, G12B, G99Z
	11	生物材料分析	G01N-033
	12	制御	G05B, G05D, G05F, G07#, G08B, G08G, G09B, G09C, G09D
	13	医療機器	A61B, A61C, A61D, A61F, A61G, A61H, A61J, A61L, A61M, A61N, H05G
化学	14	有機化学、農薬	(C07B, C07C, C07D, C07F, C07H, C07J, C40B) not A61K, A61K-008, A61Q
	15	バイオテクノロジー	(C07G, C07K, C12M, C12N, C12P, C12Q, C12R, C12S) not A61K
	16	製薬	A61K not A61K-008
	17	高分子化学、ポリマー	C08B, C08C, C08F, C08G, C08H, C08K, C08L

	18	食品化学	A01H, A21D, A23B, A23C, A23D, A23F, A23G, A23J, A23K, A23L, C12C, C12F, C12G, C12H, C12J, C13D, C13F, C13J, C13K
	19	基礎材料化学	A01N, A01P, C05#, C06#, C09B, C09C, C09F, C09G, C09H, C09K, C09D, C09J, C10B, C10C, C10F, C10G, C10H, C10J, C10K, C10L, C10M, C10N, C11B, C11C, C11D, C99Z
	20	無機材料、冶金	C01#, C03C, C04#, C21#, C22#, B22#
	21	表面加工	B05C, B05D, B32#, C23#, C25#, C30#
	22	マイクロ構造、ナノテクノロジー	B81#, B82#
	23	化学工学	B01B, B01D-000#, B01D-01##, B01D-02##, B01D-03##, B01D-041, B01D-043, B01D-057, B01D-059, B01D-06##, B01D-07##, B01F, B01J, B01L, B02C, B03#, B04#, B05B, B06B, B07#, B08#, D06B, D06C, D06L, F25J, F26#, C14C, H05H
	24	環境技術	A62D, B01D-045, B01D-046, B01D-047, B01D-049, B01D-050, B01D-051, B01D-052, B01D-053, B09#, B65F, C02#, F01N, F23G, F23J, G01T, E01F-008, A62C
機械工学	25	ハンドリング機械	B25J, B65B, B65C, B65D, B65G, B65H, B66#, B67#
	26	機械加工器具	B21#, B23#, B24#, B26D, B26F, B27#, B30#, B25B, B25C, B25D, B25F, B25G, B25H, B26B
	27	エンジン、ポンプ、タービン	F01B, F01C, F01D, F01K, F01L, F01M, F01P, F02#, F03#, F04#, F23R, G21#, F99Z
	28	繊維、製紙	A41H, A43D, A46D, C14B, D01#, D02#, D03#, D04B, D04C, D04G, D04H, D05#, D06G, D06H, D06J, D06M, D06P, D06Q, D99Z, B31#, D21#, B41#
	29	その他の特殊機械	A01B, A01C, A01D, A01F, A01G, A01J, A01K, A01L, A01M, A21B, A21C, A22#, A23N, A23P, B02B, C12L, C13C, C13G, C13H, B28#, B29#, C03B, C08J, B99Z, F41#, F42#
	30	熱処理機構	F22#, F23B, F23C, F23D, F23H, F23K, F23L, F23M, F23N, F23Q, F24#, F25B, F25C, F27#, F28#
	31	機械部品	F15#, F16#, F17#, G05G
	32	運輸	B60#, B61#, B62#, B63B, B63C, B63G, B63H, B63J, B64#
その他	33	家具、ゲーム	A47#, A63#
	34	その他の消費財	A24#, A41B, A41C, A41D, A41F, A41G, A42#, A43B, A43C, A44#, A45#, A46B, A62B, B42#, B43#, D04D, D07#, G10B, G10C, G10D, G10F, G10G, G10H, G10K, B44#, B68#, D06F, D06N, F25D, A99Z

	35	土木技術	E02#, E01B, E01C, E01D, E01F-001, E01F-003, E01F-005, E01F-007, E01F-009, E01F-01#, E01H, E03#, E04#, E05#, E06#, E21#, E99Z

技術分野によってIPCの種類が多様なため、下記表3.1.1-2の基準通り文献の選択を実施した。

表3.1.1-2　IPC番号の種類の数によって選択したルール

IPC 番号の数	選択ルール
20 種類以上	同じ IPC 番号は 2 回以上選択しないように制限
20 種類未満	同じ IPC 番号を選択出来るが、選択した IPC 番号はなるべく均等で偏らないように制限。

②選択した中国特許文献が特許庁の「中韓文献翻訳・検索システム」の公開番号に存在するかのチェックを行い、存在する文献のみ対象にした。

③選択した中国文献について、日本語のパテントファミリーのある文献は技術分野ごとに半数未満であることをチェックした。

④抽出した35技術分野の合計700文献から発明の詳細な説明に相当する部分の中から、１文献につき１文ずつ中国語の文を抽出した。抽出する中国文は、文字数が20文字以上130文字以下の文とした。

（２）基準翻訳文作成

機械翻訳の精度調査を行う際、基準翻訳文はとても重要であるため、基準翻訳文の正確さが求められている。選定した中国特許文献の700文について、日中双方ネイティブレベルかつ、特許文献の翻訳経験がある作業者が、各自得意な技術分野を担当して基準翻訳文を作成した。また、中国文献を選定する際、日本語のパテントファミリーのある文献は対応中国文に該当する日本語文を付けて、不備があるかのチェックを行い、翻訳を行った。最後に作成した基準翻訳文を翻訳者同士で訳文のチェックを行った。

（３）機械翻訳文準備

「中韓文献翻訳・検索システム」及び無料のオンライン機械翻訳サービス「みんなの自動翻訳」を利用して、抽出した700文の評価文に対する機械翻訳文を作成した。

①「中韓文献翻訳・検索システム」の機械翻訳文

※「中韓文献翻訳・検索システム」の機械翻訳について、一括翻訳が出来ないため、下記a～cの作業を700回繰り返した。

a.「中韓文献翻訳・検索システム」の公報番号索引機能を利用して、評価文の各公報番号の照会を行った。

b.照会結果から原文を表示し、原文から評価文を検索した。

c.該当評価文に対応する訳文をb処理のa処理の照会結果から特定した。

②「みんなの自動翻訳」の機械翻訳文

抽出した700文の評価文に対して一括翻訳を実施し、機械翻訳文の訳を作成した。

なお、「みんなの自動翻訳＠TexTra®[8]」は国立研究法人情報通信研究機構がWebで提供（利用には登録が必要）している統計翻訳エンジンである。

中日翻訳について、いくつかのエンジンが利用可能であるが今回は「JPO特許：中国語-日本語」を使用した。

「みんなの自動翻訳＠TexTra®」翻訳サーバスペックは下記の通りである。

CPU：2.60GHz (32Core)／メモリ：128GB／HDD：542GB

[8] https://mt-auto-minhon-mlt.ucri.jgn-x.jp/content/menu/

3.1.2 人手評価方法

（1）「内容の伝達レベル」評価

「内容の伝達レベル」の評価について、表3.1.2-1の評点基準で行った。

表3.1.2-1 内容の伝達レベルの5段階評価

評点	評点基準
5	すべての重要情報が正確に伝達されている。(100%)
4	ほとんどの重要情報は正確に伝達されている。(80%～)
3	半分以上の重要情報は正確に伝達されている。(50%～)
2	いくつかの重要情報は正確に伝達されている。(20%～)
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。(～20%)

「内容の伝達レベル」の評価について、下記の通り実施した。

a. 各評価者は「特許文献機械翻訳の品質評価手順」の「内容の伝達レベル」基準を熟読し、意見交換を行い、評価手順の理解を行った。

b. 評価対象の700文の中から、35各技術分野別に各2文ずつ、合計70文を選定した。

c. 評価者を2チームに分けて、チーム毎に選定した70文の評価を実施した。

d. 両チームで評価した結果、評点の差が2点以上に生じた文について、お互いに意見交換を行い、意識を合わせ、再評価を実施した。

e. 意識合わせを行った上で、700文から評価を行った70文以外の630文に対して2チームで評価を行い、最終的に平均値の評価結果を取得した。

（2）「重要技術用語の翻訳精度」評価

「重要技術用語の翻訳精度」の評価について、表3.1.2-2の評点基準で行った。

表3.1.2-2 重要技術用語の評価

評点	評点基準
A(適訳語)	人手翻訳に照らし、技術的に同義かつ一般的に用いられる訳語である。
B(可訳語)	技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。
C(誤訳語)	誤訳である。
D(不訳語)	未知語、訳漏れである。

「重要技術用語の翻訳精度」の評価について、下記の通り実施した。

a. 「重要技術用語の翻訳精度」の評価対象の単語抽出作業は、中国語・日本語双方堪能で、特許の専門分野の単語知識を有する担当者で行った。

b. 重要技術用語について、１文から1単語の抽出を基本とした。ただし、文中に技術用語に相応しい単語がない場合、同じ分野の別の文で２つ抽出し、一つの分野で抽出された重要技術用語の数は20個以上とした。
c. 特許翻訳経験の長い評価者二人は「重要技術用語の翻訳精度」の基準を熟読し、意見交換を行い、評価手順の理解を行った。
d. 各評価者は技術用語の評価を行い、評価結果が一致してない用語について、評価結果が一致するまで再評価を行った。

3.1.3 自動評価方法

評価文 700 文の自動評価に使用した BLEU 及び RIBES について概要を説明する。

(1) BLEU

BLEU[9] (bilingual evaluation understudy) は、現在最も良く使われる翻訳精度の自動評価指標である。

$$BLEU = BP * \sqrt[4]{p(1) * p(2) * p(3) * p(4)}$$

p(n)は参照訳に対する翻訳結果の n-gram 適合率であり、次のようになる。

$$p(n) = \frac{翻訳結果中の重複なしで参照訳と一致するn\text{-}gram数}{翻訳結果の全n\text{-}gram数}$$

このように、翻訳結果の形態素の 1 ～ 4 個の連続した並びが、どのくらい参照訳のそれらに一致するかという適合率の相乗平均になっている。ここで再現率を考慮せずに適合率のみを計算することには以下の問題がある。すなわち翻訳結果が形態素 1 個であったとして、参照訳の文中にその同じ形態素が含まれるだけで、適合率は 1 (100%)となってしまう。再現率を考慮しないことで生じる問題を回避するため、BLEU では翻訳結果が参照訳より短い場合は、短さのペナルティ BP を乗じている。ここで、翻訳結果の形態素数を t、参照訳の形態素数を r とすると、BP は次のようになる。

$$BP = \begin{cases} 1 & (t > r) \\ exp\left(1 - \frac{r}{t}\right) & (t \leq r) \end{cases}$$

翻訳結果が参照訳より長い場合、BP 項は BLEU の値に影響を及ぼさない。翻訳結果が参照訳より短い場合は、BP 項が 1 より小さくなり結果として BLEU の値が小さくなる。それぞれの適合率については、p(1)は、形態素 (単語) の訳の正確さ(adequacy)を、また p(2)～p(4)は、形態素の連続した並びの適合率を判定するため訳の流暢さ(fluency)を、それぞれ評価していると考えられる。

[9] https://en.wikipedia.org/wiki/BLEU

BLEU は評価指標が単純でまた演算コストが掛からないため、一般に広く使われている。反面、例えば翻訳結果と参照訳とで形態素 4-gram の並びが適合しない場合、どんなに 3-gram(トライグラム)までの適合率が高くても、BLEU 値がゼロになってしまうといった問題がある。このため 1 文毎の評価にはあまり使用されない。また、類義語や同義語を考慮するには、適合率の計算時に複数の参照訳を用意する必要がある。

⑵ RIBES

RIBES[10] (rank-based intuitive bilingual evaluation score) は、近年日本で開発された翻訳精度の自動評価指標である。ここでは簡単に原理だけを示す。

$$RIBES = \frac{\tau + 1}{2} * p^{\alpha} \quad (0 \leq \alpha \leq 1)$$

τ は Kendall の順位相関係数である。参照訳と翻訳結果とで形態素が対応するペアの数を n とすると、翻訳結果の形態素から参照訳への形態素の全ペア数は n 個から 2 個を選ぶ組み合わせの数になるので、そのうち昇順になるペア数を P と置けば、

$$\tau = 2 * \frac{昇順ペア数}{全ペア数} - 1 = \frac{2 * P}{\frac{n * (n-1)}{2}} - 1$$

と定義される。τ は 2 つの文の形態素の並びが完全に逆順になる時に値 -1 を取り、完全に同じ順序になる時に値 1 を取る。RIBES は、τ の値域が 0～1 となるように正規化したものだと言える。ただし、順位相関係数だけでは、語の訳の正確さについては何も考慮していない。そのため、1-gram(ユニグラム)の適合率 p を使い、語の訳の正確さについてのペナルティを課している。べき乗の α は、ペナルティをどの程度 RIBES 値に反映させるか決定するためのチューニングパラメーターである。今回の自動評価作業では、デフォルト値の $\alpha = 0.25$ とした。

RIBES は原言語もしくは目的言語が日本語となる翻訳において、従来の機械評価よりも人手評価と高い相関性を示すことが知られており、NTCIR9 の自動翻訳タスクで自動評価基準として採用されるなど高い実績を持つ。

[10] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html

3.2 人手評価結果

3.2.1 内容の伝達レベルによる評価結果

700文の特許庁翻訳とみんなの翻訳による翻訳結果を「内容の伝達レベル」の観点から人手評価した結果を以下（１）～（４）に示す。

（１） 全体(700文)の評価結果

全体(700文)の評価の平均と標準偏差を図3.2.1-1に示す。700文の評価の結果をみると、特許庁翻訳の平均値は3.15点、みんなの翻訳では、3.87点であり、みんなの翻訳の方が0.72点高かった。また、図3.2.1-2に示すように、特許庁翻訳では、3点の文の割合が一番高く、みんなの翻訳では、図3.2.1-3に示すように、4.5点の文の割合が一番高く、全体的にみんなの翻訳の方が高評点の度数が多かった。

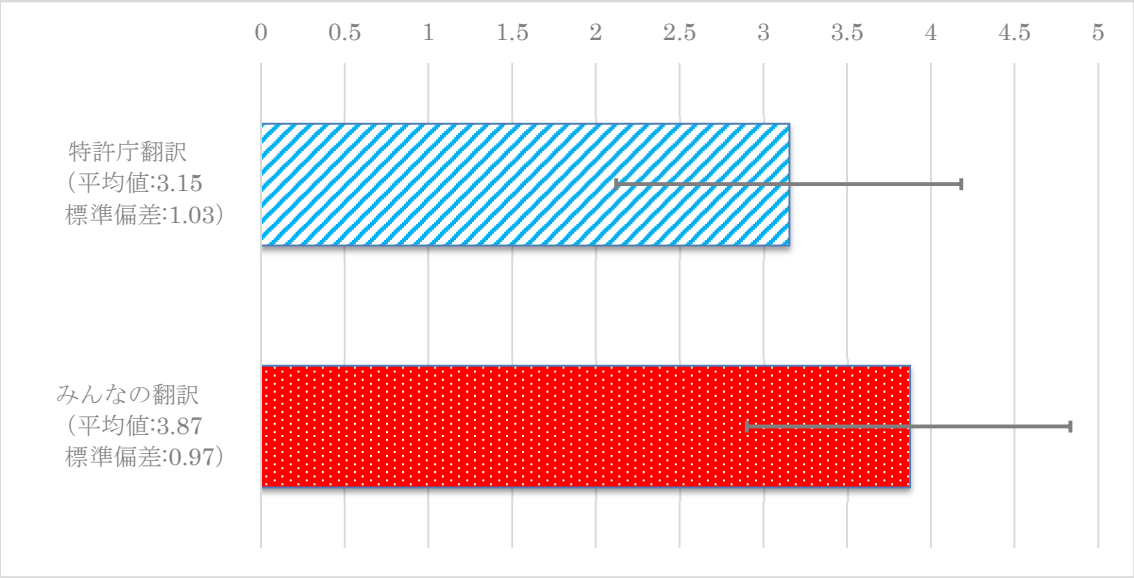


図3.2.1-1 内容の伝達レベルの平均値

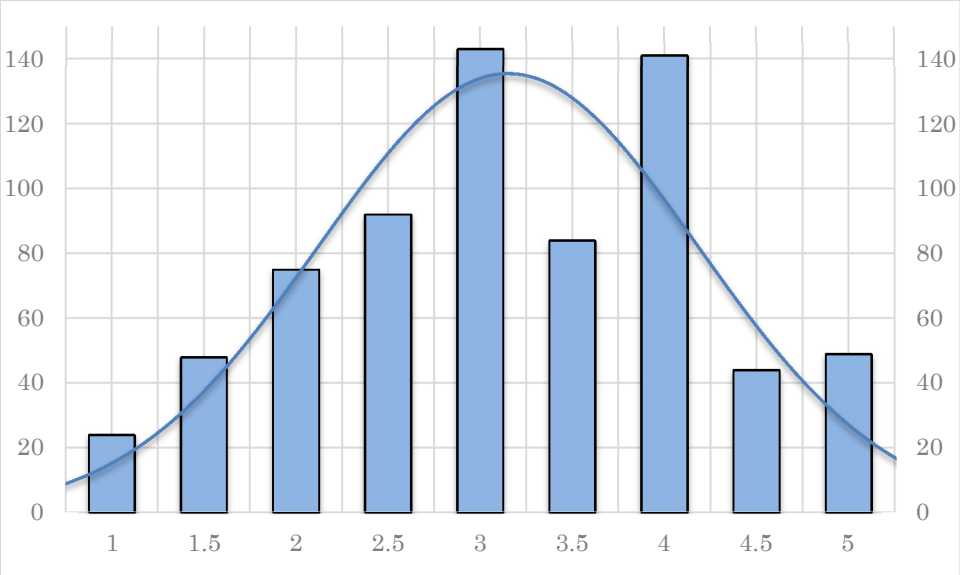


図3.2.1-2 特許庁翻訳700文の点数分布（平均値：3.15、標準偏差：1.03）

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	24	48	75	92	143	84	141	44	49
割合	3.4%	6.9%	10.7%	13.1%	20.4%	12.0%	20.1%	6.3%	7.0%

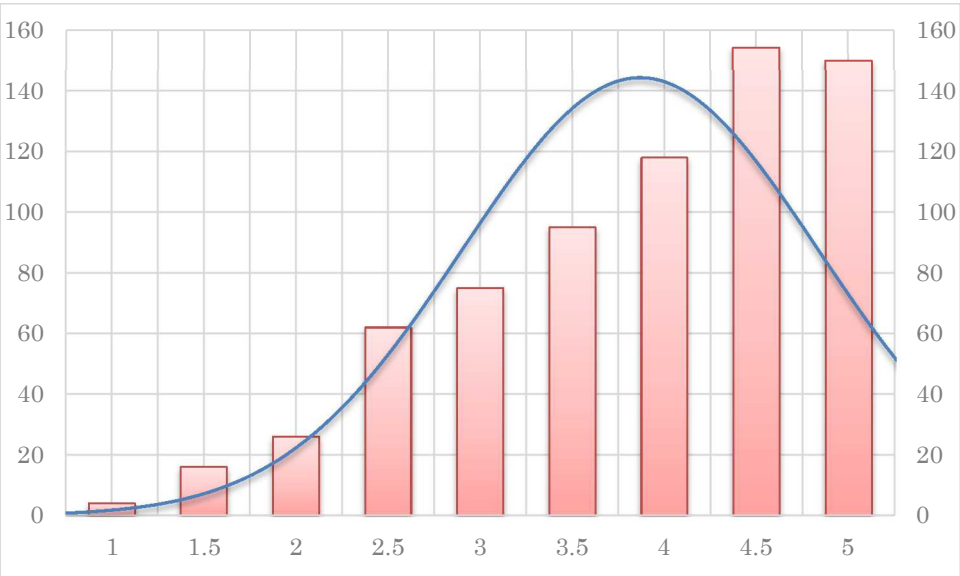


図3.2.1-3 みんなの翻訳700文の点数分布（平均値：3.87、標準偏差：0.97）

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	4	16	26	62	75	95	118	154	150
割合	0.6%	2.3%	3.7%	8.9%	10.7%	13.6%	16.9%	22.0%	21.4%

（２）セクタ毎(５分野)の評価結果

セクタ別の評価結果をみると、特許庁翻訳では、化学分野の評価値が高く、電気工学分野の評価値が低かった。みんなの翻訳でも、化学分野の評価値が高く、電気工学分野の評価値が低かった。評価の低い分野では、専門用語の誤訳や構文（係り受け）の誤りが比較的多く見られた（詳細は3.3.4 分野毎の誤訳原因参照)。

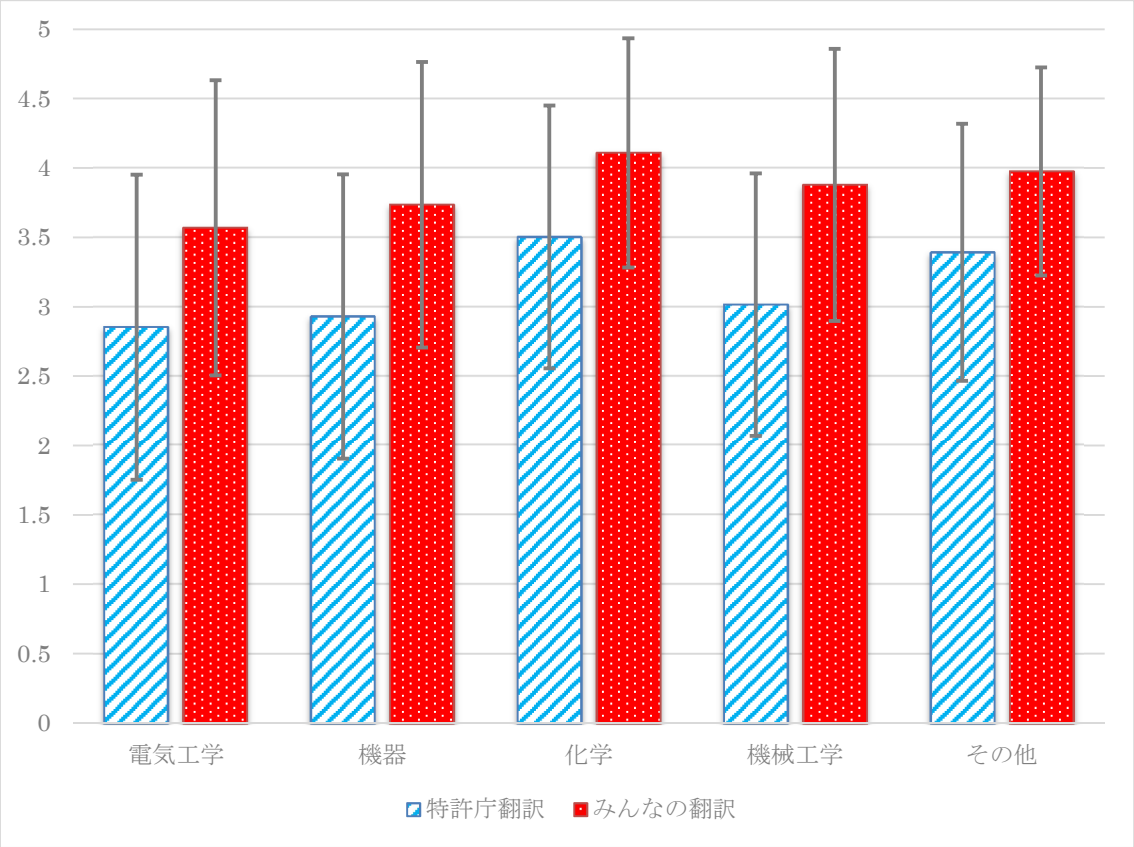


図3.2.1-4 セクタ毎の内容の伝達レベルの平均値

（３）35技術分野別の評価結果

35技術分野別の評価結果をみると、特許庁翻訳では、平均評価点数が3.0以下の評価結果が14技術分野もあった。一方、みんなの翻訳では、全ての技術分野の平均評価点数が3.0以上であった。

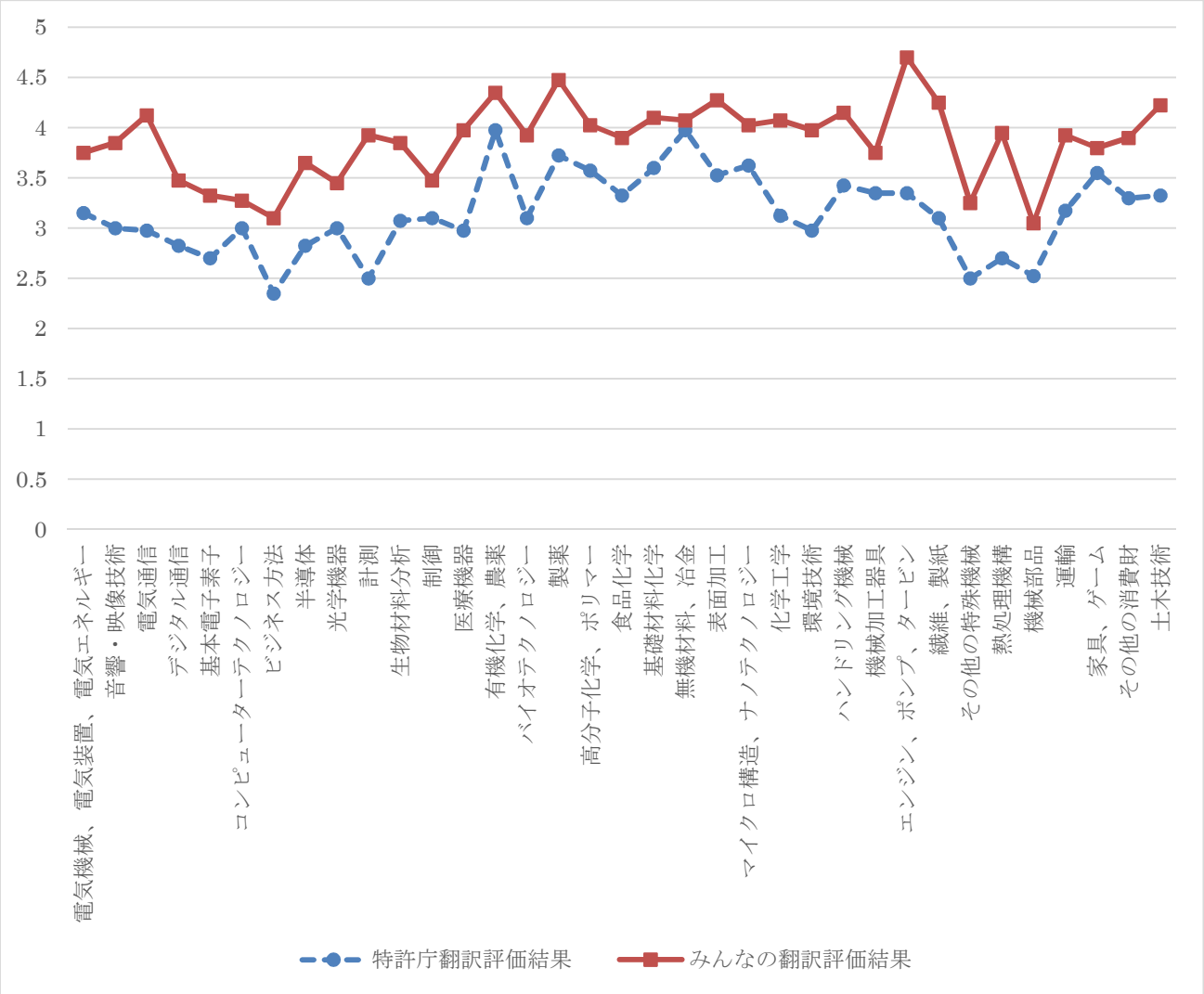


図3. 2. 1-5　35技術分野別の内容の伝達レベルの平均値

（４）文長別の評価結果

評価文700文を文字長が52文字以下の350文と53文字以上の350文の２つのグループに分けた。文長別の評価の平均値を図3.2.1-6に示す。評価の結果から、特許庁翻訳とみんなの翻訳はいずれも52文字以下の方が53文字以上より、平均評価点数が高いことが分かる。

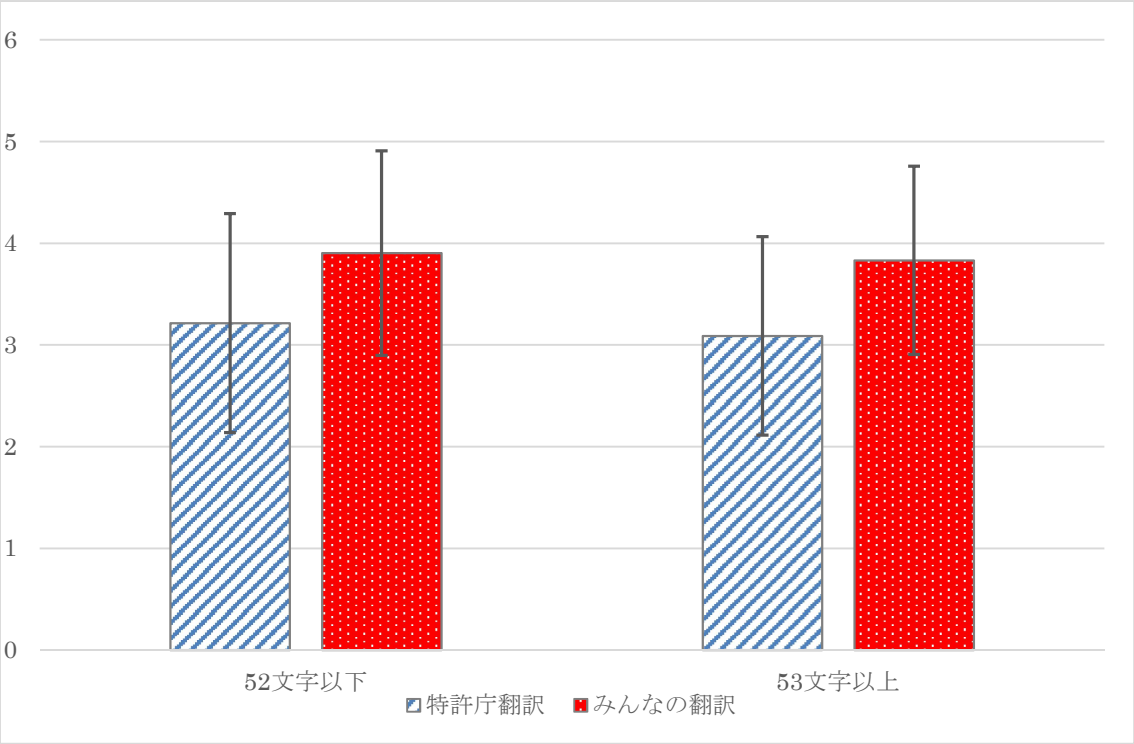


図3.2.1-6 特許庁翻訳とみんなの翻訳の文長別の平均値

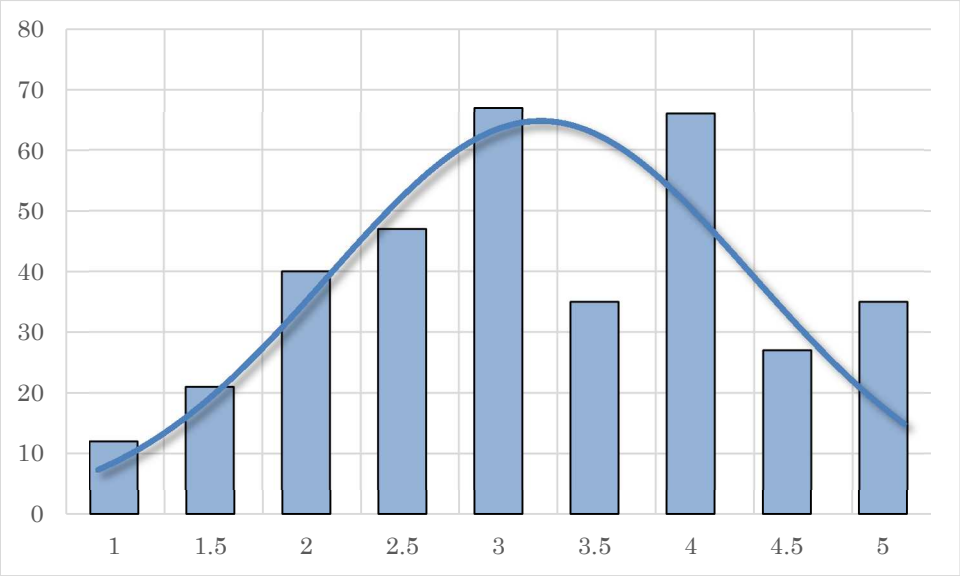


図3. 2. 1-7 特許庁翻訳の52文字以下の点数分布

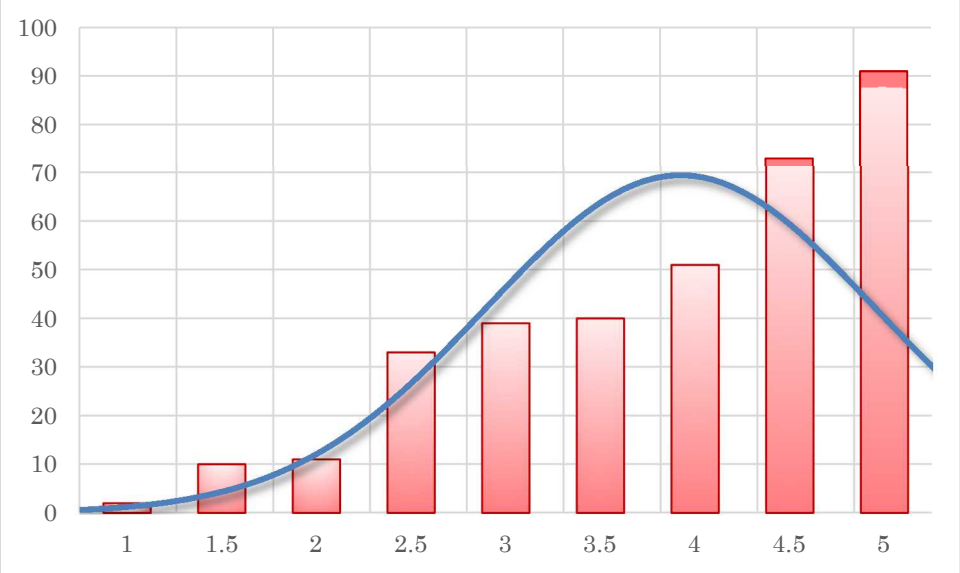


図3. 2. 1-8 みんなの翻訳の52文字以下の点数分布

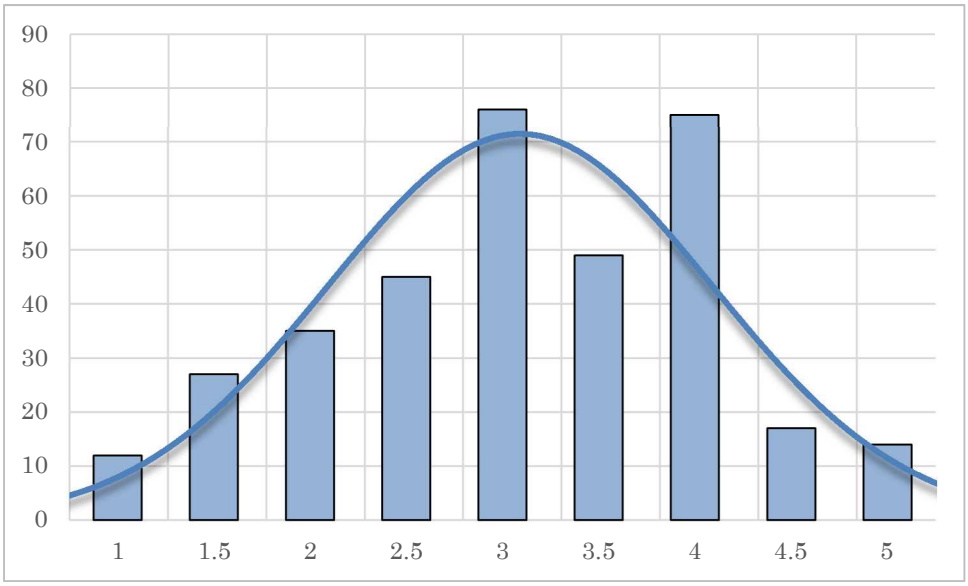


図3. 2. 1-9 特許庁翻訳の53文字以上の点数分布

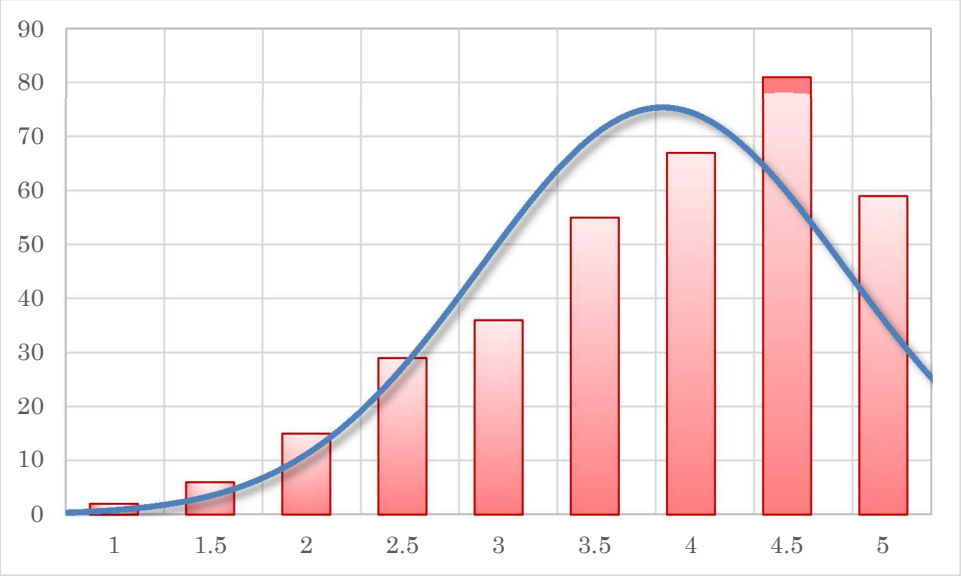


図3. 2. 1-10 みんなの翻訳の53文字以上の点数分布

3.2.2 重要技術用語の評価結果

特許庁翻訳とみんなの翻訳による翻訳結果を「重要技術用語の翻訳精度」の観点から人手評価した結果を、「A（適訳語）」「B（可訳語）」「C（誤訳語）」「D（不訳語）」をそれぞれ４～１の評点に置き換えて、以下（１）～（３）に示す。

（１）全体(700語)の評価結果

重要技術用語の全体(700文)の評価結果を図3.2.2-1に示す。結果をみると、特許庁翻訳の適訳語の割合が69.8%であり、みんなの翻訳では、82.7%であり、みんなの翻訳が特許庁翻訳より適訳語の割合が12.9%高かった。

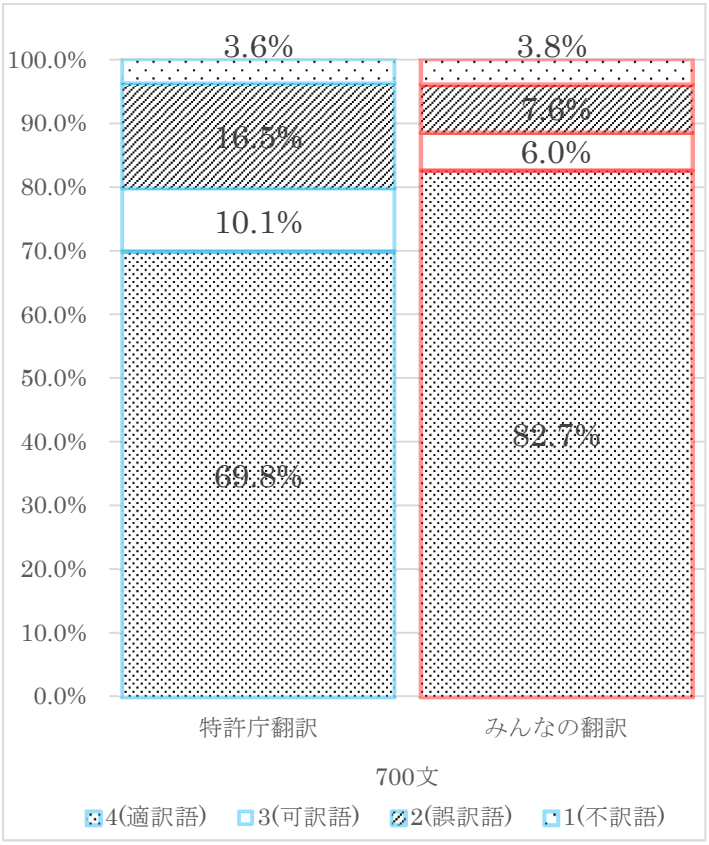


図3.2.2-1 700文の点数分布グラフ

（２） セクタ毎（５分野）の評価結果

セクタ別の評価結果をみると、特許庁翻訳では、「その他」分野の適訳語の割合が一番高く、「化学」分野の不訳語の割合が一番低い。一方、みんなの翻訳では、「その他」分野の適訳語の割合が一番高く、「電気工学」分野の不訳語の割合が一番低い。

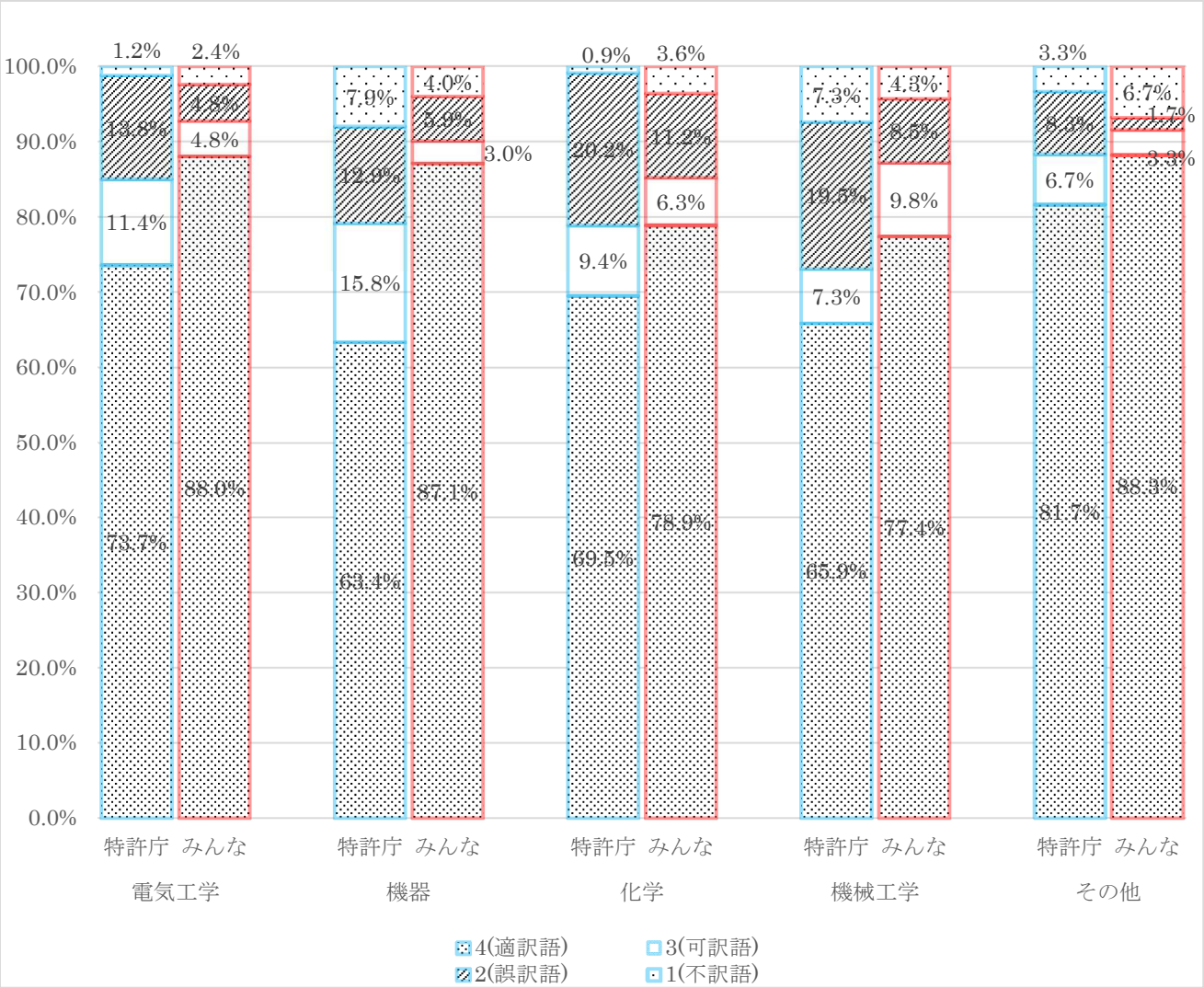


図3.2.2-2 セクタ毎の評価結果

（３）35技術分野別の評価結果

特許庁翻訳とみんなの翻訳の35技術分野別の評価結果を図3.2.2-3に示す。結果をみると、特許庁翻訳は「バイオテクノロジー」と「運輸」の平均評価点数が3.0以下で、みんなの翻訳では、3.0以下の技術分野がなかった。また、特許庁翻訳の「ビジネス方法」、「高分子化学、ポリマー」、「無機材料、冶金」、「表面加工」、「マイクロ構造、ナノテクノロジー」、「その他の特殊機械」、「家具、ゲーム」等の計7技術分野の平均評価点数がみんなの翻訳より高かった。

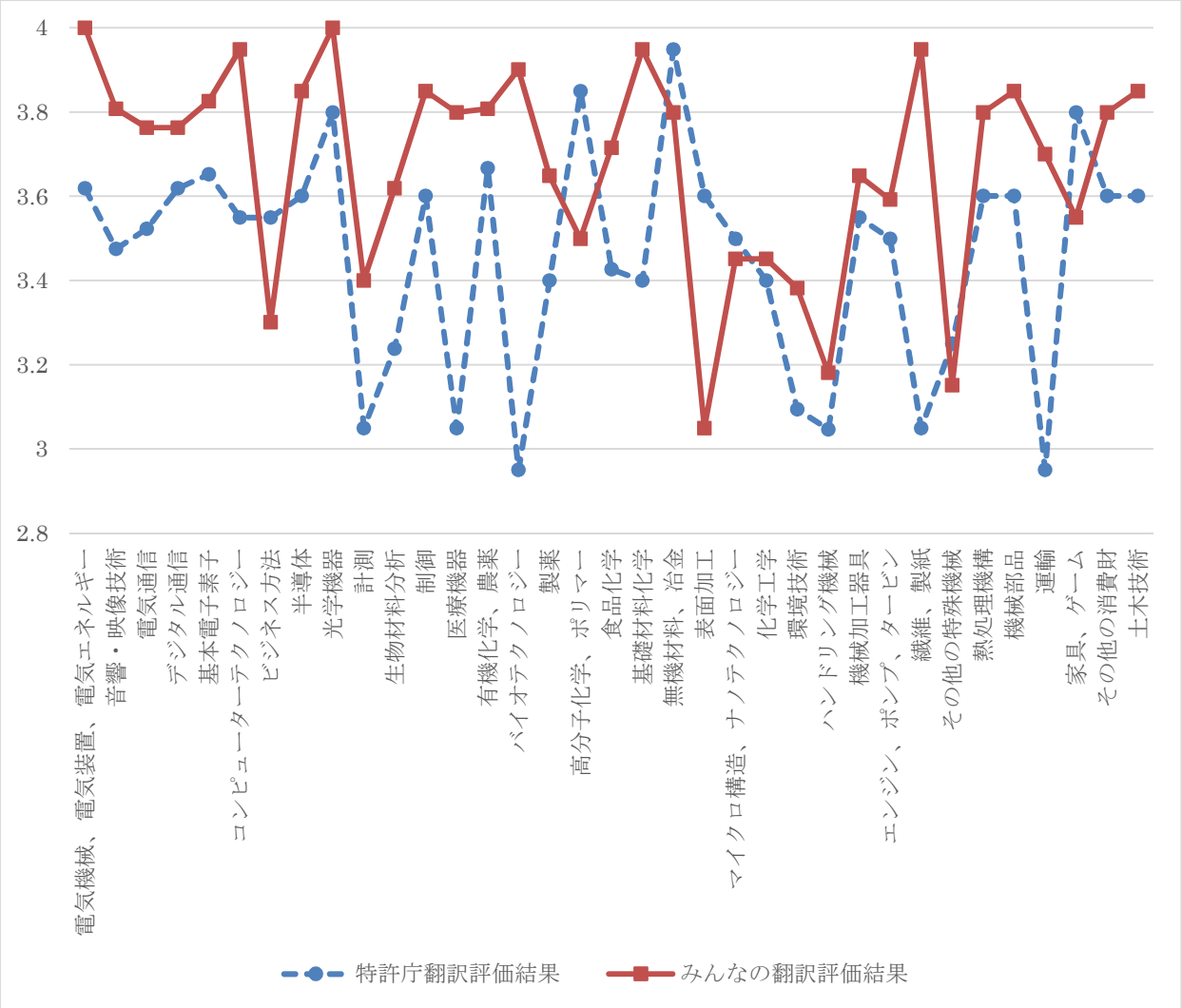


図3.2.2-3　特許庁翻訳とみんなの翻訳の35技術分野別の評価結果

3.3 特許庁翻訳結果の分析

3.3.1 重要技術用語と内容の伝達レベルの評価結果の相関関係

特許庁翻訳において、評価文700文から選択した重要技術用語の人手評価と、評価文700文の内容の伝達レベルの人手評価との間にはどの程度の相関があるか調べた。また、評価文700文のうち、重要技術用語の選択に使用された615文を使用している。

文ごとに評点の平均値をプロットする際、通常の散布図では全てのデータが0.5刻みの格子の上に乗るため、分布の様子が分かりにくい。そこで各格子点における度数を棒グラフにした。

図3.3.1-1に、700文全体の結果を示す。ピアソンの積率相関係数（以下「相関係数」）は0.1745であり、非常に弱い相関しか示さないことが分かる。内容の伝達レベルの評価に関わらず、重要技術用語については最高評価の４における度数が高い。文の翻訳精度に比較して、重要技術用語の翻訳精度が高いことが分かる。

セクタ毎（35技術分野の５大分類）の結果についても、同様の傾向であった（図3.3.1-2～図3.3.1-6）。化学については相関係数が0.34と他のセクタに比較して若干高いものの、他のセクタは相間係数が0.20以下であり、内容の伝達レベルと重要技術用語との間に強い相間は認められなかった。

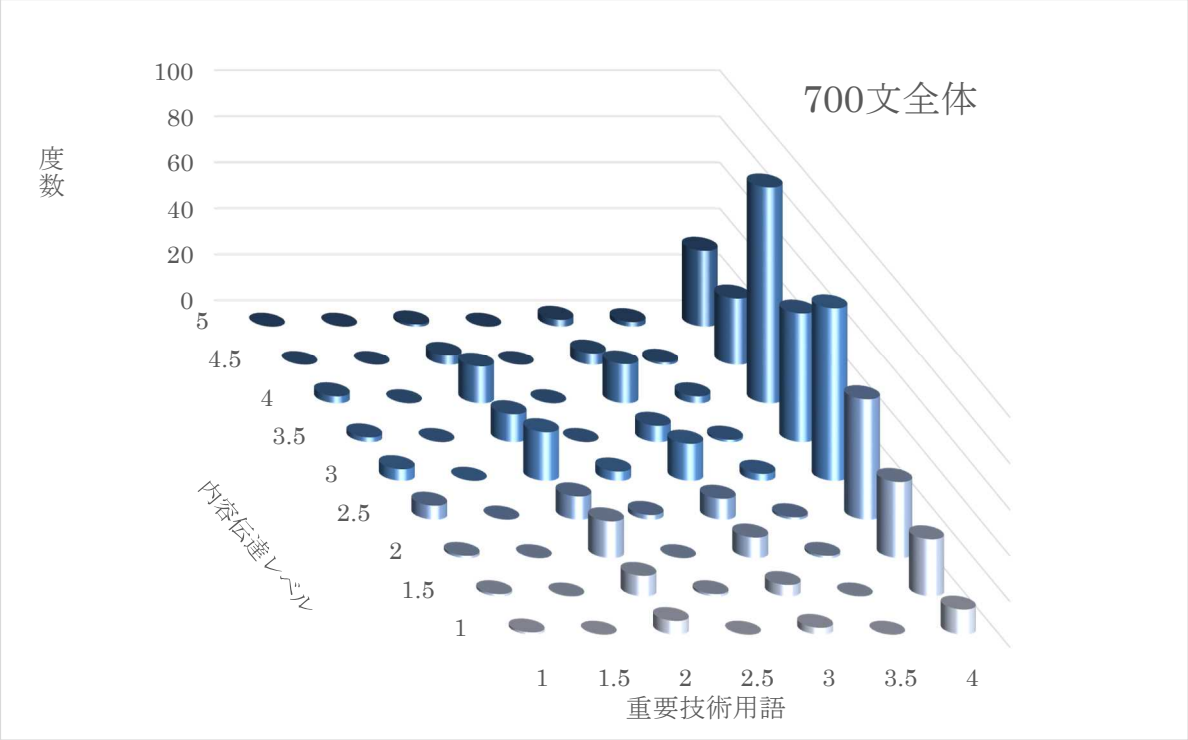


図3.3.1-1 重要技術用語と内容伝達レベルの分布 700文全体

化学については相関がある。また、**その他**についても非常に弱いが相関が認められる。これらは用語の翻訳精度を高めることで、文の翻訳精度が向上することが期待される。**電気工学・機器・機械工学**のセクタでは相関は認められなかった。

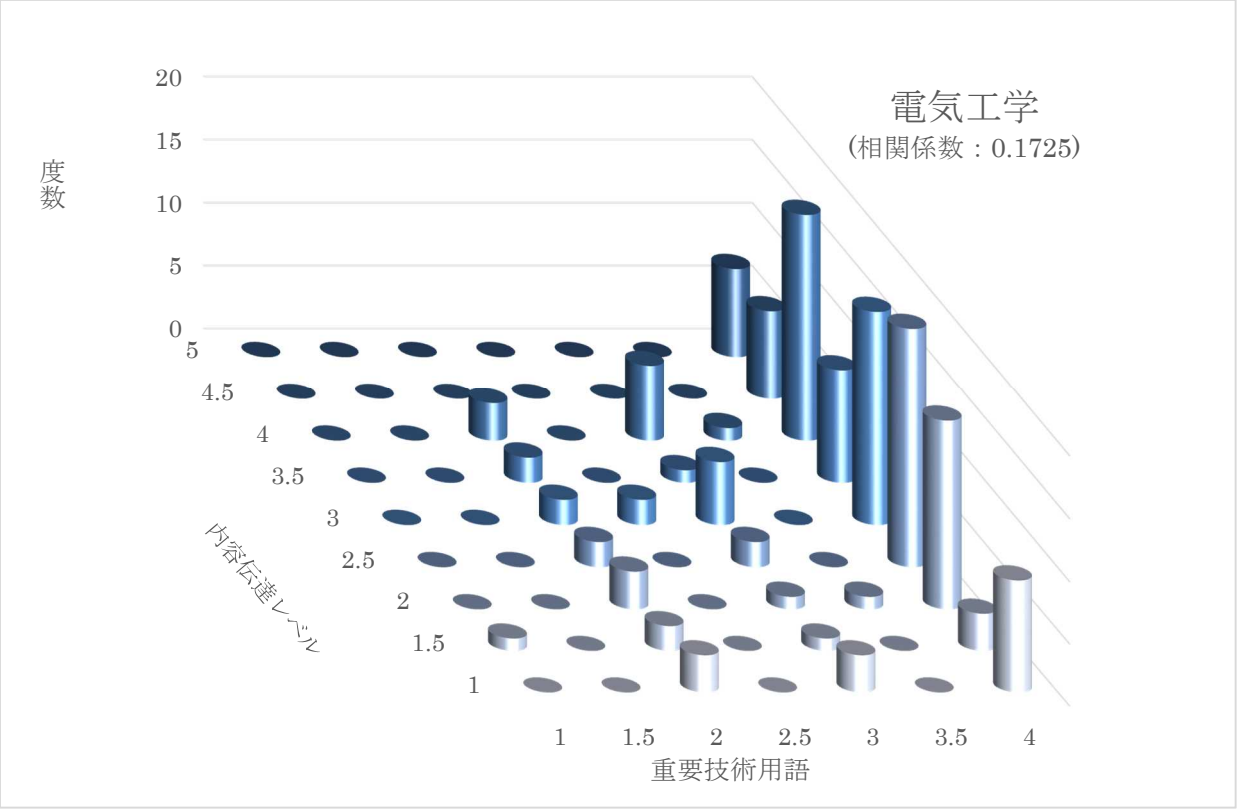


図 3.3.1-2 重要技術用語と内容伝達レベルの分布 電気工学

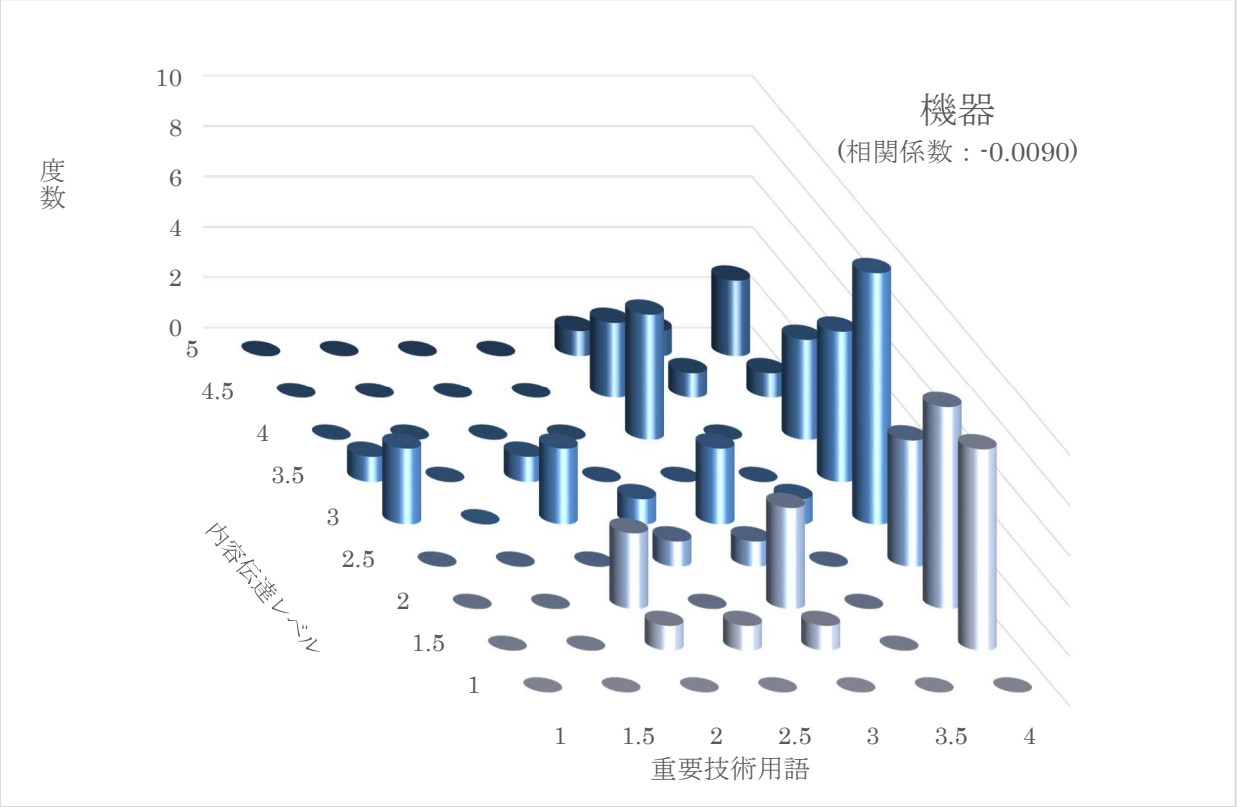


図 3.3.1-3 重要技術用語と内容伝達レベルの分布 機器

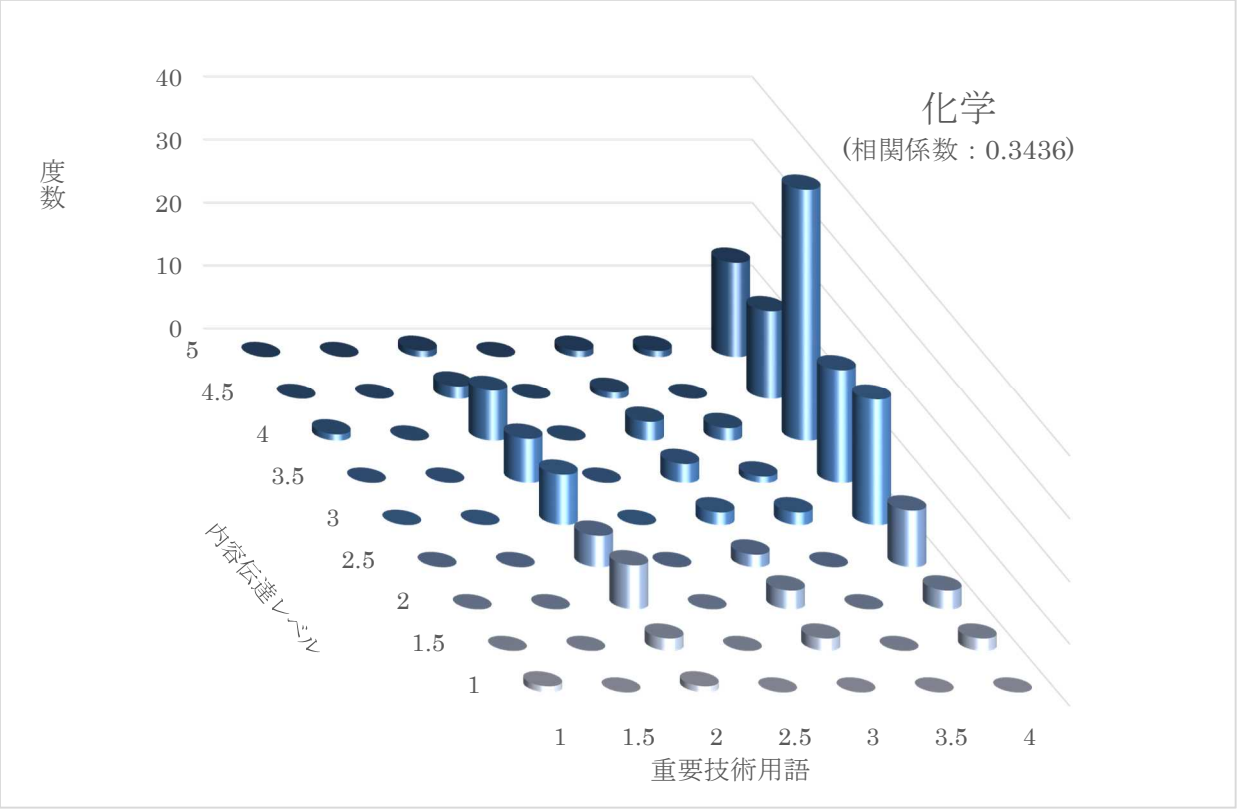


図 3.3.1-4 重要技術用語と内容伝達レベルの分布 化学

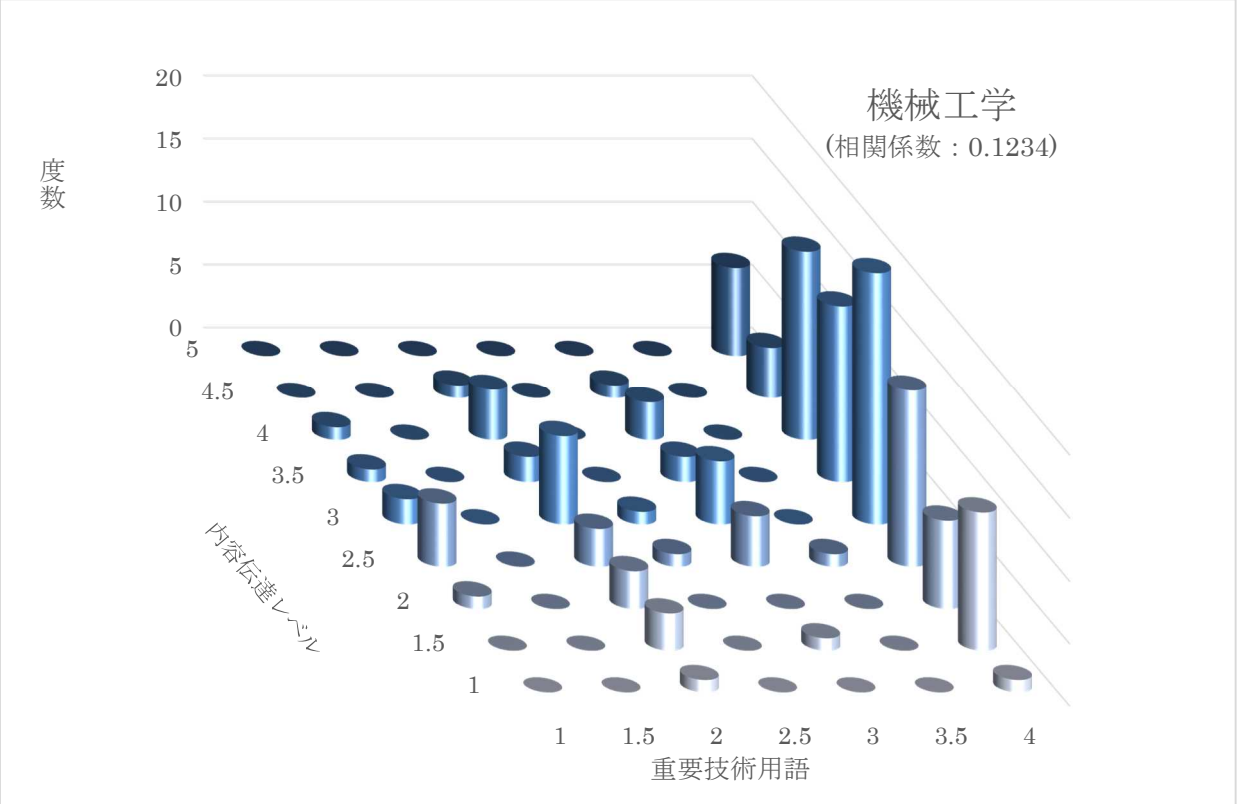


図 3.3.1-5 重要技術用語と内容伝達レベルの分布 機械工学

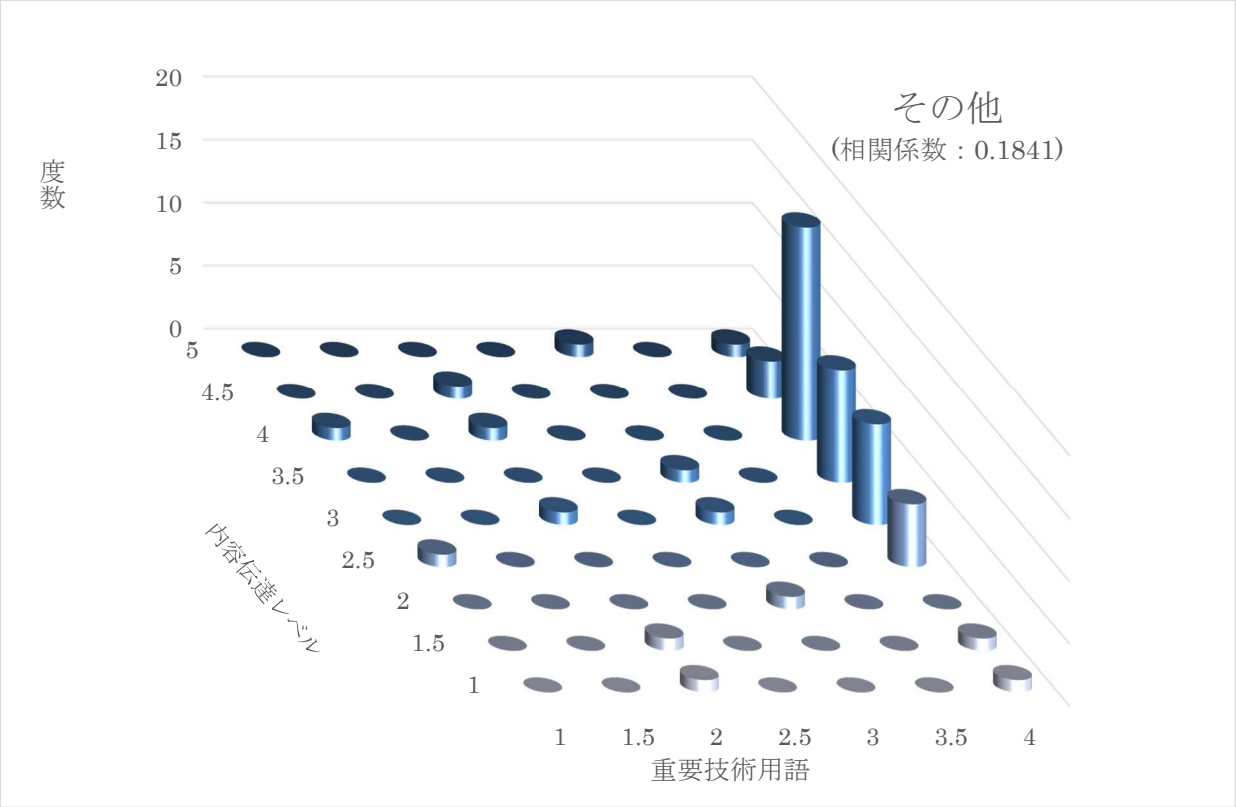


図3.3.1-6 重要技術用語と内容伝達レベルの分布 その他

相関のない場合について、重要技術用語の翻訳精度が高くても、構文その他が原因となり内容の伝達レベルが低い場合があることは十分に考えられる。一方で、作成したグラフを見ると、重要技術用語の翻訳精度が低いにも関わらず内容の伝達レベルが高い場合が見受けられる。以下に具体的な例をいくつか上げる：

表3.3.1-7 重要技術用語の翻訳精度が低く内容の伝達レベルが高い例

分野	半導体（電気工学）
原文	这种驱动电流和栅极漏电对栅介质厚度要求上的矛盾，对于传统的栅介质而言是无法回避的。
参照訳	このような駆動電流とゲートリークのゲート誘電体の厚さに対する要求における矛盾は、従来のゲート誘電体には回避できないものである。
機械翻訳	このような駆動電流とゲートは漏電してゲート誘電体の厚さの要求上の矛盾に対し、従来のゲート誘電体については回避することができない。
重要技術用語	栅极漏电：ゲートリーク

原因分析	「栅极漏电（ゲートリーク）」の技術用語を「ゲートは漏電して」に翻訳していますが、文レベルでは、大枠理解できるレベルにある。

分野	計測（機器）
原文	现有的集沙仪主要有旋风分离式，多口方口式，翻斗式等几种。
参照訳	従来の砂収集器具は主にサイクロン分離式、多重四角形口式、スキップバケット式など数種類がある。
機械翻訳	既存の集沙儀はサイクロン分離式が主にあり、マルチポート方口式、ダンプ式などの何種類か。
重要技術用語	集沙仪：砂収集器具
原因分析	「集沙仪（砂収集器具）」の技術用語を「集沙儀」に翻訳した。技術用語以外はほかの文の翻訳はよく出来ている。

分野	音響・映像技術（電気工学）
原文	因此，本发明提供一种能够有效减少假警报的移动检测方法以及使用此方法的移动检测装置。
参照訳	これによって、本発明は効率的に誤認警報を減らすことができる移動検出方法及びこの方法を用いた移動検出装置を提供する。
機械翻訳	従って、本発明は効率的縮小偽警報の移動検知方法及びこの方法を用いる移動検出手段を提供する。
重要技術用語	假警报：誤認警報
原因分析	技術用語「假警报：誤認警報」を「縮小偽警報」に誤訳しているが、意味は推測できる。

このように、重要技術用語が間違っていても、それ以外の語や構文が正しく訳され、訳文から原文の意味が推測できる場合に、内容の伝達レベルが高くなることがある。

3.3.2 誤訳原因の分析

3.3.2.1 誤訳原因の分類

特許庁翻訳の 700 文の人手評価結果から評点が 4 点以下の 623 文について、誤訳原因を検証したところ、細かく分けると計 50 の誤訳原因が観察され、7 以上の文章に観察された誤訳原因のみを 16 件に纏めた。具体的な原因とその頻度分布について、図 3.3.2.1-1 の通りである。（一の文章に同一の誤訳原因が複数回にわたり観察された場合でも、1 とカウントした。）

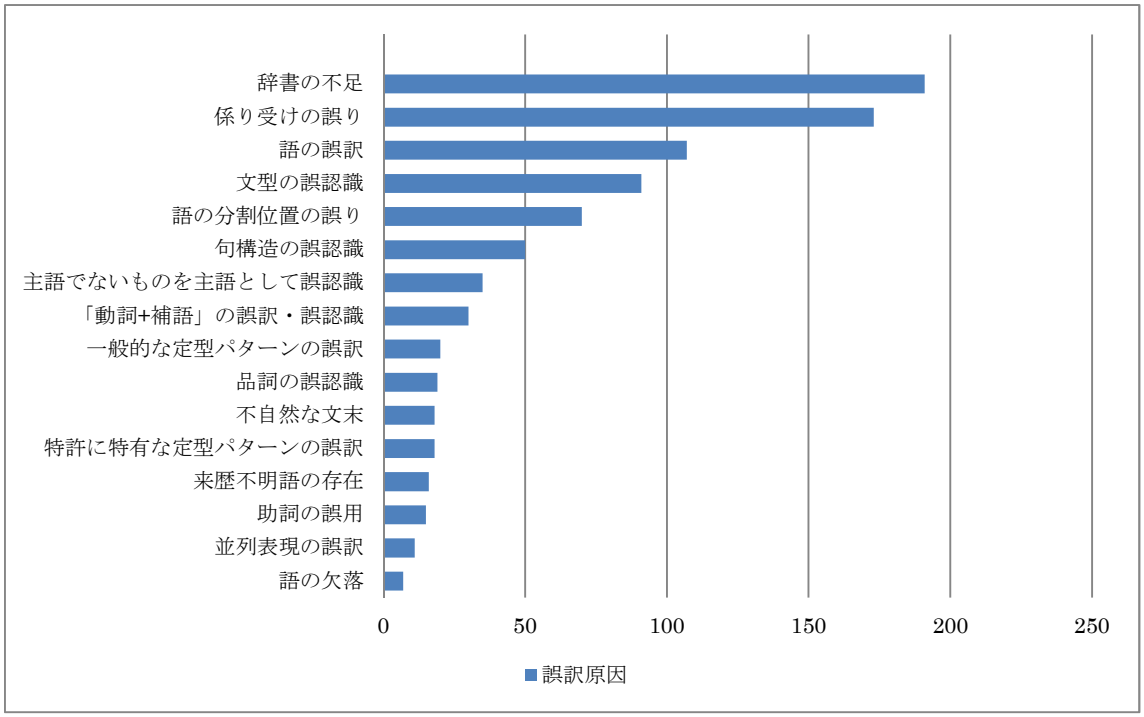


図3.3.2.1-1 全技術分野の誤訳原因分布図

図3.3.2.1-1の17パターンの誤訳原因を、その特徴に基づいてグループ化を行ったところ、下記表3.2.3.1-2の結果を得た。また、図3.3.2.1-3に、誤訳原因の大分類のグラフを示す。

表3.3.2.1-2 全技術分野の誤訳原因の分類

大分類	No	特定の原因	文章数
1. 構文解析	1-1	係り受けの誤り	173
	1-2	文型の誤認識	91
	1-3	語の分割位置の誤り	70
	1-4	句構造の誤認識	50

	1-5	「動詞+補語」の誤訳・誤認識	30
	1-6	主語でないものを主語として誤認識	35
	1-7	並列表現の誤訳	11
		小計①	460
		比率	53%
大分類	No	特定の原因	文章数
2. 辞書	2-1	辞書の不足	191
	2-2	語の誤訳	107
	2-3	助詞の誤用	15
	2-4	品詞の誤認識	19
		小計②	332
		比率	38%
大分類	No	特定の原因	文章数
3. 定型パターン	3-1	一般的な定型パターンの誤訳	20
	3-2	特許に特有な定型パターンの誤訳	18
		小計③	38
		比率	4%
大分類	No	特定の原因	文章数
4. その他	4-1	不自然な文末	18
	4-2	来歴不明語の存在	16
	4-3	語の欠落	7
		小計④	41
		比率	5%
		合計	871

次に表3.3.2.1-2の大分類の分布を図3.3.2.1-3のようにグラフ化する。

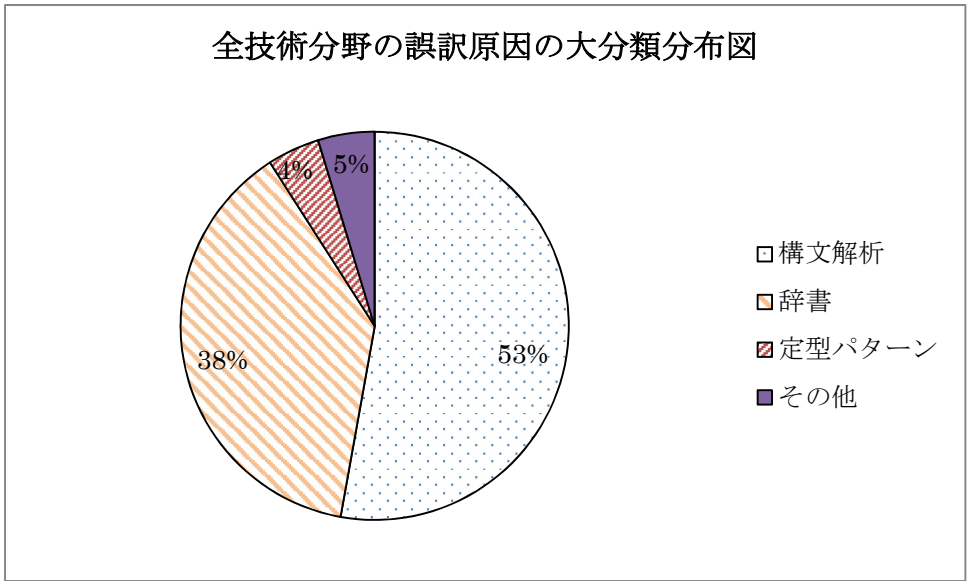


図3.3.2.1-3 全技術分野の誤訳原因の大分類分布図

図3.3.2.1-3から分かるように、構文解析に起因する誤訳が53%、辞書に起因する誤訳が38%、合計で91%を占める結果となった。ただし、これらの誤訳原因は4大分類の中でも相互に関連性があることは想像に難くない。

具体的には構文解析のNo.1-1、No.1-2、No.1-4、No.1-7に挙げた「係り受けの誤り」、「文型の誤認識」、「句構造の誤認識」、「並列表現の誤訳」は、類似する文型パターンが散見される特許文献においては定型パターンが翻訳精度の向上に貢献すると推察される。

また、同じくNo.1-3「語の分割位置の誤り」には複合語の分割位置を正しく判断できないことに起因する誤訳も含まれるため、大分類に示す辞書のNo.2-1、No.2-4に挙げた「辞書の不足」、「品詞の誤認識」が影響していると言える。

3.3.2.2 誤訳原因の検証

各誤訳原因について、文例を挙げて説明する。

1. 構文解析による誤訳

1-1. 係り受けの誤り

原文	本发明的目的是提供一种双侧定位振子，保证振子在使用过程中不会发生脱落或移位的・象。
参照訳	本発明の目的は両側で位置決めするオシレータを提供し、オシレータの使用中に脱落又は転位現象が発生しないことを保証することにある。
機械翻訳	本発明の目的は両側位置決め振動子を提供する、使用プロセスに振動子を保証して可能性がなく脱落あるいはビットシフトの現象が発生する。
分析	人手翻訳では「保証する」という動詞が「オシレータの」に直接的に係り、さらに「使用中に」と「脱落又は転位現象が発生しないことを」に並列かつ間接的に係っていることがわかる。 それに対し、特許庁翻訳では原文中の「保証する」に相当する動詞が、その直後に配置されている「振動子を」だけに係っており、それ以降の「可能性がなく」にも「脱落あるいはビットシフトの現象が発生する。」にも係っておらず、それぞれが脈絡のない独立した句や節として扱われていることが判明した。 改善策としては、中国語と日本語の語順の違いを考慮に入れて構文解析を行うことにより、句や節単位で前から順番に翻訳したのちに結合する、いわば逐次翻訳に近い手法から脱却することにより改善できると考える。

1-2. 文型の誤認識

原文	因此，如果要达到更高的・・精度，需要・石英晶体振・器的・率偏差・行温度・・。
参照訳	したがって、時計の更に高い精度を達成しようとすると、水晶発信器の周波数偏差に対して、温度補償を行わなければならない。
機械翻訳	従って、もしもっと高いクロック精度に達すること必要ならば、水晶結晶発振器の周波数偏差に温度監視して備えていた必要がある。
分析	「钟精度，需要対」に見られるように「，」で区切られた文節単位で翻訳せずに、文節をまたいで翻訳していることに誤訳の大きな一因があると考えられる。翻訳精度を向上させるためには根本的に必要な処置でありながら、比較的に改善が容易であると推測されるため、優先して善処すべき課題である。ただし、特許庁翻訳の訳文に見られるように「需要」の訳である「必要」という語が2度にわたって用いられており、係り受けに関する

	構文解析も見直す必要がある。

1-3. 語の分割位置の誤り

語の分割位置の誤りは、単語単位、句単位、文節単位のそれぞれにおいて頻出する課題である。また、単語単位での分割位置の誤りには複合語の分割位置の誤りも含まれ、さらには一の単語を複合語であるかのように分割する例と、その反対に一の複合語を複数の単語に分割する例が存在するため、詳細に検証することが必要である。

1-3-1. 単語単位での分割位置の誤り（一の複合語を複数の単語に分割する例）

原文	本・明的・施方式涉及火工技・・域，更具体地，本・明的・施方式涉及一种模・・筒及用其・・爆破・・・装置・火可靠性的方法。
参照訳	本発明の実施方式は火工技術分野に関し、具体的には、本発明の実施方式は模擬薬莢及びそれを用いて爆破弁駆動装置の発火信頼性を検出するための方法に関する。
機械翻訳	本発明の実施形態は火仕事技術分野に関し、更に具体的に、本発明の実施形態は薬莢およびそれで破裂弁駆動装置伝火信頼性を検出する方法を模擬することに関する。
分析	「模・・筒」は「模擬薬莢」と一語に訳されるべきであるのに対し、特許庁翻訳では「模擬」と「薬莢」に分割して訳されている。 さらに、「模擬」が文末で「方法を模擬することに関する。」と誤った係り受けを招いており、人手翻訳の「方法に関する。」とは異なる結果となった。 専門用語を正確に訳せていないという事実はありながらも、前半部での構文解析は正しく行われており、一の複合語を二の単語に誤って分割したために文全体の意味が通じなくなった典型的な例と言える。 前述したとおり、構文解析に起因する誤訳は辞書の不足と密接な関係があることが多く、当該事例においても「模・・筒」を「模擬薬莢」と辞書登録することにより文全体の理解度を高める可能性がある。

1-3-2. 単語単位での分割位置の誤り（複数の単語を一の単語として誤認識する例）

上記1-3-1と同じ中国語原文を用いて、上記の例とは反対に複数の単語に分割して翻訳されることが求められる句を、一の単語として解析した例を挙げる。

「爆破・・・装置・火可靠性」は「爆破弁駆動装置の発火信頼性」と分割されることが求められるが、「破裂弁駆動装置伝火信頼性」と訳されている。

原文において「爆破・・・装置」と「・火可靠性」の間に「的」が挿入されていれば両単語を「の」で接続して正しく翻訳できた可能性は否めないため、試験的に「的」を挿入

して翻訳することは検証に値する。

前述の「模擬薬莢」の辞書登録とあわせて対処することで、この文献の評点は現行の3から4に引き上げられると考察する。

1-3-3. 句単位での分割位置の誤り

原文	・・管道式・火装置，罐体・置感・装置，当罐体内・火材料不足・・出提示信息。
参照訳	作業場における配管式の消火装置であって、タンクに検知装置が設けられ、タンク内の消火材料が不足する時に提示メッセージを発信する。
機械翻訳	作業場ダクト式消化装置は、缶体は装置に感応することを設置し、缶体内に材料不足を鎮火する時示唆情報を出す。
分析	この例では、「・火材料不足」という原文を「・火材料＋不足」と分割して「消火材料が不足する」と訳すべきところを「・火＋材料不足」と分割した結果、「材料不足を鎮火する」と誤訳している。 この例のように、動詞としても名詞としても機能する単語が含まれる複合名詞、句や節の構文解析に失敗することが多く見られるため、特許庁翻訳は重点的に対応することが推奨される。

1-3-4. 節単位での分割位置の誤り

原文	2.根据权利要求1所述的电多层组件，其中开口（8）用半导体材料填充，所述半导体材料包括压敏电阻陶瓷或电阻材料。
参照訳	2、請求項１に記載の電気多層コンポーネントにおいて、前記開口（８）に前記半導体材料を充填、該半導体材料はバリスタセラミックス又は抵抗材料を含む。
機械翻訳	２．以下を特徴とする請求項１に記載の電気多層アセンブリ：開口（８）用半導体材料は充填し、前記半導体材料は感圧抵抗セラミックあるいは抵抗材を含む。
分析	「开口（8）用半导体材料」を一の名詞と解析してしまい、「場所＋動詞+名詞」の構造に基づいて「开口（8）＋用＋半导体材料」と解析することができていない。その結果、本来は目的語である名詞の「半導体材料」が主語として働いたことにより、文前半の意を損ねている。 さらに、「用」は動詞としての働きのほか、「用＋A＋B」で「AでBする」の「で」に相当する介詞としての役割も持ちあわせているため、この文献の「用＋半导体材料＋填充」に限らず複数の解釈ができる可能性を秘めている。したがって、特許庁翻訳は頻出単語である「用」の解析と用法につ

	いて、細心の注意を払って処理することが肝要である。

1-4. 句構造の誤認識

原文	由此，通过采用齿轮传动，避免了驱动器直接作用于第二传动组件，起到了缓冲的作用，保证了传动的平稳性，且提高了驱动器的使用寿命。
参照訳	したがって、歯車伝動を用いることにより、駆動装置が第二伝動アセンブリに直接作用することを回避し、緩衝の作用を果たし、伝動の安定性を保証し、かつ駆動装置の耐用年数を延ばす。
機械翻訳	これにより、歯車伝動を用いることによって、第２の駆動アセンブリにドライバ直接作用を避けて、緩衝した作用を果たして、伝動した定常性を保証して、且つドライバの実用寿命を高めた。
分析	「起到了～作用」は「～の（な）作用を果たす」という常套句であり、「～」に挿入される語を名詞または形容詞として訳すことが多いため、特許庁翻訳に見られる「緩衝した作用を果たして」ではなく、「緩衝の作用を果たし」とすることが適切である。また、「保证了传动的平稳性」に関しても、「平稳性」に係っていることから判断すると、「传动」が動詞ではなく名詞として用いられていると句構造を解析することが正しい。句構造の解析については統計翻訳の要素により改善できる可能性があると考える。

1-5. 「動詞+補語」の誤訳・誤認識

原文	3、能够将从光栅层、药膜层射下来的光发射回去，从而提高相片的亮度和三维效果。
参照訳	３、回折格子層、フィルム層から射出した光線を反射することで、写真の輝度と立体効果を向上させることができる。
機械翻訳	３、グレーティング層、薬膜層から光放射を撃ち落として帰り、これによって写真の輝度と三次元効果を高める。
分析	「V+方向補語」の構造である”射下来”、”发射回去”を「光放射を撃ち落として帰り」のように一般動詞のように訳した結果、原文とはかけ離れた訳文となってしまった。

1-6. 主語でないものを主語として誤認識

原文	现有的无线通讯技术大多是基于电磁场，即通过终端设备发出的电磁场来传输信息，对终端设备均有较高的硬件要求。

参照訳	従来の無線通信技術は多くが電磁界に基づき、即ち端末装置から送信した電磁界によって情報を伝送するため、端末装置のハードウェアに対して、いずれも比較的厳しく要求する。
機械翻訳	既存のワイヤレス通信技術は大部分電磁場に基づき、すなわち端末機器が出す電磁場によって情報を転送し、端末機器に全部比較的高いハードウェアが要求する。
分析	「硬件（ハードウェア）」は主語でないが主語として訳し、「ハードウェアが要求する」と訳文になった。

1-7．並列表現の誤訳

原文	因此，以往的技术中，难以在不使耐弯曲疲劳性降低的情况下防止界面剥离并提高耐久性。
参照訳	そのため、従来の技術では、耐屈曲疲労性を低下させずに、界面剥離を防止して耐久性を向上させるのは困難であった。
機械翻訳	従って、従来の技術に耐屈曲疲労性を低下させない場合界面剥離が且つ高耐久性を提げるのを防止しがたい。
分析	“并（及び）”は“防止界面剥离（界面剥離を防止）”と“提高耐久性（耐久性を向上）”という並列関係の二つの「句」を結んでいる。この構造が正しく理解されていないため誤訳となっている。

2．辞書による誤訳

2-1．辞書の不足

続いて、大分類の「辞書」で誤訳原因としてトップに挙げられた「辞書の不足」について、文献番号「104174467」を用いて精査する。

原文	本发明的目的在于根据塔磨机底部的形状和结构设计出一种塔磨机机壳底部的耐磨构造，解决塔磨机机壳底部的衬胶易受磨损的问题。
参照訳	本発明の目的はタワーミル底部の形状や構造に応じてタワーミルケース底部用の耐摩耗構造を設計し、タワーミルケース底部のゴムライニングが摩耗しやすいという問題を解決することにある。
機械翻訳	本発明の目的は塔グラインダーの底による形状とアーキテクチャ設計が塔グラインダーケーシングの底を出す摩擦に強い構造にあり、塔グラインダーケーシングの底のゴム張りを解決して摩耗の問題を易受する。
分析	特許庁翻訳では「塔磨机」を「塔グラインダー」、「结构」を「アーキテクチャ」、「衬胶」を「ゴム張り」と訳していることから、専門用語に対応できていないと言える。

	「结构」に関して、本文献の分野である23_化学工学においては「構造」が適切な訳語であり、「アーキテクチャ」は必ずしも誤りとは言えないものの、むしろ情報技術分野の専門用語訳として用いられるケースが多いと考えられる。 誤りがより顕著な例としては、「係り受けの誤り」で言及した上述の文献番号「104226575」は23_化学工学の分野であるために、中国語原文中の「移位」を「転位」と訳すことが求められる一方、特許庁翻訳は「ビットシフト」という情報技術分野の専門用語に訳している。 同様に、対象文献の分野とは異なる専門用語辞書を参照している例として、文献番号「104177131」は20_無機材料、冶金の分野でありながらも「瓷胎」を「磁器胎児」、「斑釉」を「斑状歯」と訳していることから、何らかの理由により医学分野の専門用語辞書が適用されているとの印象を禁じ得ない。 本入札案件のために特許庁から貸与された220万語を超える専門用語辞書は4分野に分類されているのみであり、市販されている特許文献に特化した翻訳ソフトウェアでは専門分野が30ほどに細分されている点、さらに一時に指定できる専門用語辞書の数量に制限を設けている点とは大きく異なる。市販されている製品では翻訳精度の低下を防ぐためにこのような処置をとっていることが多く、特許庁翻訳も参考に値すると考える。 特許文献の場合はその技術分野があらかじめ明確に記されているため、より精度の高い機械翻訳を実行するためには、対象となる文献の技術分野に特化した専門用語辞書を限定して翻訳するという、細やかな操作が求められる。

次に、もう一つの辞書の不足による誤訳の例を説明する。

原文	用水清洗猪脚一遍后放入到木桶中，加入10克食盐后浸泡7个小时，然后用水清洗干净并将表面的水沥干。
参照訳	水で豚足を一回洗浄した後に桶に入れ、10グラムの食塩を加えた後に7時間浸漬し、続いて水で洗浄しかつ表面の水を切る。
機械翻訳	水パージ豚足で一度後に桶中に到達することに入れて、１０の消化を助け塩後に７時間浸漬することを添加して、その後水で且つ表面水瀝乾をきれいに洗う。
分析	「10克食盐」を「10グラムの食塩」と訳すことができていない点は大いに改善の余地がある。完全な解決策とは言えないものの、解決に向けて下記の方法を提案する。辞書に「食盐」の訳語として「食塩」が登録されているかを確認する。量詞である「克」の訳語として「グラム」が登録されて

	いるかを確認し、登録されている場合は「数詞＋克」を正しく翻訳できるかの確認を行う。そのうえで、「数詞＋克＋食盐」を正しく翻訳できるかの確認を行うことにより、誤訳原因の細分化を通じた究明を行うことができる。さらに、「克」と同様に重さの単位である「斤」、「两」についても数詞や「食盐」と組み合わせて翻訳結果を検証する。

2-2. 語の誤訳

原文	同时，由于当前的评估体系不能结合用户除点击以外的网络行为数据进行评估，导致广告投放和广告呈现较难与真正的受众偏好相匹配。
参照訳	また、現在の評価システムはユーザのクリック以外のネットワーク行為データを結合して評価を行うことができないため、広告の投入と広告の表示が受け手の本当の好みに合いにくいことを引き起こす。
機械翻訳	同時に、現在のカワバタ評価システムによってユーザクリック以外のネットワーク行動データと結合することができないことは評価し、広告配置と広告提示を比較的に本当の観衆と一致することを選好するのが難しくする。
原因分析	原文の「评估体系」の日本語訳は本来「評価システム」が正しい。 特許庁の 220 万語辞書には「评估体系」の見出し語について、下記二つの訳が登録されている。 ・评估体系　　カワバタ評価システム ・评估体系　　評価システム 「評価システム」と正しい単語も登録されているが、辞書引きする際、良くない単語を選んでしまい、翻訳品質に影響を及ぼすことがある。 特に同じ見出し語について、複数の訳語が登録されている場合、もっとも相応しい訳語を選ばないと翻訳の品質に影響を及ぼす場合がある。

2-3. 助詞の誤用

原文	优选的是，所述声控感应器监测到宝宝哭声，控制机械手自动开启进入工作状态，推动摇篮晃动。
参照訳	最適化したのは：前記の音声コントロールセンサは赤ちゃんの泣き声を監視測定し、マニピュレーターの動作モードを自動的に起動させ、揺りかごを揺り動かす。
機械翻訳	好ましくは、前記音声コントロールインダクタは赤ちゃんの泣き声にモニタリングし、制御機の手は自動で進入作動状態を開封し、ゆりかごのガタツキを推進する。

分析	「赤ちゃんの泣き声にモニタリングし」の「に」は使用誤りである。

2-4. 品詞の誤認識

原文	目标表达的实例包含拼写某一词的字符序列、详细说明基因的遗传碱基对序列、形成图像的一部分的图片或视频文件中的位序列、形成程序的一部分的可执行文件中的位序列或形成歌曲或口语短语的一部分的音频文件中的位序列。
参照訳	対象表現の実例は、ある単語を綴る文字配列を含み、遺伝子を特定する遺伝子塩基対配列、画像の一部を形成する画像ファイルもしくはビデオファイルにおけるビット配列、プログラムの一部を形成する実行可能ファイルにおけるビット配列、または、歌もしくは話された言葉の一部を形成するオーディオファイルにおけるビット配列を詳しく説明する。
機械翻訳	目標表現した実例はいずれかの語を表音文字で表記する文字シーケンスを含み、遺伝子の遺伝塩基対配列、画像形成一部分の画像あるいはビデオファイル中のビットシーケンス、形成プログラムの一部分の実行可能ファイル中のビットシーケンスあるいは形成の歌あるいは口語語句の一部分のオーディオファイル中のビットシーケンスを詳細説明する。
分析	本文の中の”形成”は「形成する」という動詞としての用法であるが、これを名詞のように翻訳した結果、訳文の正確性を大きくそいでいる。

3. 定型パターン

3-1. 一般的な定型パターンの誤訳

原文	本发明涉及一种有机物处理装置和有机物处理系统，确切的说是提供一种能够处理有机物餐厨垃圾的食物垃圾无害化分解回收再利用装置。
参照訳	本発明は有機物処理装置と有機物処理システムに関し、正確に言えば有機物キッチンゴミを処理できる食物ゴミ無害化分解回収リサイクル装置を提供する。
機械翻訳	本発明は有機物処理装置と有機物処理システムに関し、適確であることは有機物キッチンゴミを処理することができる食べ物残り生屑を提供すると言えて再利用装置を回収することを分解することを示する。
分析	慣用句的な言い方の”确切的说（正確に言えば）”が適切に翻訳されていない。

3-2．特許に特有な定型パターンの誤訳

原文	为了解决以上问题，本发明提供了一种具有气凝胶隔热层的新型卷烟加热器，其具有隔热保护、降低热能损失的作用。
参照訳	上記課題を解決するため、本発明は熱損失を低減し、熱保護効果があるエーロゲル絶縁層を有する新規巻きたばこ加熱器を提供する。
機械翻訳	解決するため上問題で、本発明はエアロゲル熱遮断層を有する新型タバコ加熱器を提供した、それは断熱保護、低下の熱エネルギー損失の作用を有する。
分析	“为了解决以上问题（上記課題を解決するため）”の特許に特有な定型パターンの翻訳がよく翻訳されていない。

4．その他

4-1．不自然な文末

原文	更优选方案中，上述输送机构还包括支座和前压辊架，支座固定于升降台上，前压辊架一端与支座可转动连接，前压辊可转动安装在前压辊架另一端。
参照訳	さらに優先選択の解決手段において、上記輸送機構はまたブラケットと前部プレスローラフレームを含み、ブラケットは昇降台に固定され、前部プレスローラフレーム片端はブラケットに回転可能に接続され、前部プレスローラは回転可能に前部プレスローラフレームの他端に取り付けられる。
機械翻訳	さらに優先傾向中に、上記の搬送機構はまたキャリと前加圧ローラフレームを含み、キャリは昇降台上に固定し、前加圧ローラフレーム一端とキャリは回転連結することができて、前加圧ローラフレーム他端回転取付に前加圧ローラは。
分析	文が中途半端な終わり方となっており、正しい意味が伝わらなくなっている。

4-2．来歴不明語の存在

原文	设计优势，在设计中将传统的电极吸附改为磁场转向，这其中的优势相当明显，即，在使用过程中，永磁体不会因为其中出现粉尘而出现电阻，即其除尘强度不会随着除尘时间而减弱。
参照訳	設計上の利点は以下のとおりであり、設計では従来の電極吸着を磁気転向に変え、その優位性は明らかであり、即ち、使用中に、永久磁石は粉塵により電気抵抗が生じることがなく、即ち除塵効果は除塵時間の経過に伴って弱まらない。

機械翻訳	優勢を設計し、設計に従来電極吸着磁界に変えることを操舵し、このその中の優勢はかなり明らかであり、すなわち、使用プロセスに、永久磁石は粉塵がそのうち現われることとするので抵抗が現われる可能性がなく、すなわちその除塵強度は除塵時間についていってい弱まる可能性がない。
分析	来歴不明語”操舵”が訳文に混入している

4-3．語の欠落

原文	所述滤波腔体为圆锥形腔体也可以采用其它形状的腔体，滤波腔体与通风孔一一对应，滤波腔体与通风孔同轴且相通，位于通风孔的另一端。
参照訳	前記のフィルタキャビティは円錐形かその他形状のキャビティで、フィルタキャビティは通風口と一対一で対応して、通風口と同軸で連通して、通風口の他方の端部に位置する。
機械翻訳	前記フィルタキャビティは円錐形キャビティのために同じく採用することができる他の形状のキャビティは、フィルタするキャビティと通風孔全単射は、フィルタキャビティと通風孔インラインと連通は、通風孔の他端に位置する。
分析	“一一对应”（一対一で対応）が訳されていない。

3.3.3 主要誤訳原因の評価

前節で明らかになった上位五つの主要誤訳原因は「辞書の不足」、「係り受けの誤り」、「語の誤訳」、「文型の誤認識」、「語の分割位置の誤り」である。主要誤訳原因の個数・平均点分布と点数分布について、図3.3.3-1、図3.3.3-2に示す。

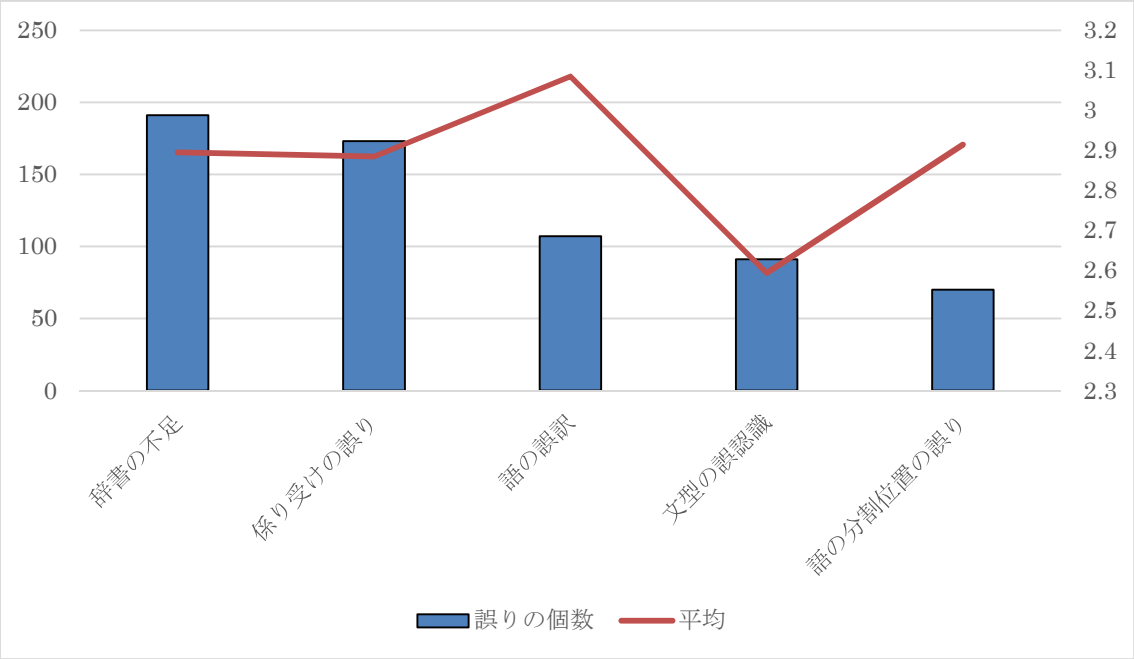


図3.3.3-1 主要誤訳原因の個数・平均点分布

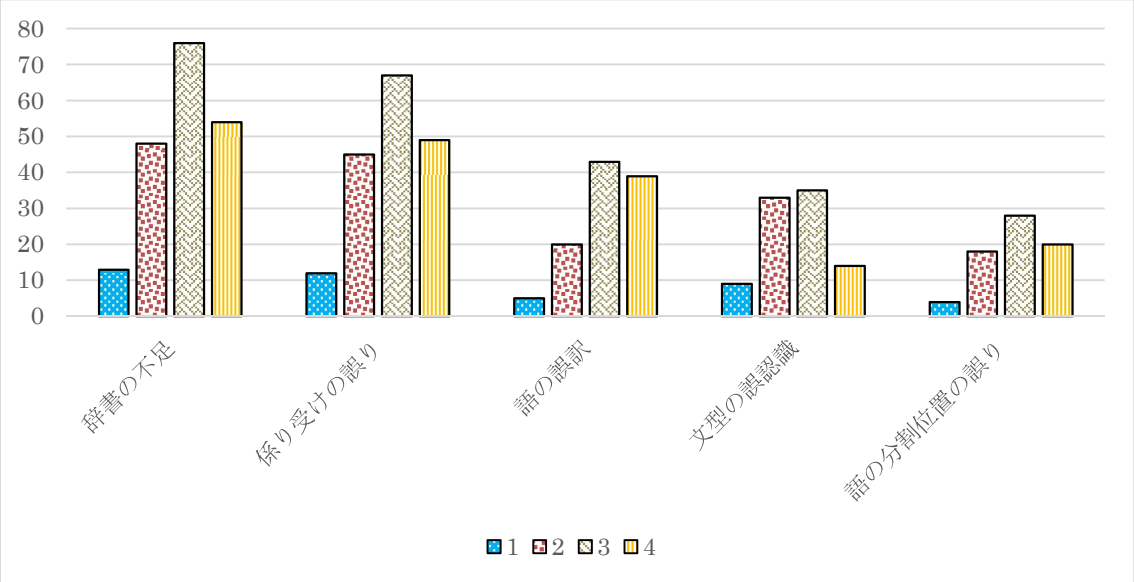


図3.3.3-2 主要誤訳原因の点数分布

図3.3.3-1から「辞書の不足」、「係り受けの誤り」の件数が一番多く、図3.3.3-2から全ての誤訳原因について、評点3の件数が一番多いことが分かる。また、「辞書の不足」、「係

り受けの誤り」、「語の分割位置の誤り」については、平均評点が2.90程度である一方、「語の誤訳」の平均評点は3.08、「文型の誤認識」の平均評点は2.59であった。これは、内容伝達レベルに与える「語の誤訳」の影響は比較的小さく、「文型の誤認識」が与える影響が大きいことを意味する。また、これらの誤訳原因について、セクタ毎で平均評点を算出したところ（図3.3.3-3～図3.3.3-7）、「機械工学」及び「その他」のセクタと比較して、「電気工学」、「機器」及び「化学」の分野は、誤訳原因の有無による差が大きくなる傾向が見られた。

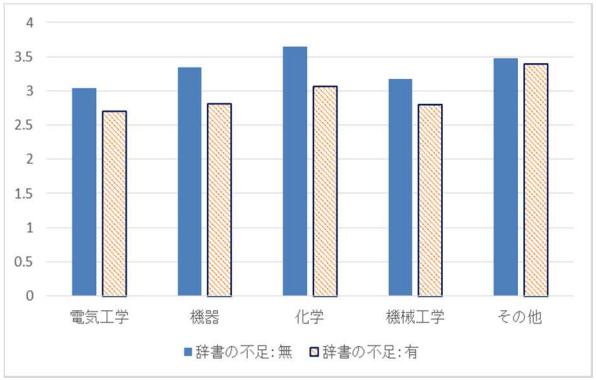


図3.3.3-3 誤訳原因：辞書の不足

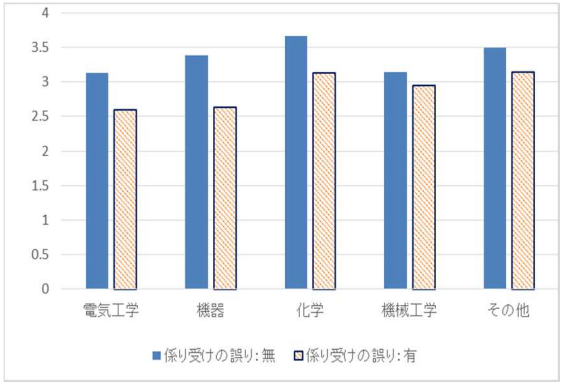


図3.3.3-4 誤訳原因：係り受けの誤り

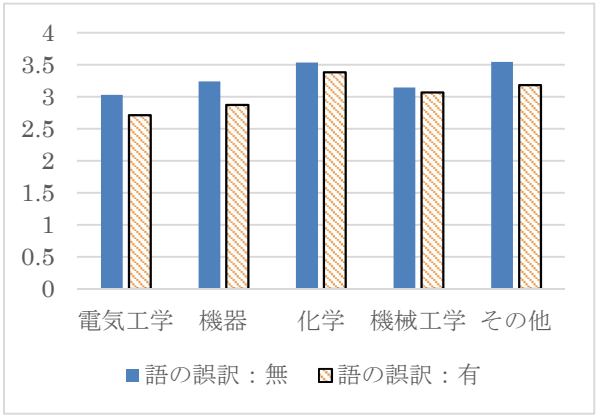


図3.3.3-5 誤訳原因：語の誤訳

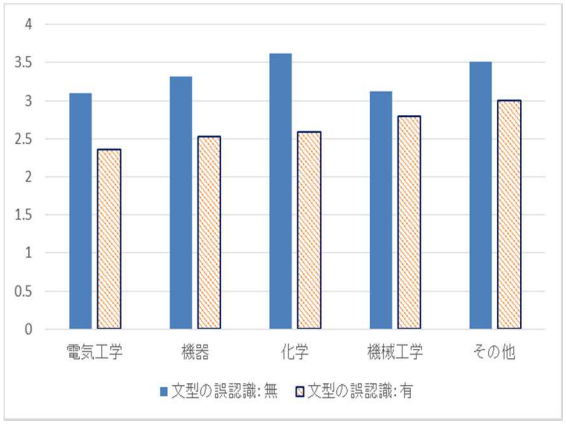


図3.3.3-6 誤訳原因：文型の誤認識

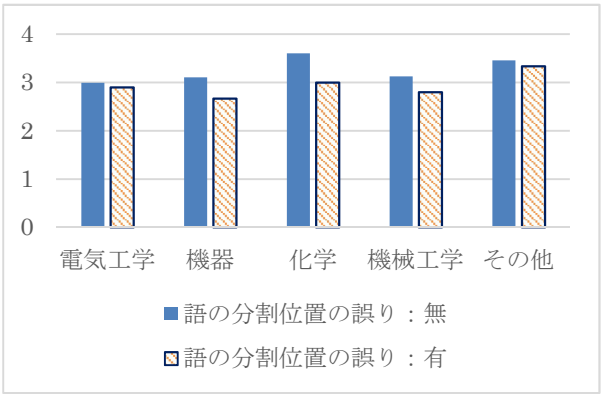


図3.3.3-7 誤訳原因：語の分割位置の誤り

3.3.4 セクタ毎の誤訳原因

セクタ毎の誤訳原因を図3.3.4-1に示す。3.2.1節でセクタ毎では電気工学の翻訳精度評価が一番低く、化学の翻訳精度評価が一番高かった。一方、図3.3.4-2から、化学と電気工学は「辞書の不足」の差が大きいことが分かる。従って、電気工学の訳質が低い大きな原因の一つとして、辞書の不足が要因となっていることが考えられる。

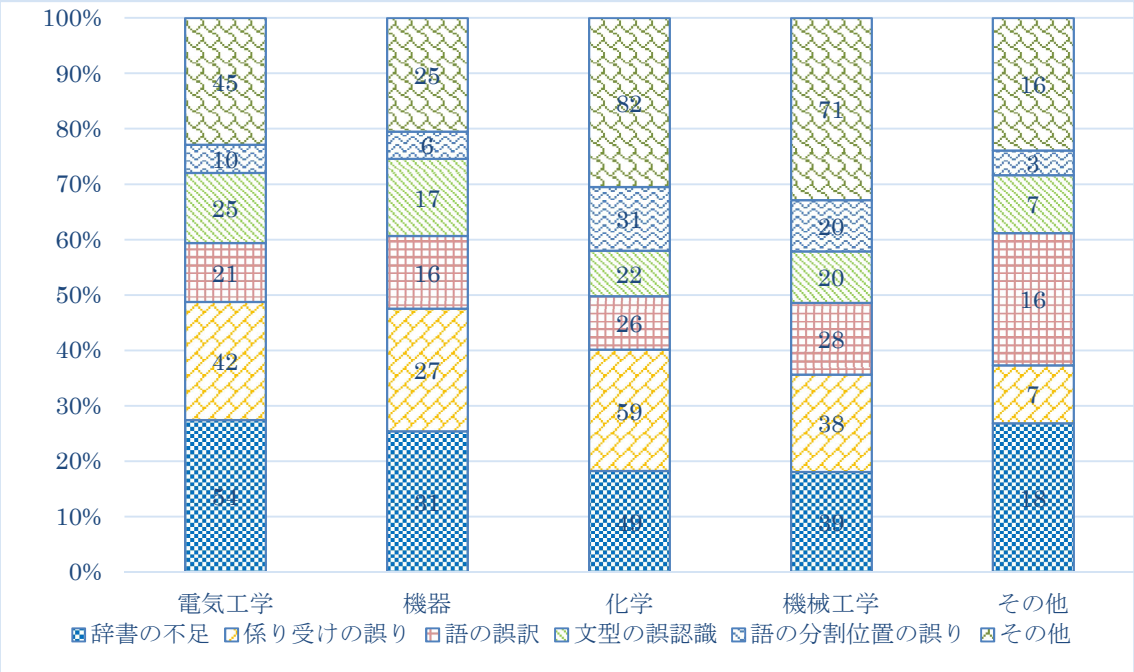


図3.3.4-1 セクタ毎の誤訳原因分布図

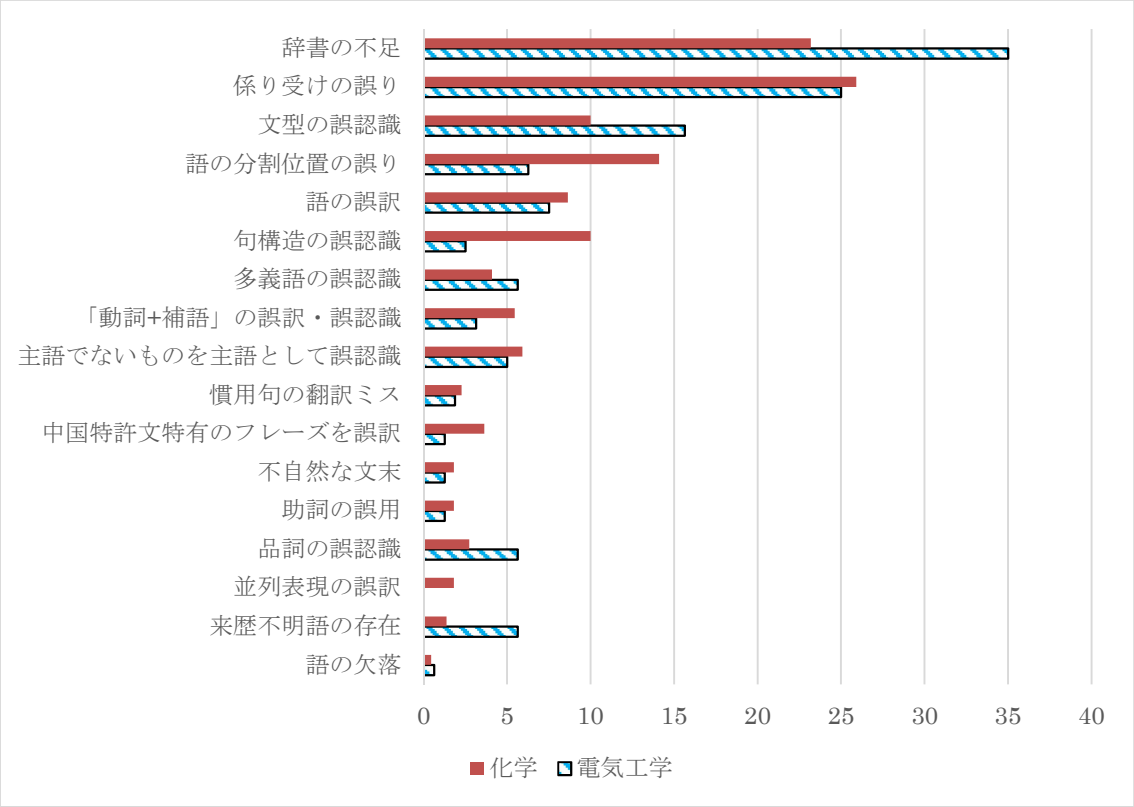


図3.3.4-2 化学と電気工学の誤訳原因分布図

続いて、35技術分野について、各誤訳原因が技術分野によって大きく発生率が変わるのか、技術分野による差が少ないのか分析する。図3.3.4-3に、35技術分野のうち、内容の伝達レベルの評価が最も高かった上位５分野をＡ群、最も低かった下位５分野をＢ群として、それぞれについて主要誤訳原因の割合を示す。

表3.3.4-3 内容の伝達レベルの上位・下位５分野一覧

上位 5 分野（A群）	下位 5 分野（B群）
14. 有機化学、農薬	05. 基本電子素子
16. 製薬	07. ビジネス方法
19. 基礎材料化学	10. 計測
20. 無機材料、冶金	29. その他の特殊機械
22. マイクロ構造、ナノテクノロジー	31. 機械部品

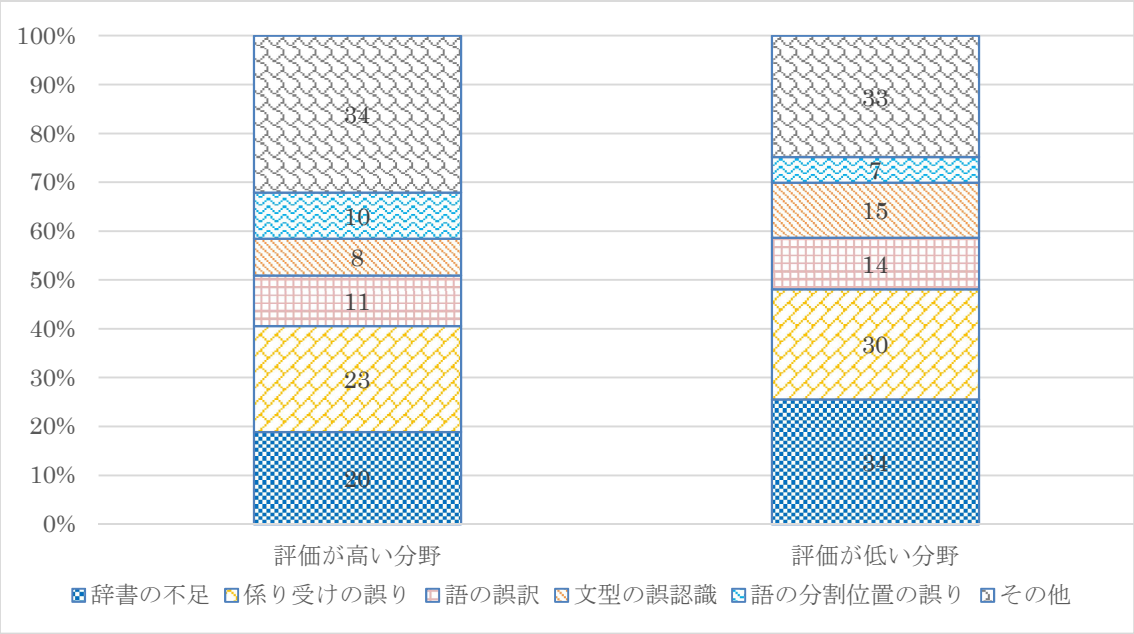


図3.3.4-4 35技術分野の評価が高い分野と低い分野の主要誤訳原因分布図

図3.3.4-1と図3.3.4-2を比較して分かるように、技術分野毎の比較でも、セクタ毎の比較と同様の傾向が見られる。すなわち、比較的評価の低かった分野Ｂ群のいずれの技術分野も、比較的評価の高かった分野Ａ群に比較して、辞書の不足による誤訳の発生割合が大きい。これらの結果は、辞書登録の不足が、分野毎の翻訳品質のバラツキの主要な要因になっていることを意味する。

以上により、計測、医療機器、環境技術、その他の特殊機械の分野など「内容の伝達レベル」と「重要技術用語の翻訳精度」による人手評価の低い分野について、辞書登録を進めることによって、品質が改善されるものと考えられる。

3.4 自動評価結果・分析

3.4.1 BLEUによる自動評価結果

評価文700文を特許庁翻訳とみんなの翻訳でそれぞれ翻訳し、その翻訳結果を自動評価指標のBLEUにより評価した。

（1）700文全体のBLEUによる評価

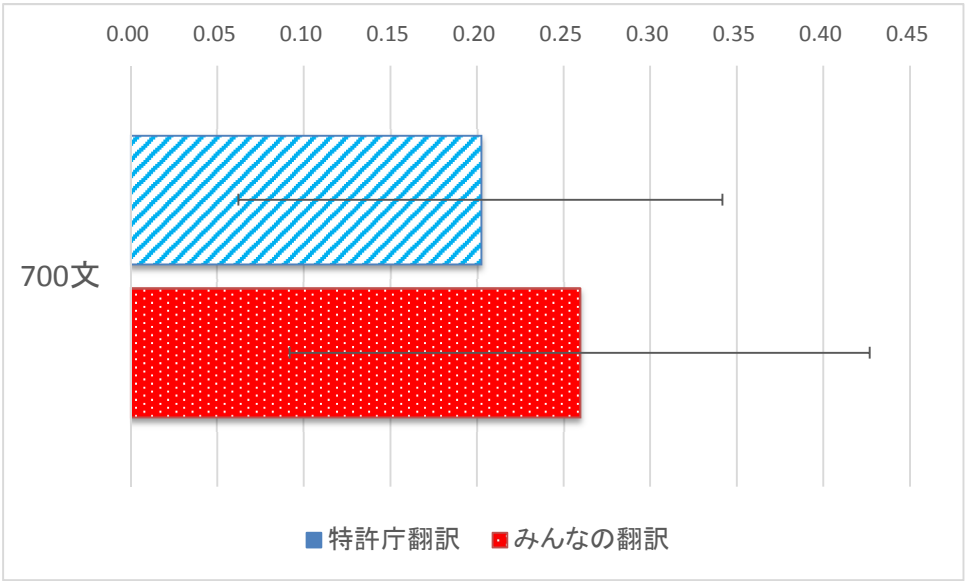


図3.4.1-1 700文のBLEUによる評価の平均値

	特許庁翻訳	みんなの翻訳
平均値	0.2020	0.2593
標準偏差	0.1397	0.1675

みんなの翻訳の方が特許庁翻訳よりBLEUによる評価の平均値が高かった。分布を調べるために、評価値について0.05間隔で度数グラフを作成した。

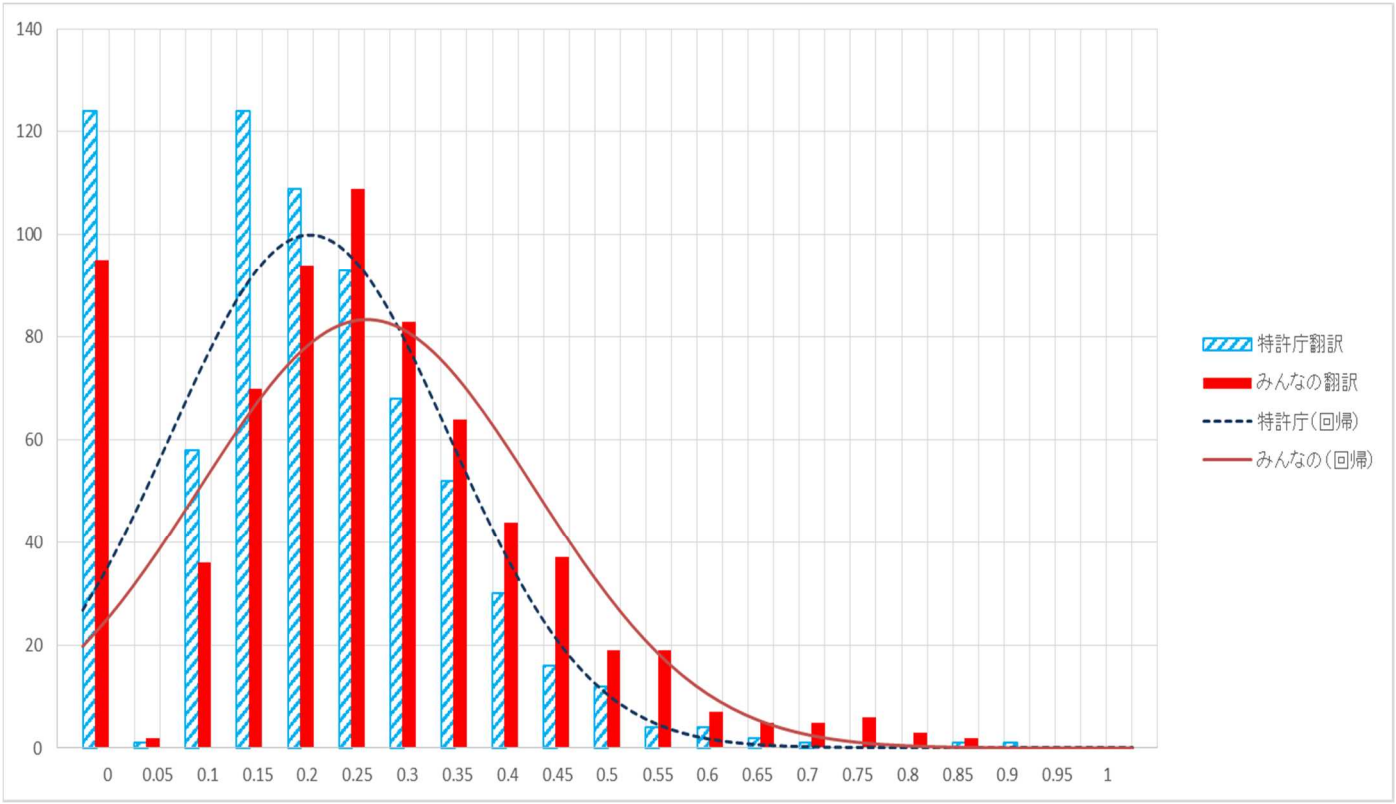


図3.4.1-2 BLEU評価700文の分布グラフ

（２）セクタ毎のBLEUによる評価

次にセクタ毎にBLEUによる評価の平均値をとった。

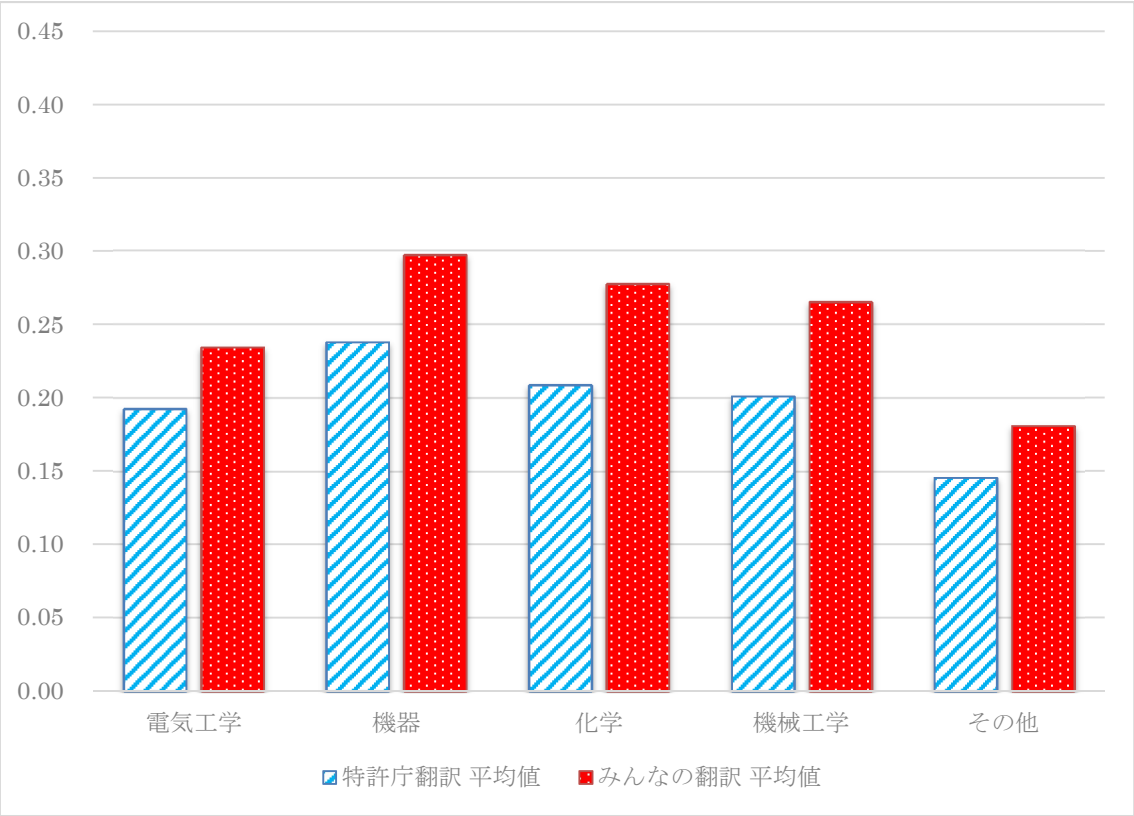


図3.4.1-3 セクタ毎のBLEUによる評価の平均値

		電気工学	機器	化学	機械工学	その他
特許庁翻訳	平均値	0.1926	0.2379	0.2087	0.2012	0.1452
	標準偏差	0.1408	0.1396	0.1424	0.1381	0.1069
みんなの翻訳	平均値	0.2343	0.2970	0.2774	0.2652	0.1811
	標準偏差	0.1601	0.2011	0.1558	0.1661	0.1315

BLEUによる評価では、全体にみんなの翻訳のほうが特許庁翻訳より高い評価値となったが、セクタごとに見れば機器・化学・機械工学・電気工学・その他の順に高い評価値となっていることは、特許庁翻訳・みんなの翻訳の両方で同じ傾向を示した。

（３）35技術分野毎のBLEUによる評価

最後に35技術分野毎にBLEUによる評価の平均をとった。

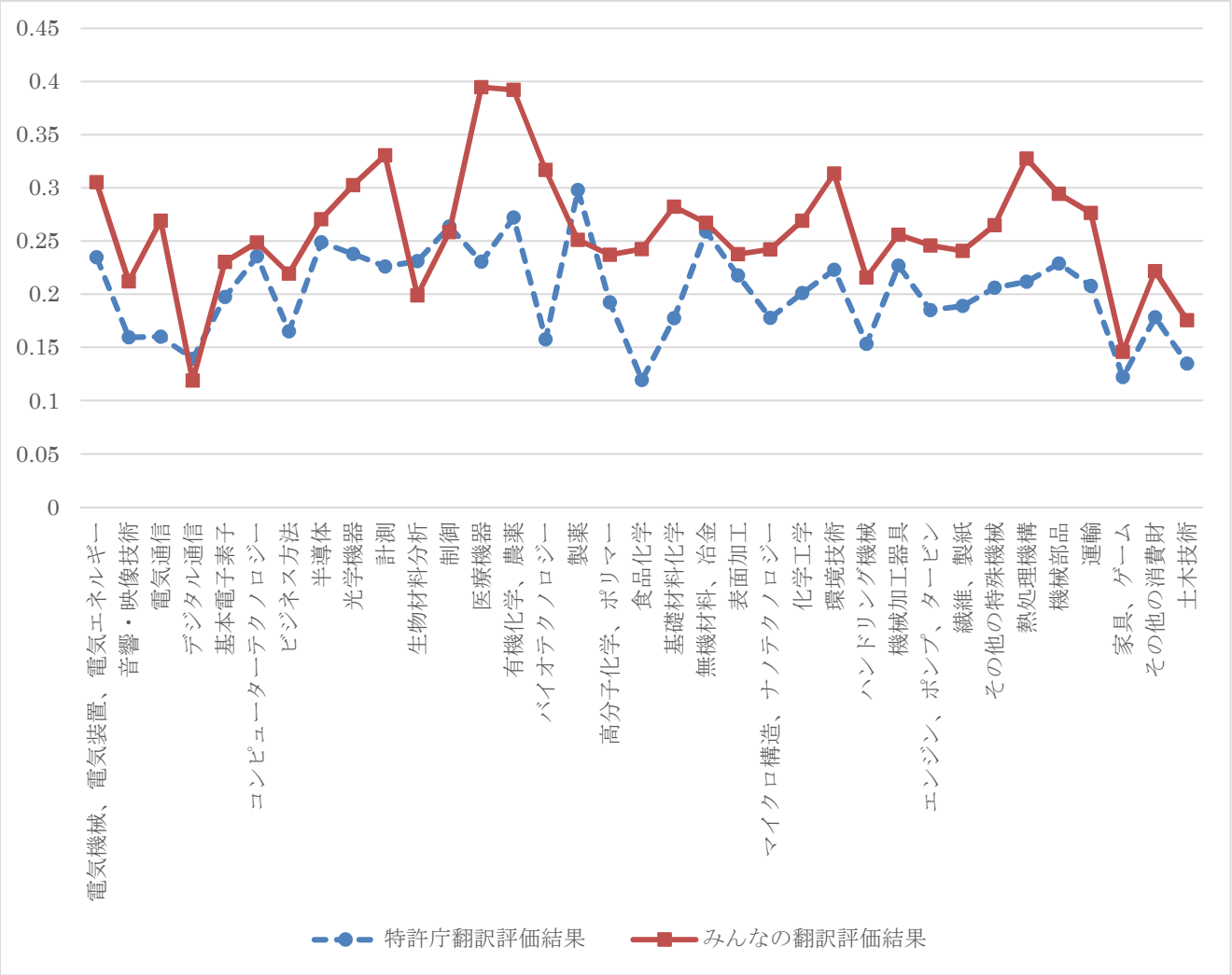


図3. 4. 1-4　35技術分野毎のBLEUによる評価の平均値

3.4.2 RIBESによる自動評価結果

評価文700文を特許庁翻訳とみんなの翻訳でそれぞれ翻訳し、その翻訳結果を自動評価指標のRIBESにより評価した。

（１）700文全体のRIBESによる評価

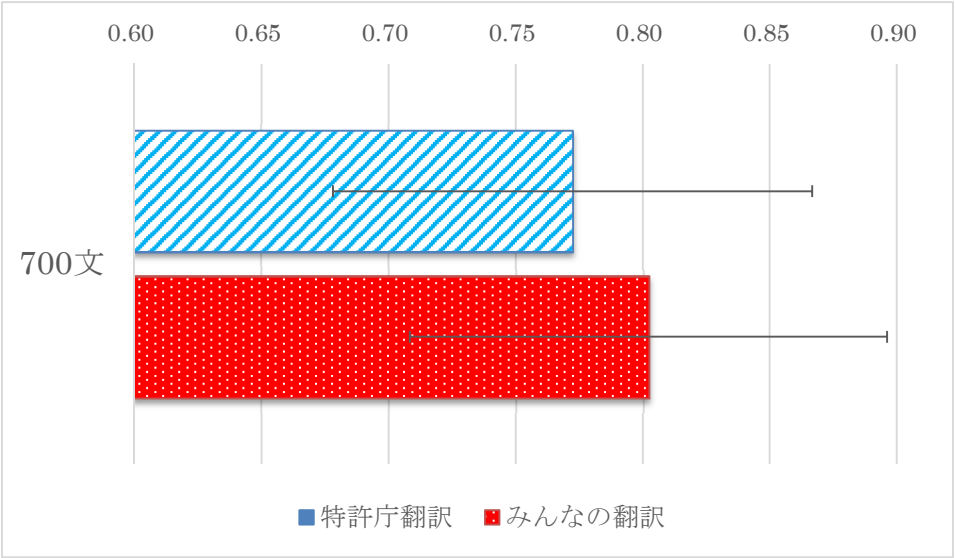


図3.4.2-1 700文のRIBESによる評価の平均値

	特許庁翻訳	みんなの翻訳
平均値	0.7724	0.8022
標準偏差	0.0942	0.0939

みんなの翻訳の方が特許庁翻訳よりBLEUによる評価の平均値が高かった。分布を調べるために、評価値について0.05間隔で度数グラフを作成した。

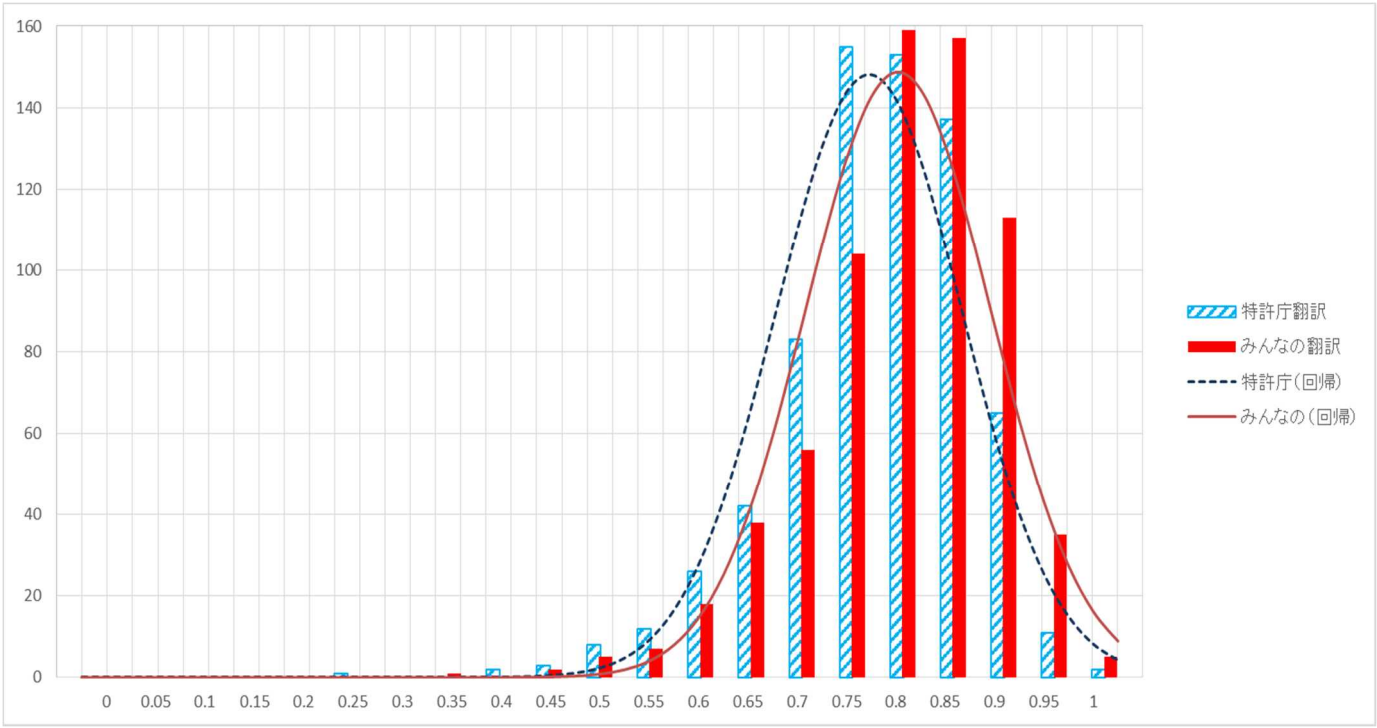


図3.4.2-2 RIBES評価700文の分布グラフ

（２）セクタ毎のRIBESによる評価

次にセクタ毎にRIBESによる評価の平均値をとった。

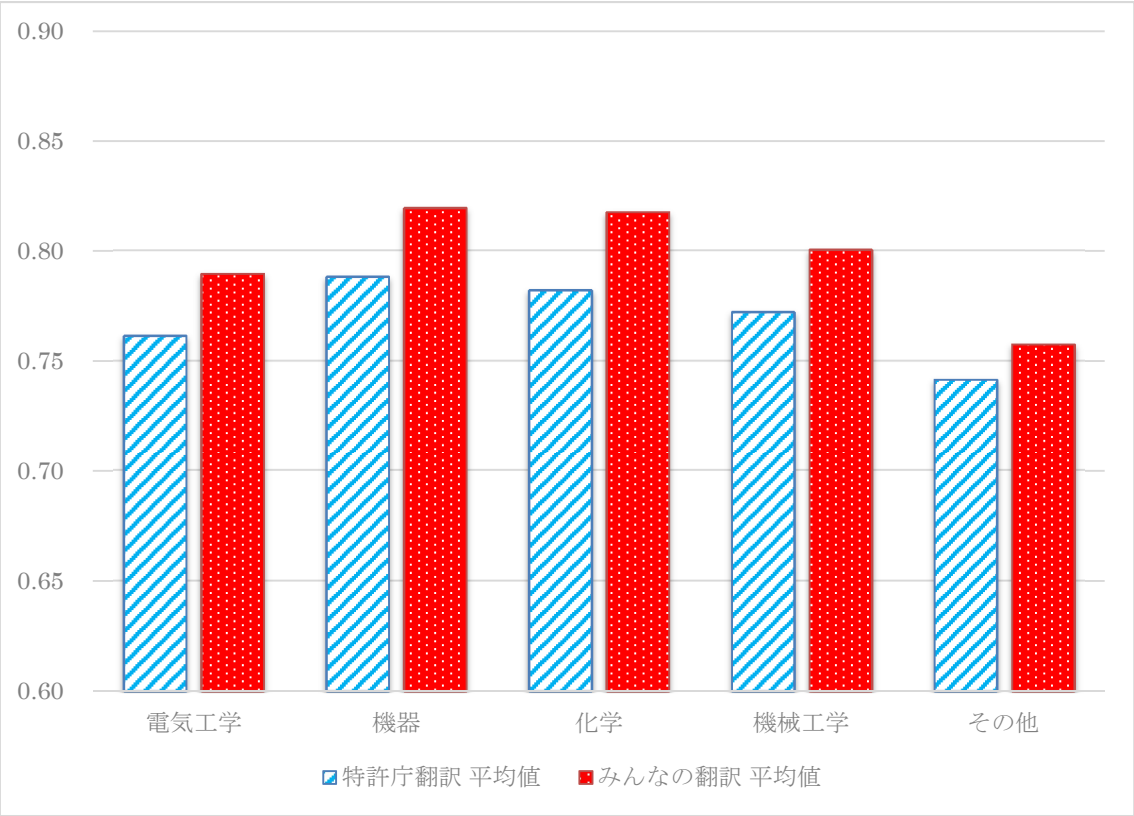


図3.4.2-3 セクタ毎のRIBESによる評価の平均値

		電気工学	機器	化学	機械工学	その他
特許庁翻訳	平均値	0.7613	0.7882	0.7820	0.7721	0.7413
	標準偏差	0.0972	0.0962	0.0917	0.0965	0.0724
みんなの翻訳	平均値	0.7894	0.8193	0.8174	0.8004	0.7574
	標準偏差	0.0966	0.0906	0.0876	0.0971	0.0848

RIBESによる評価では、全体にみんなの翻訳のほうが特許庁翻訳より高い評価値となったが、セクタごとに見れば機器・化学・機械工学・電気工学・その他の順に高い評価値となっていることは、特許庁翻訳・みんなの翻訳の両方で同じ傾向を示した。

（３）35技術分野毎のRIBESによる評価

最後に35技術分野毎にRIBESによる評価の平均をとった。

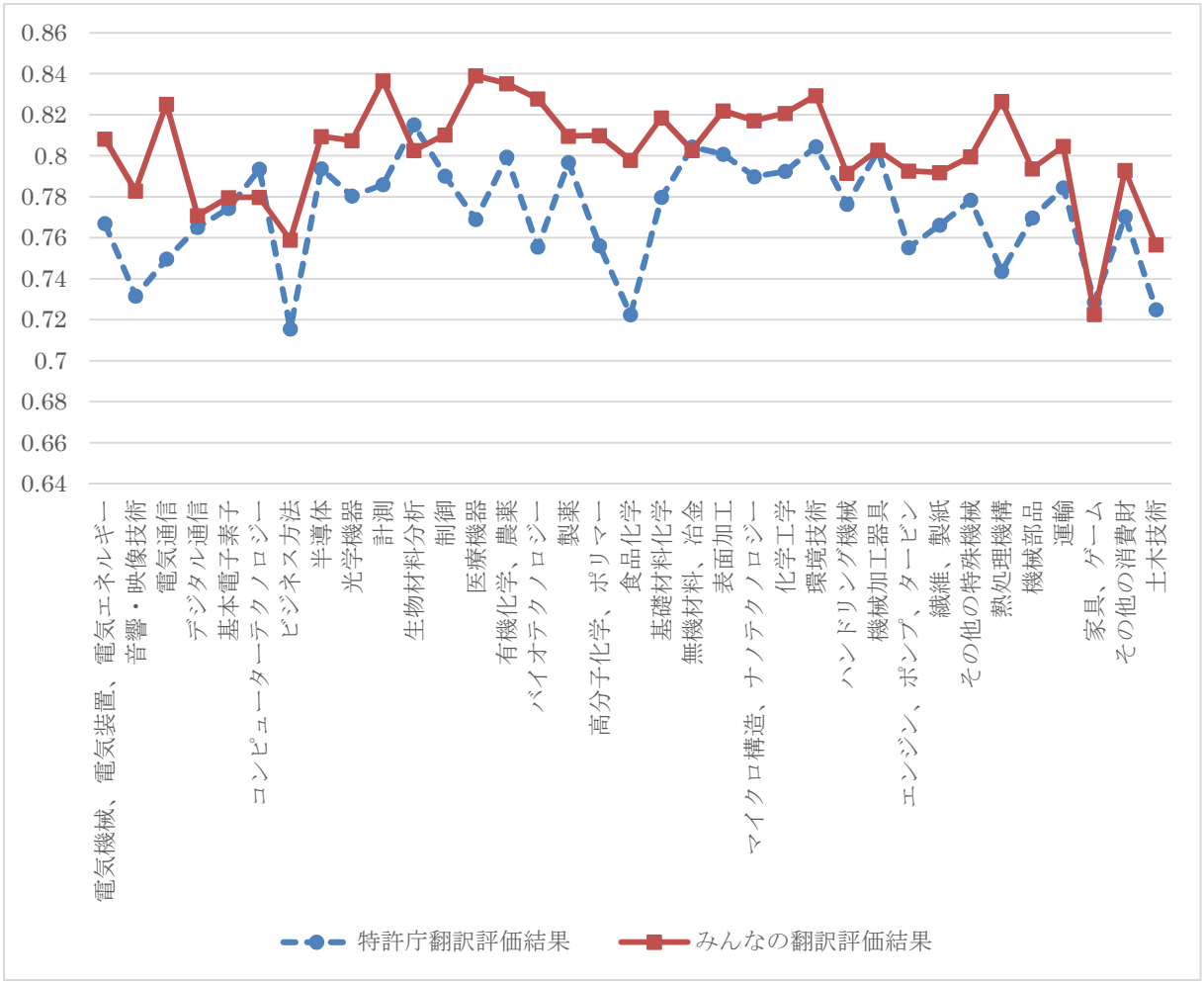


図3. 4. 2-4　35技術分野毎のRIBESによる評価の平均値

3.4.3 自動評価結果の分析

3.4.3.1 人手評価と自動評価の相関関係

内容の伝達レベルの人手評価と機械による自動評価はどれだけ相関性があるかピアソンの積率相関係数を求める。

ピアソンの積率相関係数の数値について、相関性は表3.4.3.1-1に示す。

表3.4.3.1-1 ピアソンの積率相関係数の相関性

相関係数の値(r)	相関性
0	相関なし
0<| r |≦0.2	ほとんど相関なし
0.2<| r |≦0.4	低い相関あり
0.4<| r |≦0.7	相関あり
0.7<| r |<1.0	高い相関あり
1.0 または-1.0	完全な相関

（1）全体(700文)についての相関

700文全体の相関係数を求めた結果を図3.4.3.1-2に示す。700文の評価の結果をみると、人手評価と自動評価の相関係数は、RIBESの方がBLEUより高く、みんなの翻訳の方が特許庁翻訳よりBLEUとRIBESの相関係数が共に高かった。

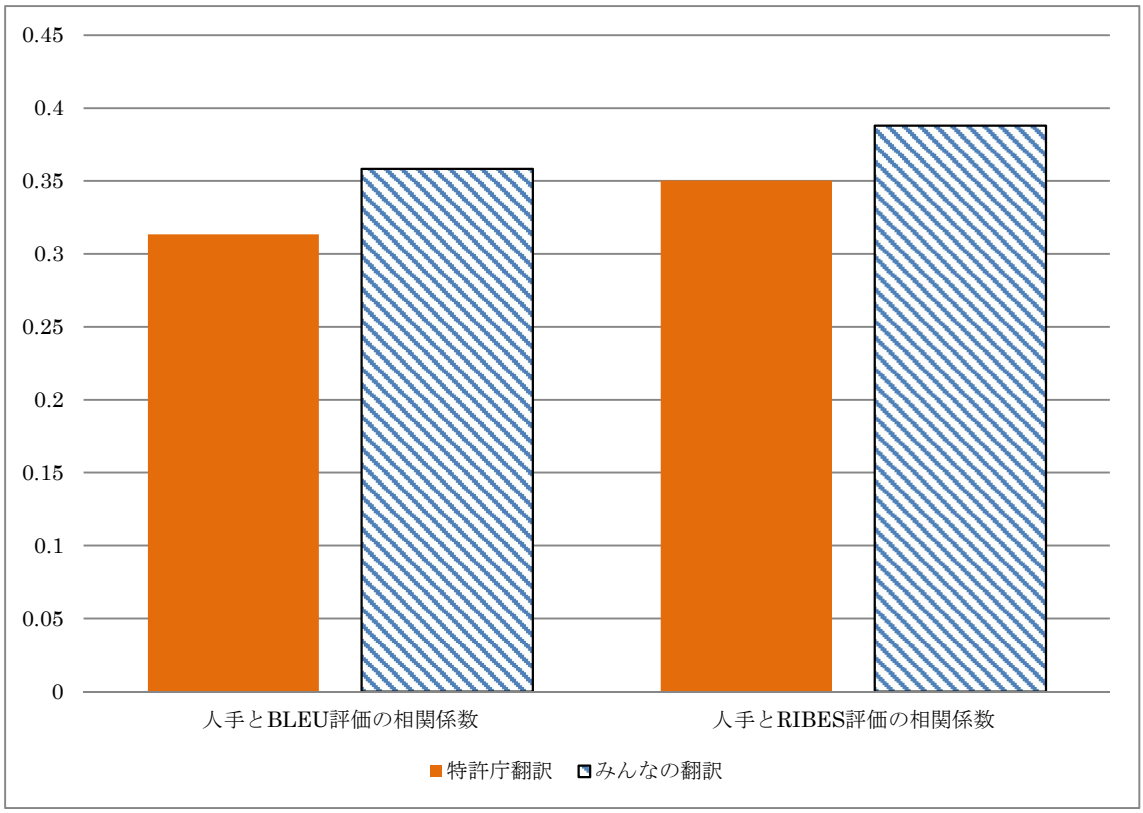


図3.4.3.1-2　内容の伝達レベルの人手評価と自動評価の全体相関係数

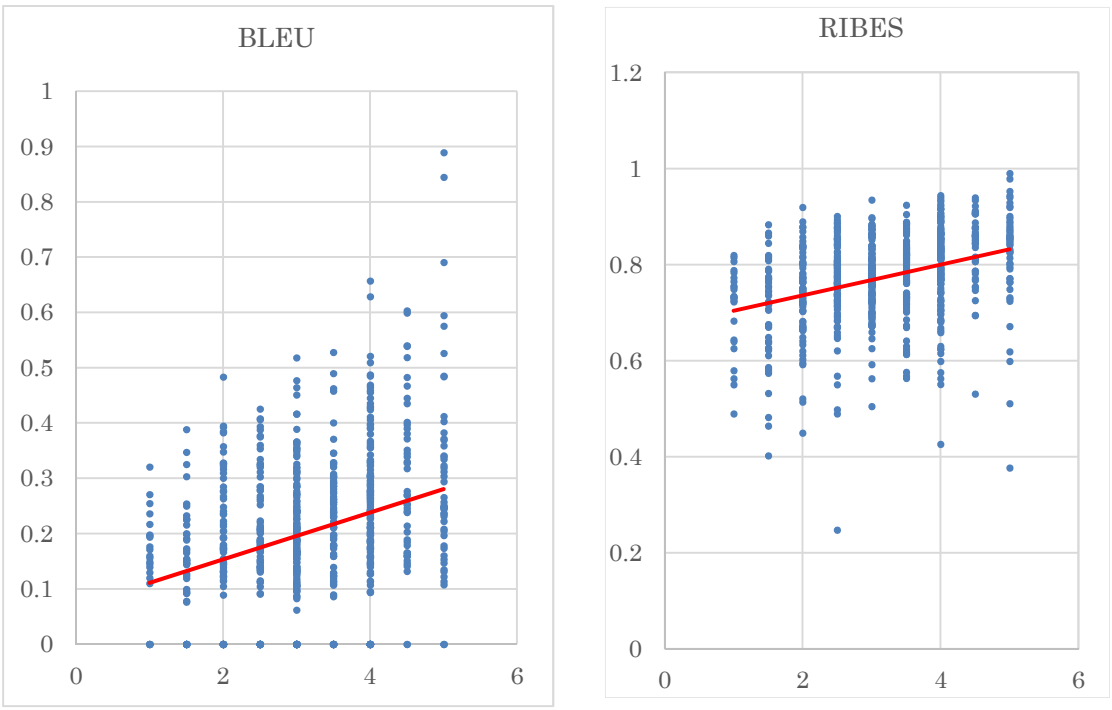


図3. 4. 3. 1-3　特許庁翻訳の人手評価と自動評価の散布図

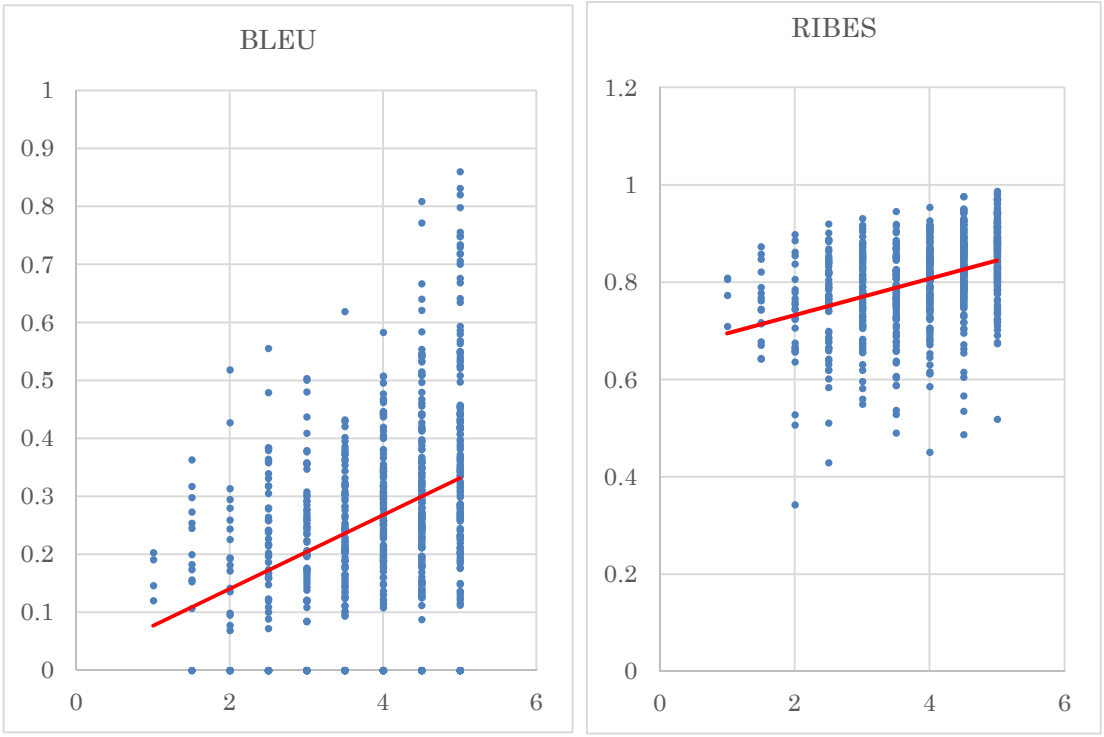


図3. 4. 3. 1-4　みんなの翻訳の人手評価と自動評価の散布図

表3.4.3.1-5　内容の伝達レベルの人手評価と自動評価の全体相関係数

人手と BLEU 評価の相関係数		人手と RIBES 評価の相関係数	
特許庁翻訳	みんなの翻訳	特許庁翻訳	みんなの翻訳
0.3134	0.3585	0.3503	0.3880

評価文700文の人手評価と自動評価の相関係数が 0.2＜ r ≦0.4 にあるので、人手評価と自動評価の間には低い相関があることが分かる。また相関係数はRIBESのほうがBLEUより、人手評価との相関が強い。

（2）セクタ毎(5分野)についての相関

セクタ毎の人手評価とBLEUの相関係数を求めた結果を図3.4.3.1-6に示す。セクタ別の評価結果をみると、特許庁翻訳では、電気工学、機械工学 と その他 分野の相関係数が みんなの翻訳 より高い。電気工学 は 特許庁翻訳 と みんなの翻訳 いずれも0.4以上で、みんなの翻訳 では機器が0.6以上になった。また、その他 分野は特許庁翻訳 と みんなの翻訳 いずれも0.1以下であることが分かる。

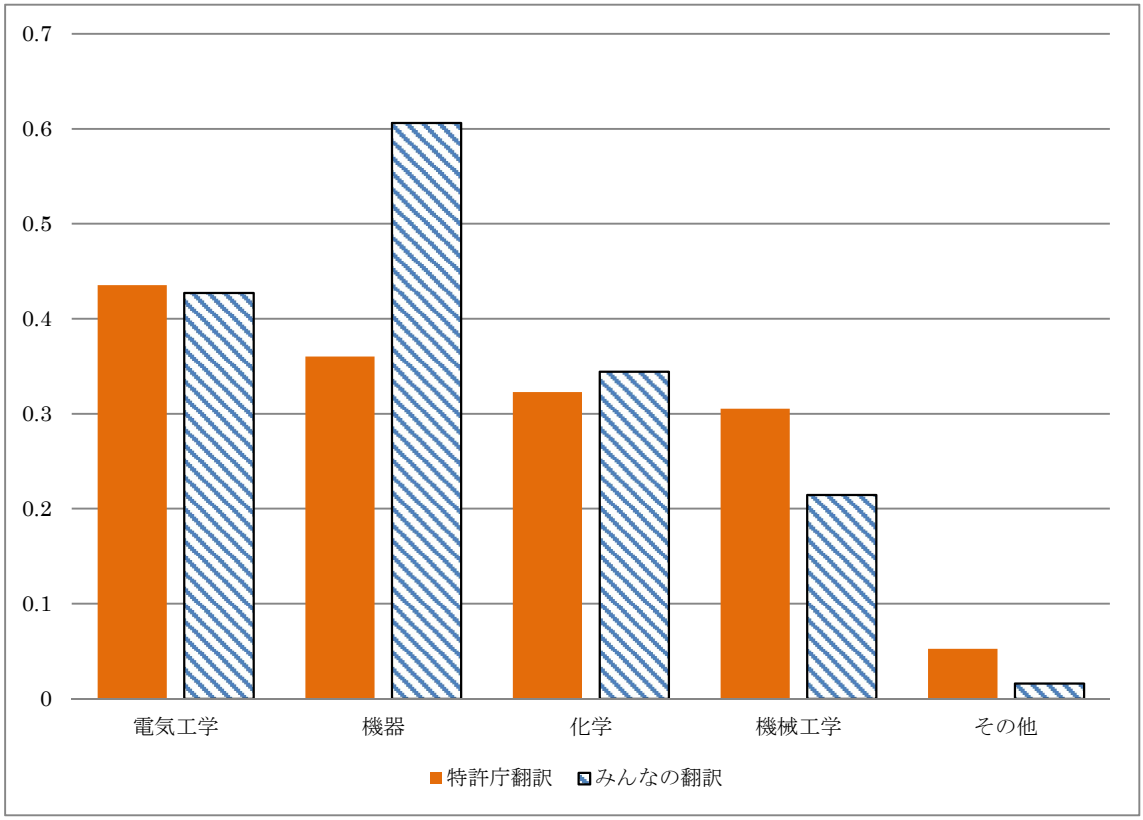


図3.4.3.1-6 人手評価とBLEUのセクタ毎の相関係数

セクタ毎の人手評価とRIBESの相関係数を求めた結果を図3.4.3.1-7に示す。セクタ別の評価結果をみると、特許庁翻訳では、機械工学とその他分野の相関係数がみんなの翻訳より高い。電気工学は特許庁翻訳とみんなの翻訳のいずれも0.4以上で、みんなの翻訳では機器が0.5以上で、特許庁翻訳の機械工学が0.4以上になった。また、その他分野は特許庁翻訳とみんなの翻訳いずれも0.2以下であることが分かる。

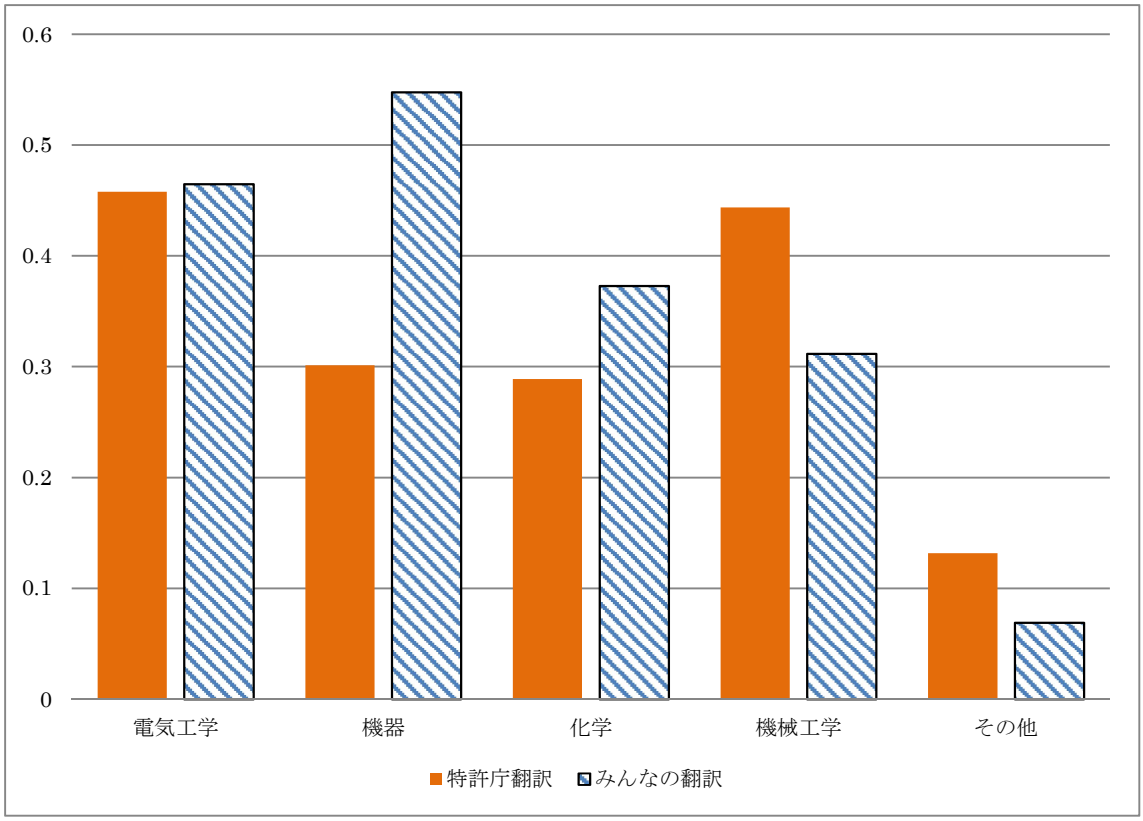


図3.4.3.1-7 人手評価とRIBESのセクタ毎の相関係数

表3.4.3.1-8 内容の伝達レベルの人手評価と自動評価のセクタ毎の相関係数

	人手と BLEU 評価の相関係数		人手と RIBES 評価の相関係数	
	特許庁翻訳	みんなの翻訳	特許庁翻訳	みんなの翻訳
電気工学	0.4355	0.4272	0.4579	0.4647
機器	0.3601	0.6061	0.3012	0.5476
化学	0.3230	0.3442	0.2889	0.3729
機械工学	0.3052	0.2147	0.4437	0.3118
その他	0.0525	0.0163	0.1319	0.0691

35 技術分野毎の特許庁翻訳とみんなの翻訳の人手評価と自動評価の相関係数のグラフを以下に示す。

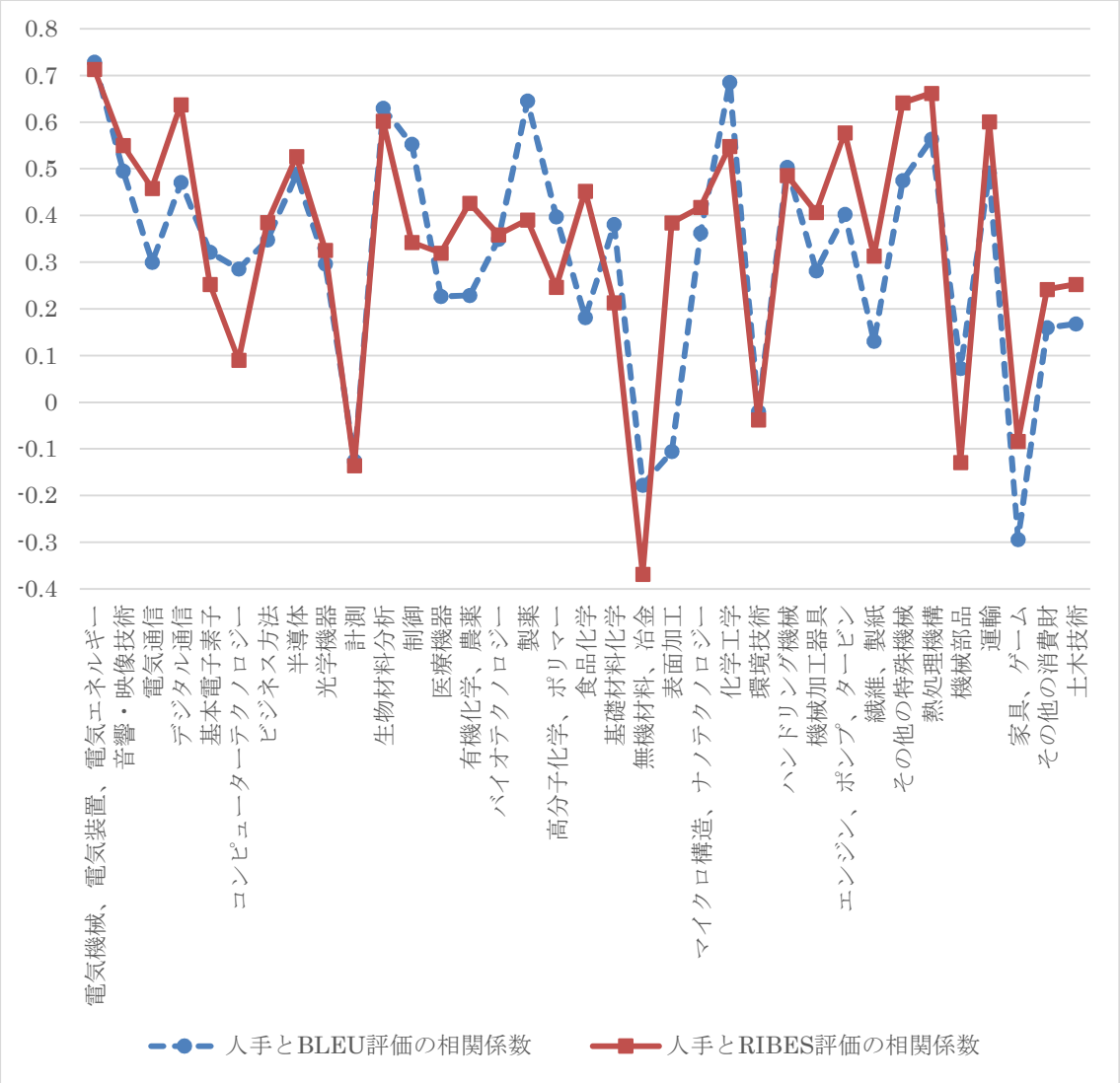


図3. 4. 3. 1-9 特許庁翻訳の技術分野毎の人手評価と自動評価の相関係数

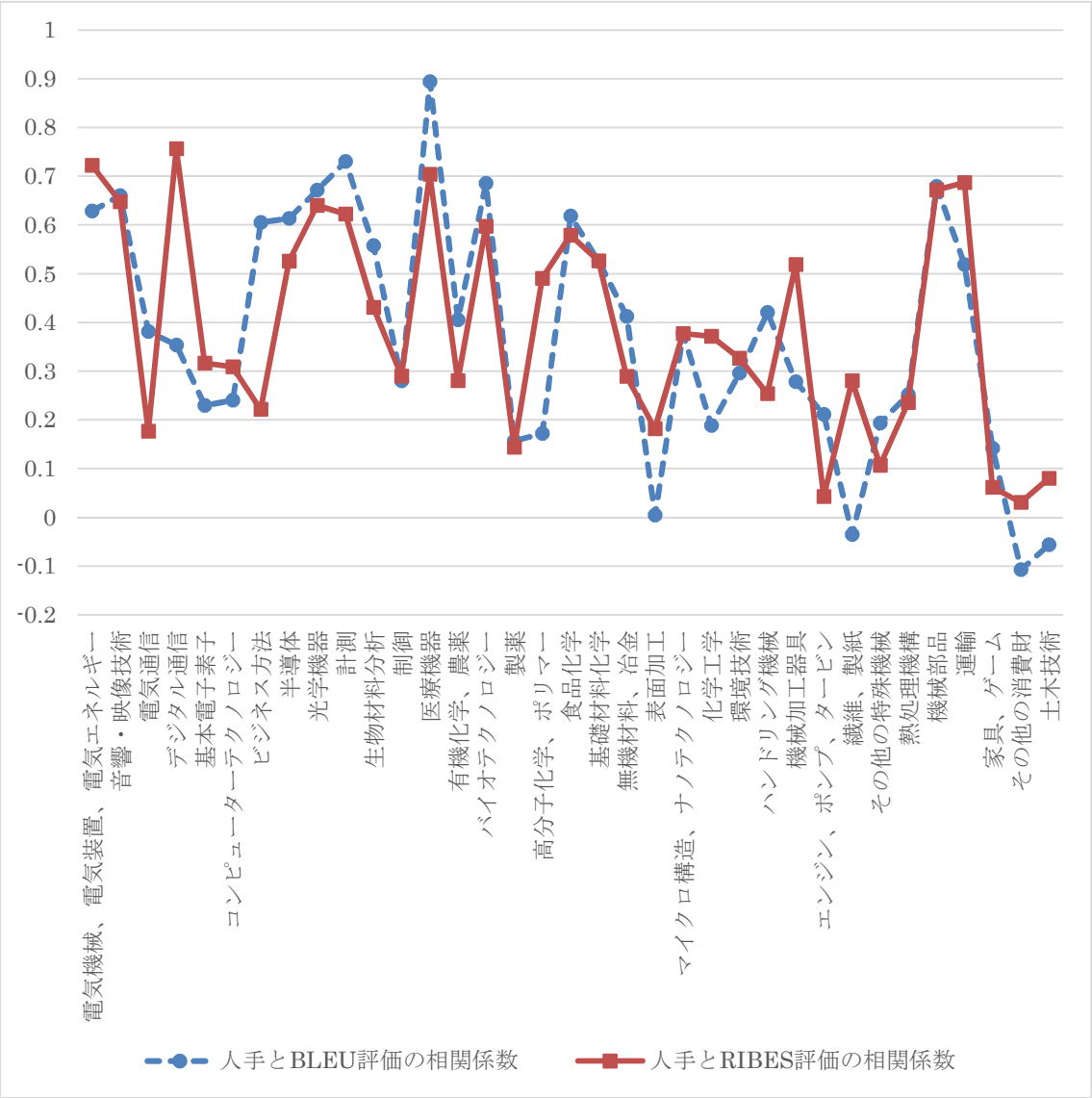


図3.4.3.1-10 みんなの翻訳の技術分野毎の人手評価と自動評価の相関係数

相関関係を調べる際の、35分野の分野毎のサンプル数は、20文と少なく有意水準5%で、何らかの関係があるというためには相関係数[11]は0.443763以上であることが必要である。

したがって、特許庁翻訳で相関があると言える技術分野は、人手評価とBLEUの自動評価が12技術分野、人手評価とRIBESの自動評価が13技術分野である。一方、みんなの翻訳で、前者が13技術分野、後者が14技術分野である。

[11] http://www.biwako.shiga-u.ac.jp/sensei/mnaka/ut/rtable.html

3.4.4 自動評価の利用について

BLEUは一般に広く使われている自動評価ツールである。反面、例えば訳文と参照訳で４個の連続した単語の並びが全く一致を見ない時など（＝良い訳でもしばしば起こり得る）には、スコアが０になってしまうという欠点があり、１文毎の評価には使用しないことが多い。また、同義語を考慮するためには、参照訳を複数用意する必要がある。

自動評価の大きなメリットとして「コストが安価である」「決定した基準について再現性のある一意的な評価を与える」などが挙げられる。

いかにして自動評価を人手評価に近づけるかについては、機械翻訳関連の重要なタスクであり成果が待たれる。本事業に特化した一つの案として、既にある評価基準を使うのではなく本事業のためのカスタムメイドの評価基準を開発することが考えられる。

具体的には、用語の正確さにはBLEU1（形態素１個の適合率）を使用し、構文の正確さには語順が関係していると見てRIBESを使用するなど「特許文献機械翻訳の品質評価手順」に従い評価の側面ごとに異なったアルゴリズムを用いることも考えられる。

今回の自動評価からの結果を見ると、BLEUのスコアが0になっている所が多かった為、BLEU評価の利用について、上記述べたように形態素の適合率を調整して利用しなければならないと考えられる。

3.5 特許文献機械翻訳の品質評価の課題・問題点

3.5.1「特許文献機械翻訳の品質評価手順」の改善点

人手評価は自動評価より、正確であるが、評価者によって評価の差を生じることがある。評価同士の評価の差を少なくするために、評価手順はとても重要である。

「特許文献機械翻訳の品質評価手順」の「内容の伝達レベルの評価」の5段階評価基準をもっと細かく細分したほうが各評価者の評価の差が縮まると想定する。

例えば、現在の評価基準で評点4（ほとんどの重要情報は正確に伝達されている。[80%～]）について、どんな項目が重要情報か明確してないため、評価者それぞれの評価が違う結果になる可能性があるので、評価項目をもっと細かく分けて、各項目の採点を取り、全ての評価項目の合計を最終評点として纏めたほうが良い。

例えば、下記の表3.5.1-1のように評価対象の項目を細分して評価したほうが良い。

表3.5.1-1 翻訳品質評価項目

評価項目	語順	重要単語	未翻訳単語	助詞	品詞	句構造	・・・	合計
評点	5	3	5	2	3	2	X	XX
コメント	この欄には各細分項目に関する説明を書く							

上記表のように、翻訳で誤訳の原因と思われる項目をなるべく細分化し、各細分項目について評価を行うことで、人手評価の品質が上がると同時に、細分項目の採点結果から、どの原因が誤訳を招いたかの分析にもつながることが出来る。

また、重要単語の評価も細分項目にあるので、重要技術用語の翻訳精度評価作業が容易になると考察する。

4. 対訳辞書データの作成・分析

4.1 概要

（1）中日対訳コーパス

2014 年公開の中国公開特許公報と対応する日本公開特許公報を、パテントファミリーを利用して探し出し XML データを解析して 10,811,899 文対の中日対訳コーパスを作成した。また、2012・2013 年公開の中国公開特許公報の要約と対応する和文抄録について、XML およびテキストデータを解析して 4,340,535 文対の中日対訳コーパスを作成した。その過程において、37,632 件の中日公開特許公報番号リストと、867,622 件の中国公開特許公報・和文抄録番号リストを作成した。

（2）中日対訳辞書

上記により作成した 15,152,434 文対の中日対訳コーパスを解析して、対訳辞書候補を抽出。人手確認により 50,144 語の対訳辞書を作成した。

さらに、特許庁より貸与された中国語の「未知語リスト」から 7,000 語を選定し、日本語訳を付け対訳辞書を作成した。以上で作成した 2 種類の辞書を、特許庁 2,210,294 語の辞書に統合し、2,267,438 語の辞書を作成した。また統合された辞書について頻度・複合語数の調査を行った。

4.2 対訳コーパスの作成

4.2.1 テキストデータの抽出

(1)対応する公開特許公報のテキストデータ抽出

DOCDB の backfile から、CN（中国）もしくは JP（日本）とラベリングされた XML について family-id を収集した。また、DOCDB の weekly update から Amend もしくは CreateDelete とラベリングされた XML について family-id を収集した。
以下に DOCDB のデータ例を示す。

表 4.2.1-1　DOCDB データの例（抜粋）

```
<exch:exchange-document country="JP" ... family-id="52302513" ...>
<exch:bibliographic-data>
<exch:publication-reference data-format="docdb">
  <document-id lang="ja">
  <country>JP</country>
  <doc-number>2015008723</doc-number>
  <kind>A</kind>
  </document-id>
</exch:publication-reference>
</exch:bibliographic-data>
</exch:exchange-document>
```

公開特許公報（種別 A）は、CN が 4,731,144 件、JP が 12,483,618 件あった。
そのうち CN が 2014 年以降の公開となっているもので絞り込んだところ、
family-id が 71,772 件、それに含まれる CN が 73,344 件、JP が 76,555 件あった。

これを基に今回特許庁から貸与された公開特許公報 XML データを取り出したところ、family-id が 34,710 件、それに含まれる CN が 35,297 件、JP が 36,293 件となった。
ひとつの family-id に対して CN もしくは JP が複数含まれる場合があるが、その組み合わせをそれぞれペアとして数えた場合の中日文献対の総数は 37,873 件であった。

以上の情報をまとめ、対応する文献の文献番号を対とした対応中国・日本公開特許公報番号リストを作成した。以下に番号リストのデータ例を示す。

表 4.2.1-2　番号リストの例（抜粋）:

pub-num(CN)	pub-date(CN)	pub-num(JP)	pub-date(JP)	ipc	fam-id
CNA103477788	2014.01.01	JPA2013255491	2013.12.26	A01D34/73	48626242
CNA103478016	2014.01.01	JPA2013255431	2013.12.26	A01K41/00	49818886
CNA103478137	2014.01.01	JPA2011032260	2011.02.17	A01N43/80	42668923

またこの番号リストを基に取り出した中国もしくは日本公開特許公報の XML データより「名称」「要約」「請求項」「明細書」をテキストデータとして抽出した。

中国・日本公開特許公報の対については XML データ中で左記構造の並びが対応していない場合があるが、それぞれの内容を別々のテキストとして取り出すことで、この問題に対処することができた。以下に中国公開特許公報 XML のデータ例を示す。

表 4.2.1-3　中国公開特許公報 XML の例（抜粋）:

```
<cn-patent-document lang="ZH" country="CN">
  <cn-bibliographic-data>
    <cn-publication-reference>
      <document-id>
        <country>CN</country>
        <doc-number>104255087</doc-number>
        <kind>A</kind>
        <date>20141231</date>
      </document-id>
    </cn-publication-reference>
    <classifications-ipcr>
      <classification-ipcr>
        <text>H05K    9/00     (2006.01)</text>
      </classification-ipcr>
    </classifications-ipcr>
    <invention-title> 线束 </invention-title>
   <abstract>
     <p num="1">线束(10)具备：三根电缆(9)，用于传导对称三相交流电；...</p>
   </abstract>
 </cn-bibliographic-data>
```

```
  <application-body lang="CN" country="CN">
    <description>
      <invention-title id="tilte1">线束</invention-title>
      <p id="p0004" num="0004">铺设于以机动车为代表的车辆的线束有时候具备包围电线...</p>
    </description>
    <claims>
      <claim id="cl0001" num="0001">
        <claim-text>导电性的三个电磁屏蔽部件(1)，隔开间隔地并列排列...<br/></claim-text>
      </claim>
    </claims>
  </application-body>
</cn-patent-document>
```

(2)対応する要約のテキストデータ抽出

2012・2013 年公開の中国公開特許公報の「要約」部分の和文抄録について、特許庁より貸与を受けた。この 867,622 件の和文抄録より、「名称」「要約（抄録本文)」のそれぞれをテキストデータとして抽出した。

次に 2012・2013 年公開の中国公開特許公報の XML ファイル 867,622 件より、「名称」「要約」のそれぞれのテキストデータを抽出した。以下に和文抄録のデータ例を示す。

表 4.2.1-4　和文抄録の例（抜粋）:

```
<PATDOC>
<SDOBI>
<B110>103476239</B110>
<B511>H05K13/04</B511>
<B542>自動的に回路基板を挿入および取り外す自動部品挿入機</B542>
</SDOBI>
<SDOAB>
<P>抄録文</P>
<P>自動的に回路基板を挿入および取り外す自動部品挿入機であって...</P>
</SDOAB>
</PATDOC>
```

この過程で 867,622 件の対応中国公開特許公報・和文抄録文献番号リストを作成した。

4.2.2 抽出した特許文献の対訳文アライメント

抽出した文献対を解析して対訳コーパスを作成した手順について述べる。

テキストデータを分析し、対応する文を推定する自動文アライメント処理を行う際には、対訳文アライメントツール（国立研究開発法人・情報通信研究機構より提供）を使用した。本ツールは、特許庁の前年度の調査でも使用実績があり、自動文アライメントについて高い精度を有するツールであるため、対訳コーパスならびにそこから作成される辞書について高品質が期待されることが選定の理由となった。アライメントの精度をさらに上げるため、対訳文アライメントツールは特許庁より貸与された約 220 万語対訳辞書であらかじめ学習させた。

技術内容が対応する中日公開特許公報の対から取り出した「名称」「要約」「請求項」「明細書」のテキストデータから対訳文アライメントツールによって作成された中日対訳コーパスの文対数は 10,811,899 件となった。また、技術内容が対応する中国公開特許公報の要約と和文抄録の対から取り出した「名称」「要約」のテキストデータから対訳文アライメントツールによって作成された中日対訳コーパスの文対数は 4,340,535 件となった。以上、総計 15,152,434 文対を持つ中日対訳コーパスを作成した。アライメントスコアの文数は、以下のようになった。

表 4.2.2-1　アライメントスコアの分布：

0.04 以上：	11,403,478 文対	75.3%
0.06 以上：	7,883,480 文対	52.0%
0.08 以上：	4,800,392 文対	31.7%

対応付けられた総計 15,152,434 件の対訳文対には、(a)精度を示すアライメントスコア (b)中国公開特許の公開番号 (c)日本公開特許の公開番号もしくは和文抄録の文献番号 (d)特許文献の国際特許分類(IPC)を付した。また、対訳文対が「要約」「請求項」「明細書」「和文抄録」のいずれの部分から作成されたものであるか区別できるようにするために、文献番号にサフィックスを付けた。以下に対訳文対の例を示す：

表 4.2.2-2　対訳文対の例：

0.23590158 ||| CNA103828145-JPA2013084923_abs ||| H01S3/121 ||| E00 ||| 9 ||| 1-1 ||| 激光光源单元 13 射出彼此不同的多个波长的脉冲激光。 ||| レーザ光源ユニット１３は、相互に異なる複数の波長のパルスレーザ光を出射する。

作成した対訳コーパスでは、一つの文献ファイルを「要約」「請求項」「明細書」「和文抄録」に分割した。またそれぞれのファイルにおいて、文アライメント精度の目安となるアライメントスコアに基づき、対訳文対を精度の高い順に並び替えた。

特許庁から貸与された特許公報その他のデータは、圧縮した状態でテラバイト単位の容量であったため、必要なデータだけを検索して SSD に展開するなどした。また処理に際しては、LAN で接続された Linux マシン 4 台（総論理コア数 : 24）に並列で実行できるようなシェルスクリプトを作成・使用した。

4.3 対訳辞書の作成

4.3.1 対訳コーパスに基づく対訳辞書作成

(1)対訳コーパスに基づく対訳辞書候補データ作成

前節において作成した対訳コーパスを、対訳用語抽出ツール（国立研究開発法人・情報通信研究機構より提供）により解析し、対訳となっていると思われる 1 つ以上の連続した名詞もしくはサ変名詞の並びを抽出した。この際の前処理として、作成した対訳コーパスのうち、対訳文アライメントツールのアライメントスコアが 0.08 以上となるものを使用した。さらに対訳コーパスの全文対を 35 技術分野（表 3.1.1-1）に分類した。

抽出された対訳辞書候補データについては、人手確認による対訳辞書データ作成をより効率的・効果的行うために、さらに以下のいくつかの処理に掛けて絞り込んだ。

・対訳コーパスからあらためて用語抽出を行った。具体的には、(a)中国語、日本語をそれぞれ形態素解析　(b)統計的機械翻訳システム Moses を使用して、中・日の品詞情報付きフレーズテーブルを作成 (c)フレーズテーブルのうち名詞句もしくは未定義語と判定されるものだけを抽出 (d)先の対訳用語抽出ツールによる対訳辞書候補データとの積集合を抽出 というステップの処理結果をあらためて対訳用語候補データとした。
・明らかに誤りと思われる用語（中に助詞を含む語など）あるいはノイズと思われるデータ（記号で始まるものなど）を取り除いた。

また、特許庁から貸与を受けた対訳辞書 2,210,294 語と重複するデータを排除した。こうして得られた辞書候補データは、出現頻度が高く複合語の語数が少ない語対を優先的に人手確認に掛けるため、ソートした。また、候補データに付与した全 35 技術分野によって候補を分割し、人手による評価作業によって用語の補強が必要とされる割合の高い 12 技術分野（表 4.3.1-2）だけを使用した。

以上の手順で、92,657 語の対訳辞書候補データを作成した。以下に例を示す。

表 4.3.1-1　対訳辞書候補データの例

中国語	日本語	技術分野番号
光伏发电	太陽光発電	10, 23, 29
光伏发电站	太陽光発電所	10, 23
光伏发电系统	光起電力システム	10
光伏发电系统	太陽光発電システム	10
光伏发电装置	太陽光発電装置	13, 24
光伏提水	太陽光発電揚水	29
光伏电池板	太陽光発電パネル	10, 32
光伏组件	太陽電池モジュール	10, 19, 23, 25
光伏组件	太陽電池部品	32
光传导作用	光伝導作用	10
光传感	光センサー	10
光传感	光検出	10
光传感部	光センサー部	10
光作用	光作用	23
光侦检器	光検出器	13

(2)対訳辞書候補データの人手確認

3章の「内容の伝達レベル」と「重要技術用語の翻訳精度」の人手評価結果から共に評価が低い4分野（計測、医療機器、環境技術、その他の特殊機械）の辞書を作成することを特許庁の担当者と協議の上、決定した。

なお、上記４分野で作成した辞書が5万語未満の場合、「重要技術用語の翻訳精度」の評価が低い技術分野から5万語の辞書データを作成することを決定した。

辞書作成対象技術分野は以下の通りである。

表4.3.1-2 辞書作成技術分野

優先順位	辞書作成分野
1	計測
2	医療機器
3	環境技術
4	その他の特殊機械
5	バイオテクノロジー
6	運輸
7	ハンドリング機械
8	繊維、製紙
9	生物材料分析
10	製薬
11	基礎材料化学
12	化学工学

つまり、人手確認用対訳辞書候補データの「計測」分野から辞書を作成し、5万語に達した時点で作業を終了する。

なお、「計測」分野の人手確認用対訳辞書候補データを全て使用して作成した辞書が5万語未満の場合、表4.3.1-2で示した「計測」の次に優先順位が高い「医療機器」分野の辞書を作成する。

表4.3.1-2で示した12技術分野を全て使用して、作成した辞書が5万語未満の場合、別途特許庁の担当者と協議の上、対象技術分野を決める必要がある。

今回の5万語辞書は表4.3.1-2に示した12技術分野全てを応用して作成した。

上記4.3.1(1)で作成した92,657語の対訳辞書候補データについて、人手による確認作業を行い、50,144語の中日対訳辞書データを作成した。

また、人手チェックで不採用になった単語について、不採用理由を記録し、「人手確認により対訳辞書データから除外したデータ」を作成した。

除外したデータの例については、以下の通りである。

表4.3.1-3 不採用データの例

不採用理由	見出し語	訳語	備考
D1:見出し語自体が不適切	与外杯	外側カップ	見出し語の「与」が不要である
D2:訳語自体が不適切	像素	素	訳語に「画」が抜けている 「画素」が正しい
D3: 見出し語と訳語が対応していない	元素周期表	周期表	見出し語と訳語は用語として問題ないが、見出し語の「元素」が訳語で対応してないため、不採用とする
D4: 辞書に登録する用語として不適切	井上博夫	井上博夫	人名、地名等は対象外とする
	周围区	周辺領域	見出し語と訳語は対応しているが、見出し語が「周围区域」になる場合、「周辺領域域」と翻訳される可能性があるので、不採用とする

人手確認用の品質担保の為、以下の（ｉ）、（ii）の作業を行った。

（ｉ）人手確認を行う人手確認用対訳辞書候補データのうち5%(約5千語)については、異なる２名の訳語確認者が対訳辞書データへの採否の判断をそれぞれ行い、判断の異同を確認し、判断が異なる場合にはいずれの訳語確認者が誤っているのかを分析した。

（ii）人手確認を行い採用した対訳辞書データ及び特許庁が貸与する対訳辞書を統合し、その中のうち本事業で新たに採用した対訳辞書データに含まれる一つの見出し語に対してＮ個以上の対応する訳語が存在するもの、又は本事業で新たに採用した対訳辞書データに含まれる一つの訳語に対してM個以上の対応する見出し語が存在するものについて、Ｎ個以上の訳語又はM個以上の見出し語の妥当性を再確認した。

ここで、Ｎ及びMの値については、再確認する訳語と見出し語の対が、5,000語対程度

になるように、特許庁と協議して決定した。

（ｉ）及び（ⅱ）の作業の結果については、特許庁に報告するとともに、必要に応じて策定した採用基準の修正、判断を誤った訳語確認者への指導、判断を誤った訳語確認者が確認した語の再確認等を行った。

4.3.2 中国語未知語リストからの辞書作成

(1) 中国語未知語データの選定

特許庁から貸与された「未知語」を含むサンプル文（「中韓文献翻訳・検索システム」において未知語と推定された語を含む中国語の文）を複数参照し、前節で作成した対訳辞書データおよび特許庁から貸与された対訳辞書のいずれも含まれないことを確認した。

さらに、記号など特定の文字で始まるデータを排除し、また表 4.3.2-1 のように中国語形態素解析を利用して一つもしくは複数の名詞のみからなる文字列を抽出することによりさらなる絞り込みを行い、これを人手確認用未知語候補データとした。候補データには４分野（化学・電気・機械・物理）の分野情報とサンプル文ならびに今回作成した対訳コーパスにおける頻度情報を付加し、人手作業の利便性を高めた。候補データの例を表 4.3.2-2 に示す。

表 4.3.2-1　構成語が全て名詞と解析された未知語の例

光伏|NN 主站|NN 智能|NN 体|NN
光伏|NN 产品|NN 生产国|NN
光伏|NN 人|NN
光伏|NN 侧|NN
光伏|NN 充电器|NN
光伏|NN 光伏|NN 接口|NN 变换器|NN
光伏|NN 光热|NN 组件|NN 结构图|NN
光伏|NN 农业|NN 大棚|NN 顶|NN
光伏|NN 冰|NN
光伏|NN 区|NN
光伏|NN 单元|NN 卡|NN
光伏|NN 厂|NN
光伏|NN 发电|NN 体系|NN
光伏|NN 发电厂|NN

表 4.3.2-2　人手確認用未知語候補データ（頻度・分野情報、例文付き）の例

頻度	候補	分野情報	例文
272577	感器	C00/E00/M00/P00	昆虫触角的化学【感器】对化合物进行识别...
106653	储器	C00/E00/M00/P00	其中歧管 31 通过适当的管子与不同的液体【储器】(单独标记)连接...
42944	面形	C00/E00/M00/P00	且其【面形】是一种非轴对称的回转面...
41765	流电	C00	具体而言【流电】地涂覆平面产品(例如晶片)以便产生半导体元件...
40768	出端	C00/E00/M00/P00	烘房【出端】与智能切面机能过锚链连接...
36701	缘层	C00/E00/P00	LED 灯铝基板的【缘层】是铝基板最核心的技术...
35068	缘膜	C00	受精卵发育成游泳的幼体，叫做【缘膜】幼体。
28162	苯乙	C00/E00/P00	醛类中的糠醛、十四醛、3-甲基-2-丁烯醛、【苯乙】...
26520	电装	C00/E00/M00/P00	属于电子产品的【电装】工具技术领域。
26449	子部	C00/E00/M00/P00	下表面与所述导流【子部】非连通式连接；
26241	吡咯	C00	...哌嗪-1-基乙氧基)-2(5H)-【吡咯】...
25839	后进	C00	经 7 天自然风干【后进】 窑烧结...
24493	氢氧	C00/E00/M00/P00	本实用新型涉及一种【氢氧】发生装置...

(2)未知語の人手による日本語訳付け

「4.3.2(1) 中国語未知語データの選定」処理で、頻度順に並び替えた未知語のリストから7,000語の見出し語を選び、訳付け作業を行った。

未知語を4分野に振り分けた上で中国語・日本語双方に堪能であり、対象の未知語を含む各サンプル文の属する技術分野について十分な技術的知識を有する者により訳付けを行った。

① 未知語の例文

選定した見出し語が適切であるかを判断するために、以下のように見出し語と見出し語が現れる例文の一部を抽出した。

頻度	中国語	採用判定	例文1	例文2	例文3	例文4
23168	子装		2008年，陈彬制备了加在面包预加粉当中的粉末油脂，像方便面调料那样的小袋【子装】，让酵母粉和粉末油脂一起加入这个预加粉末中，方便了面包的起泡和保持营养的完全，做出理想的面包。	所述杯体上设有可以测量温度的温度计以及测量杯【子装】水容量的刻度线。		
22821	聚氨		水溶聚合物也可以是如在US5,338,477中描述的聚醚【聚氨】亚甲基膦酸酯或者膦基聚丙烯酸。	把5%的乙基纤维素和3%的【聚氨】树脂溶解在34.5%的松油醇中，至充分溶解后加入10%的邻苯二甲酸二乙酯、47.5%的蓖麻油，混合搅拌均匀，得到有机粘合剂待用。	输送器设备100包括一或多个有弹性的带状物部分，例如第一及第二带状物部分102A、102B(第一及第二带状物部分例如由【聚氨】甲酸酯(polyurethane)或其他适合的弹性聚合物材料制成)，第一及第二带状物部分102A、102B通过迭接接合连接(图中示出了一迭接接合104)。	透水性混凝土试件采用济南产山水牌的42.5的普通硅酸盐水泥，【聚氨】基类高效减水剂，以粒径为10-20mm、5-10mm、2.5-5mm等的碎石作为粗骨料用一般洁净自来水作为拌合水制作。
22719	所选		但给药剂量可随着病人的需要、欲治疗的感染的严重性、【所选】化合物等而变化。	对于已创建的集群20，如果需要动态添加机器，则通过中心管理服务器10从资源池50的机器中挑选一台或几台加入集群，然后启用拟部署在【所选】机器上的服务进程；	在此描述的附图仅用于【所选】实施例而不是所有可能实施方式的例示目的，并且不意欲限制本公开的范围。	根据实际【所选】器件型号计算得ωn=1/(100kΩ×1μF)=10rad/s。
22510	置时		为了保证面罩前置或者后【置时】，配光面型不变且都为蝙蝠翼，面罩需要选择低分散度的面罩。			
22186	解质		开拓各种不同规格的解电常数并加以控制，而且【解质】损耗更低的新型材料代替自由空气，是实现天线和滤波器平面化、微型化和智能化的主要技术途径。	造型【解质】(关节能自由运动的金属覆盖体，一种盔甲)；	在这里，气体传感器元件是在以高温加热的情况下使用的，使得固体【解质】主体被激活。	

図4.3.2-3 未知語採用ファイル

②不適切な見出し語

未知語リストはプログラム等でたくさんの不要データを除外したものの、プログラムで除外できていないものも含まれているので、選定は慎重に行う必要がある。

未知語リストから見出し語を選ぶ際、下記の判断を行った。

A)未知語が適切であるか

B)適切な未知語は名詞であるか

C)適切な未知語は辞書登録しても良いか

上記の判断に基づき、除外したデータの例を以下に示す。

表4.3.2-4 除外した未知語データの例

判断基準	未知語	文例	備考
A)未知語が適切であるか	・・合	从・施例4和・施例5的・果可以看出，与・・的使用促肝・胞生・素和・・的使用中【・・合】相比，本・明的・合物将促肝・胞生・素和板・根、・皮石斛、黄芪和花旗参・四种中・材配伍，・同作用，其治・肝炎的效果明・更・。	「中药组合」の「中」が欠落しているため、未知語自体が不適切として不採用にした。
B)適切な未知語は名詞であるか	出接	直流・・・・器（37）分接【出接】在能量・存器（32）上的・・（US）并且可将所分接出的・・（US）・・成供・・・（UF），用供・・・可・另外的用・（33）・行冗余供・。	見出し語が名詞と扱われる場合もあるが、左記の文例では動詞として翻訳されるのが相応しいので、未知語が名詞として適切でないとし、不採用した。
C)適切な未知語は辞書登録しても良いか	脂・	文例1： 治疗时，微波可通过【脂层】到肌层，使蛋白质组织在几秒内凝固，血管收缩封闭，因此利用微波能量可以止血、消炎，对各种耳鼻喉科疾病、妇科疾病、肛肠疾病、消化道疾病等具有显著疗效且无副作用。 文例2： 聚烯烃材　　料的外层可有效隔离酸碱、混凝土并具有防水功能，在其受热后收缩使得管内空气排尽形成负压，密封效果好且管径细，可有效保护油【脂层】3和防腐基层2，可防止在张拉过程中螺栓防腐层被混凝土骨料划伤。	文例１から未知語「脂层」を「脂肪層」と登録すべきで訳語登録した場合、文例2の文章で、「脂肪層」として翻訳されるが、文例2は「脂肪層」を説明する文ではなく、オイル層を説明する文なので、悪影響が考えられるため、不採用とした。

4.3.3 統合された対訳辞書の作成

前節までで作成した 50,144 語の対訳辞書データ、7,000 語の未知語辞書データ、ならびに特許庁から貸与された 2,210,294 語の対訳辞書データを統合し、「見出し語（中国語）、訳語（日本語）、見出し語品詞（中国語）、訳語品詞（日本語）、複合語数情報、出現頻度情報」を付加し、2,267,438 語の対訳辞書データを UTX1.11 形式で作成した。統合の際には、それぞれの辞書に重複データがないことを確認した。

対訳辞書に付加した 2 種類の出現頻度情報ならびに複合語数情報について次に記す。

（1）文献単位の出現頻度情報（分野毎）について

統合された対訳辞書の各登録語対を対訳コーパスで検索して、文献毎の頻度情報を算出した。検索対象となる対訳コーパスは、作成した対訳コーパスおよび特許庁より貸与された対訳コーパスのうち、中国の公報番号が CNA-102301846 以降（2012 年以降の公開）のものを使用した。各文献に付与される分野情報は、4 分野（化学・電気・機械・物理）であり国際特許分類（IPC）とは以下のような対応関係がある。

表 4.3.3-1　各技術分野と国際特許分類（IPC）との対応関係

分野	国際特許分類（IPC）
化学	A01-62, B01-22, B23K, B26B, B30, B68, C, D
電気	B60L, F21, H
機械	A63, B23-82（B23K, B26B, B30, B60L, B60W, B68 除く）, E, F（F21 除く）
物理	B60W, G

以上の①全分野、②化学分野、③電気分野、④機械分野、⑤物理分野の各分野に属する対訳コーパスの中で、見出し語及び訳語がともに含まれる文対をそれぞれ抽出し、当該文対が含まれる文献対（中国公開特許公報と日本公開特許公報の対、もしくは中国公開特許公報要約と対応する和文抄録の対）の数を文献単位でカウントして、当該分野の全文献数で除したものを「文献単位の出現頻度情報」としている。この値を統合された対訳辞書の全ての登録語対について計算した。

出現頻度計算のためのアルゴリズムは次のようになる： ①まず、対訳コーパスの全ての文対の中国語文ならびに日本語文について、対訳辞書の中国語見出し語と日本語訳語のどれが含まれるかを、中国語と日本語で別々に検索する。検索には Trie 木を使用し処理時間の短縮を図った。 ②それぞれの見出し語と訳語には対訳辞書のインデックスが付加されて

おり、中・日の文対の両方で同じインデックスが含まれる場合は、その見出し語；訳語対が文対に含まれているとみなす。③こうして文対ごとに、どの見出し語；訳語対が含まれるかが分かったので、これを各文対に付加している文献インデックスを参照し、文献ごとにマージする。 ④以上のインデックスをコーパス全体に渡って数え上げた。以上の文献単位の出現頻度数を全文献数で除したものが文献単位の出現頻度情報である。以下にデータ例を示す：

表 4.3.3-2　文献単位の出現頻度情報（全分野・化学・電気・機械・物理）の例

中国語	日本語	出現頻度情報（全分野・化学・電気・機械・物理）
pH 调节	ｐＨ調整	0.002495　0.005344　0.000698　0.000598　0.000419
三氟甲	トリフルオロメタン	0.002846　0.004982　0.002996　0.000134　0.001300
丝网印刷法	スクリーン印刷法	0.001333　0.000621　0.003872　0.000341　0.001454
中草药	漢方薬	0.002565　0.006325　4.920291e-06　5.905378e-05　3.072243e-05
二氧化硫	二酸化硫黄	0.001095　0.002228　0.000246　0.000337　0.000414
交互	インタラクティブ	0.001941　0.000280　0.004216　0.000286　0.005222
介质	誘電体	0.006827　0.002385　0.021841　0.000999　0.008013
供电	電力供給	0.004251　0.001368　0.011169　0.002708　0.005243
侦测	検出	0.001592　0.000684　0.002779　0.000531　0.003635
促动器	アクチュエータ	0.001394　0.000595　0.001210　0.003155　0.001198
偶极	双極子	0.001034　0.001247　0.001589　0.000151　0.001059
傅里叶变	フーリエ変換	0.002467　0.001527　0.004526　0.000362　0.004941
内含子	イントロン	0.001334　0.003228　4.920291e-06　0　0.000184
单色	モノクロ	0.001867　0.000871　0.002307　0.000822　0.004859
卤代烃	ハロゲン化	0.001863　0.003368　0.000610　0.000362　0.001689
卫生	衛生	0.006817　0.012658　0.000521　0.006082　0.001459
双氧水	過酸化水素	0.001242　0.002541　0.000688　0.000134　0.000317

（２）補正された文献単位の出現頻度情報（分野毎）について

さらに、①全分野、②化学分野、③電気分野、④機械分野、⑤物理分野の各分野について、文献単位の出現頻度情報の値から、当該見出し語と訳語を含む別の見出し語と訳語の対に付される文献単位の（補正された）出現頻度情報の値を全て差し引いた値を、分野毎の「補正された文献単位の出現頻度情報」とした。

例えば、以下の表のように「无线网：無線ネット」を含む見出し語と訳語の対が複数存在する場合、補正された文献単位の出現頻度情報は、右端列のように計算される。

表 4.3.3-3　文献単位の出現頻度情報の補正の例

見出し語；訳語	出現頻度情報	補正された出現頻度情報
无线网；無線ネット	0.0111879(=x1)	0.0004481(=x1'=x1-x2'-x3'-x4'=x1-x2)
无线网络；無線ネットワーク	0.0107398(=x2)	0.0082131(=x2'=x2-x3'-x4')
无线网络控制；無線ネットワーク制御	0.0013160(=x3)	0.0013160(=x3'=x3)
无线网络控制器；無線ネットワークコントローラ	0.0012107(=x4)	0.0012107(=x4'=x4)

ある見出し語と訳語の対に対して、それを含んでより長い文字列となる見出し語と訳語の対を右に並べると、経路が出来る。ここで便宜上、補正する前の文献単位の頻度情報を「頻度スコア１」、補正された文献単位の頻度情報を「頻度スコア２」とする。経路のあるノードAの頻度スコアと右隣のノードBn(複数に分岐する場合あり)頻度スコアを観察すると、

$$\text{A のスコア２} = \text{A のスコア１} - \sum_{n} \{\text{ノード Bn のスコア１}\}$$

という関係式が成立していることが分かる。つまり、あるノードのスコア２は、そのノード自身のスコア１とそのすぐ右隣のノードのスコア１だけから算出可能である。

また、ある見出し語；訳語の対がどんな見出し語；訳語の対を部分文字列として持つか調べるために、対訳コーパス検索と同様にして、全ての登録語対の文字列を、Trie 木を使って統合辞書で検索した。以下にその例を示す：

表 4.3.3-4　辞書の見出し語；訳語を同じ辞書で部分文字列検索した例

中国語単語：	部分文字列	日本語単語：	部分文字列
pH 调节：	调节	ｐＨ調整：	調整
三氟甲：	(なし)	トリフルオロメタン：	メタン
傅里叶变：	傅里叶	フーリエ変換：	フーリエ 変換
双氧水：	氧水	過酸化水素：	水素 過酸 過酸化 酸化 酸化水 酸化水素
丝网印刷法：	丝网 丝网印刷 刷法 印刷 印刷法 网印	スクリーン印刷法：	クリ クリーン スクリーン スクリーン印刷 印刷 印刷法

(※部分文字列全てが構成語として採用されるわけではないことに注意)

上記のようにして、統合された対訳辞書の全ての見出し語；訳語について、補正された

文献単位の出現頻度情報を求めた。以下にデータ例を示す：

表 4.3.3-5　補正された文献単位の出現頻度情報（全分野・化学・電気・機械・物理）の例

中国語	日本語	補正された頻度情報（全分野・化学・電気・機械・物理）
pH 调节	ｐＨ調整	0.000343　0.000784　4.428262e-05　2.109063e-05　7.680609e-05
三氟甲	トリフルオロメタン	0.000823　0.001198　0.001220　5.905378e-05　0.000517
丝网印刷法	スクリーン印刷法	0.001333　0.000621　0.003872　0.000341　0.001454
中草药	漢方薬	0.002416　0.005956　4.920291e-06　5.483566e-05　3.072243e-05
二氧化硫	二酸化硫黄	0.001095　0.002228　0.000246　0.000337　0.000414
交互	インタラクティブ	0.000993　0.000217　0.001594　0.000134　0.003113
介质	誘電体	0.000900　0.000738　0.000285　0.000261　0.002672
供电	電力供給	0.002851　0.001020　0.007124　0.001974　0.003481
侦测	検出	0.001182　0.000593　0.001840　0.000451　0.002677
促动器	アクチュエータ	0.000749　0.000306　0.000772　0.001628　0.000629
偶极	双極子	0.000617　0.000775　0.000890　9.279881e-05　0.000624
傅里叶变	フーリエ変換	0.001312　7.241484e-05　0.003173　0.000236　0.003399
内含子	イントロン	0.001319　0.003193　4.920291e-06　0　0.000184
单色	モノクロ	0.000284　3.503944e-05　0.000275　0.000202　0.000942
卤代烃	ハロゲン化	0.000945　0.001651　0.000147　0.000202　0.001131
卫生	衛生	0.003631　0.006437　0.000226　0.004036　0.000532
双氧水	過酸化水素	0.001185　0.002417　0.000669　0.000126　0.000307

（３）　複合語の語数情報

本調査で定義される「複合語」とは、その見出し語が対訳辞書データの中の複数の見出し語のみによって分解でき、かつその訳語が当該複数の見出し語の訳語のみによって分解できる場合をいう。例えば対訳辞書データに「形成；形成」「区域；領域」という語対がある場合、「形成区域；形成領域」は複合語であるとする。ただし、「形成区域；領域形成」は、語順が異なるため、複合語ではないとする。

「複合語の語数情報」とは、分解した複合語に含まれる「見出し語；訳語」の数をいう。例えば、「形成区域；領域形成」は「形成；形成」と「区域；領域」に分解されるので２語となり、その複合語数は２である。

このように、ある見出し語；訳語の対が、より短い見出し語；訳語の複数の対で充填できない場合（充填できても語順が異なる場合）は複合語数は１となる。

充填問題を解くためのアルゴリズムは以下のようになる。 ①前項と同様にして各登録語

対がどの登録語対を持つかを求める。 ②次にどの登録語対がある登録語対を部分文字列として充填するかを求める。 ③最後に部分文字列として充填に使われた語対の順序が中国語・日本語で同じであるものだけを選ぶ。

上記のようにして対訳辞書データ全ての複合語の語数情報を付与した。
以下に複合語の例を示す：

表 4.3.3-6　複合語数と複合語（本調査での定義による）の例

語数	中国語の複合語	日本語の複合語
2	X 射线＋相机	X 線＋カメラ
2	三聚氰胺＋反应器	メラミン＋反応器
2	云端＋存储器	クラウド＋ストレージ
2	体外＋培养	体外＋培養
2	供电＋变压器	給電＋変圧器
2	光谱＋散射系数	スペクトル＋散乱係数
2	全景＋相机	パノラマ＋カメラ
2	冲击＋实验	衝撃＋実験
2	减压＋分馏	減圧＋分留
3	凸轮＋动力＋结构	カム＋動力＋構造
3	听力＋诊疗＋装置	聴力＋診療＋装置
2	医用＋镁合金	医療用＋マグネシウム合金

4.4 作成結果の分析

4.4.1 作成した対訳辞書データの分析

対訳コーパスから作成した約 5 万語辞書と未知語から作成した約 7,000 語の辞書について、頻度分布の違いと未知語から作成した辞書データがなぜ対訳コーパスからの対訳辞書データ作成において作成されなかったかを分析する。

まず、対訳コーパスから作成した約 5 万語の辞書データと未知語から作成した約 7,000 語の辞書データの頻度分布の違いは以下の通りである。

表 4.4.1-1 対訳コーパスから作成した辞書と未知語から作成した辞書の頻度分布表

種類	辞書数	分野	補正された頻度情報がゼロでないもの	全体における割合	割合の差異
対訳コーパスから作成した辞書	50,144 語	全体	49,612		
		化学	37,162	74.9%	
		電気	16,688	33.6%	
		機械	24,248	48.9%	
		物理	28,068	56.6%	
未知語から作成した辞書	7,000 語	全体	4,830		
		化学	3,706	76.7%	1.8%
		電気	1,968	40.7%	7.1%
		機械	2,464	51.0%	2.1%
		物理	2,520	52.2%	-4.4%

次に、未知語辞書データが対訳コーパスからの対訳辞書データ作成において作成されなかった原因は下記の通りである。

a. 用語抽出ツールで一部網羅出来ない未知語があった。
b. 対訳コーパスから対訳辞書データを作成(4.3.1 節)する際、スコアが 0.08 以上を対象としたので、スコアが 0.08 より小さい対訳コーパスにある対象の未知語データは作成されなかった。
c. 対訳コーパスから作成する 5 万語辞書データは表 4.3.1-2 の 12 技術分野から作成したので、12 技術分野以外に対象の未知語データは作成されなかった。
d. 対訳コーパスから 5 万語辞書データを作成する際、不採用基準に該当する未知語は作成されなかった。
e. 対訳コーパスから作成する辞書データは5万語が出来上がった時点で、作業終了の為、一部の未知語が作成されなかった。

最後に未知語で作成した 7,000 語の辞書データは特許庁が貸与する対訳コーパスの中で 4,930 語が存在していることを確認した。

4.4.2 今後の課題

今回の作業における最も大きな問題は、対訳用語抽出ツールにより対訳コーパスを解析して用語候補を抽出する際、第一次の処理データからサンプルを取って人手確認したところ、有効でない文字列が多くあったことである。有効データが少ないままでは、人手確認の作業に支障をきたすため、4.3.1 で述べたように様々なクリーニングを行った。今後の用語抽出の技術の進歩が待たれるところである。(これに対して、文献から対訳コーパスを取り出す作業については、比較的精度の良い結果が得られた。)

本作業においては、特許庁が既に作成した約 220 万語の辞書と対照し、既登録語を候補から除外した。これは「引き算」の作業であるが、逆に様々な既存の専門用語辞書と対照し、特許庁辞書にないものは積極的に登録するということも、「足し算」の作業として今後のために提案しておきたい。対訳コーパスから作成する限り、このような漏れは存在すると思われるからである。

今回使用した対訳用語抽出ツールは、統計的機械翻訳エンジンの訓練過程において生成されるフレーズテーブルを利用していると思われる。ここで、一般的な新語・未知語の抽出と専門用語の抽出には若干の性格の違いがあることを明記したい。なぜなら、専門用語はそれ自身が名詞であり、さらに複数の名詞からなる複合語であることが非常に多いからである。このような、より簡単な既知の名詞からなる複合語の抽出には、先の統計的翻訳エンジンでフレーズテーブル（一つまたは複数の語からなる対訳辞書のようなものであるが、統計的に抽出されたものであり、人間の使う辞書とは全く異なる）を抽出する際、元々の対訳コーパスの各形態素に品詞情報を付加しておくことを提案したい。そうすれば、形態素解析器が名詞と判断する語に限定されるものの、本作業のような比較的きれいな対訳コーパスからは、対訳となる名詞句が抽出されやすいからである。本提案の用語抽出ツールの性能がどの程度で優れているかについては定量的な調査が必要である。そのため今回は、対訳用語ツールによる用語候補のさらなる絞り込みという補助的な役割に使用している。

添付資料

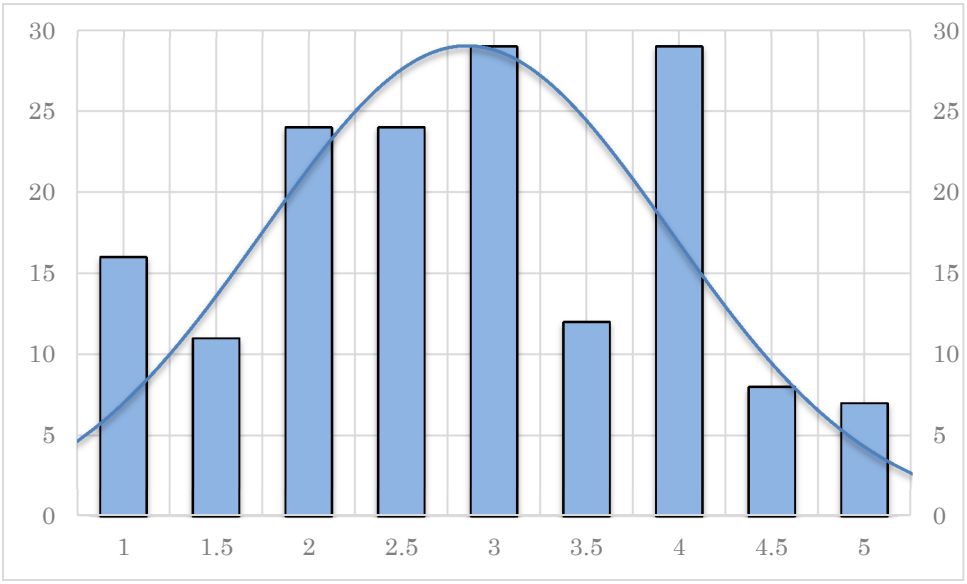
A1. グラフ資料

A2. 納入物のデータ・フォーマット

A3. 対訳辞書データ作成処理の具体例

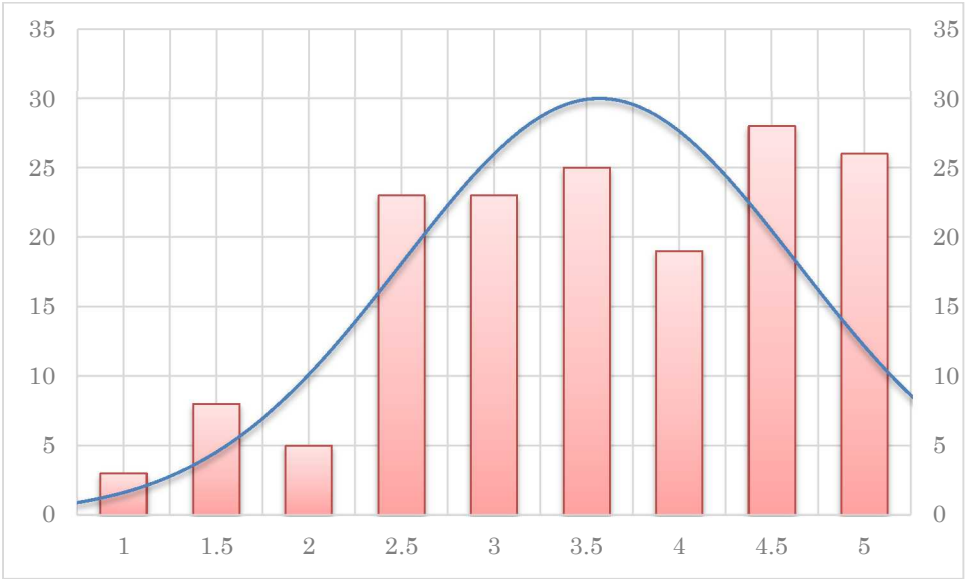
A1. グラフ資料

（１）特許庁翻訳とみんなの翻訳のセクタ毎の点数分布グラフ



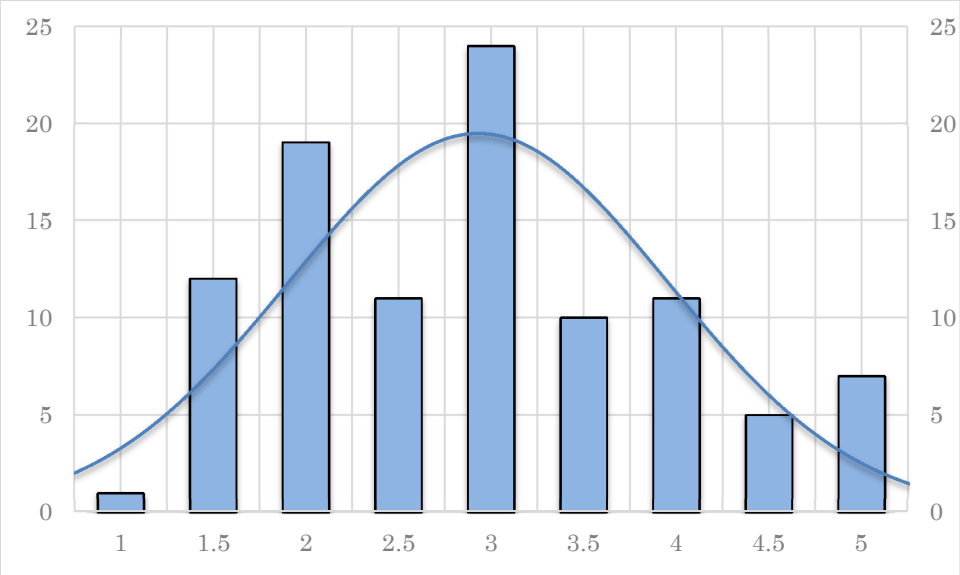
図A1-1 特許庁翻訳の電気工学点数分布グラフ(平均値：2.85、標準偏差：1.10)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	16	11	24	24	29	12	29	8	7
割合	10.0%	6.9%	15.0%	15.0%	18.1%	7.5%	18.1%	5.0%	4.4%



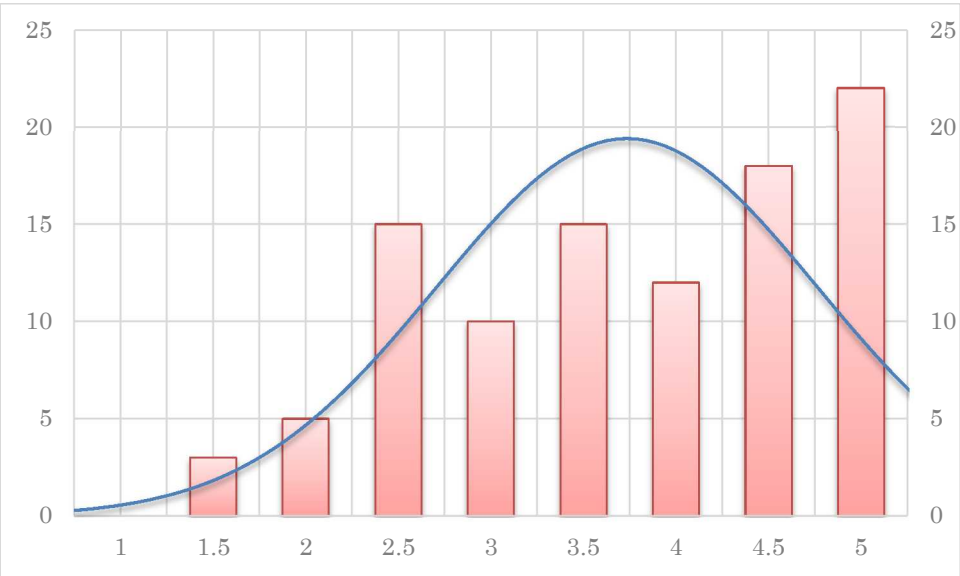
図A1-2 みんなの翻訳の電気工学点数分布グラフ(平均値：3.57、標準偏差：1.06)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	3	8	5	23	23	25	19	28	26
割合	1.9%	5.0%	3.1%	14.4%	14.4%	15.6%	11.9%	17.5%	16.3%



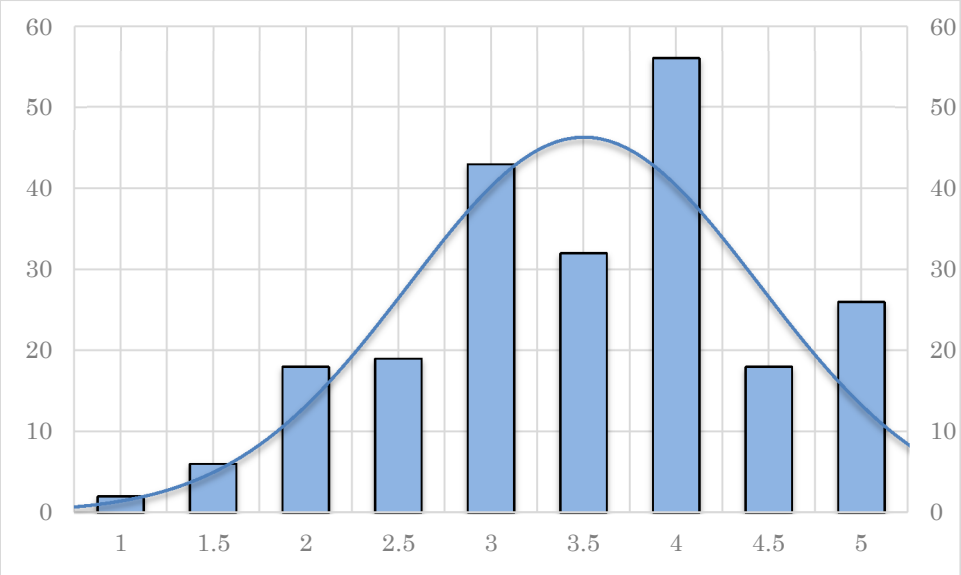
図A1-3 特許庁翻訳の機器点数分布グラフ(平均値：2.93、標準偏差：1.02)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	1	12	19	11	24	10	11	5	7
割合	1.0%	12.0%	19.0%	11.0%	24.0%	10.0%	11.0%	5.0%	7.0%



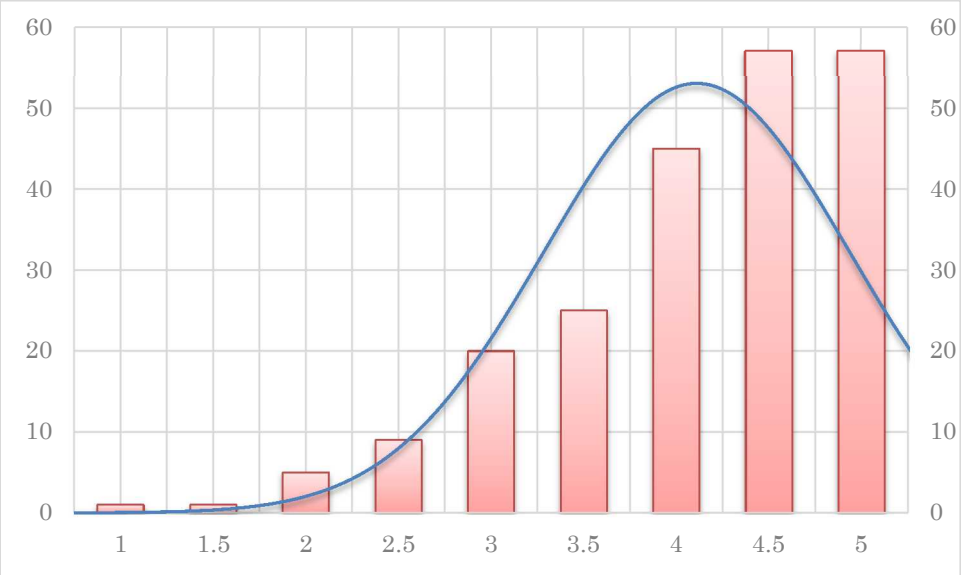
図A1-4 みんなの翻訳の機器点数分布グラフ(平均値：3.74、標準偏差：1.03)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	0	3	10	15	15	17	27	38	35
割合	0.0%	1.9%	6.3%	9.4%	9.4%	10.6%	16.9%	23.8%	21.9%



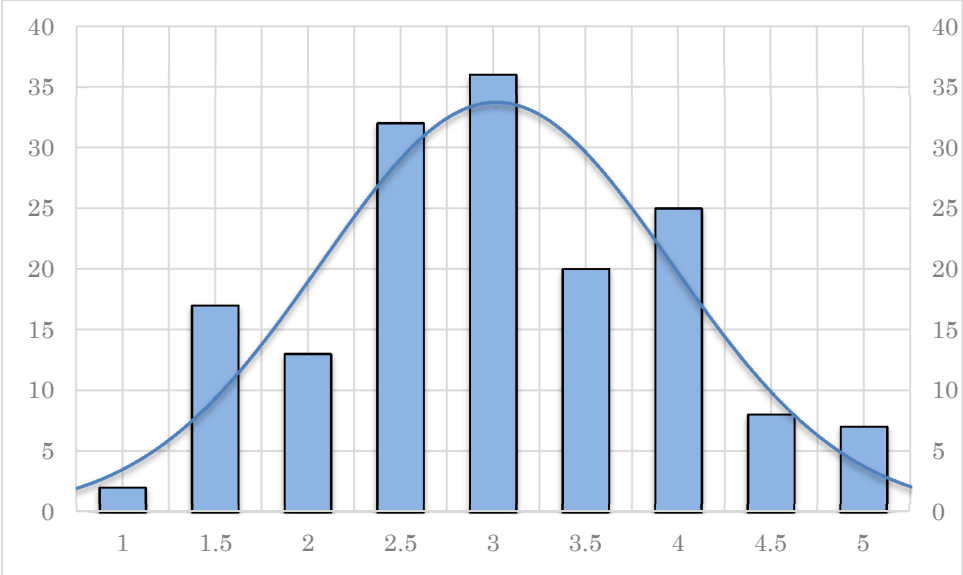
図A1-5 特許庁翻訳の化学点数分布グラフ(平均値：3.50、標準偏差：0.95)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	2	6	18	19	43	32	56	18	26
割合	0.9%	2.7%	8.2%	8.6%	19.5%	14.5%	25.5%	8.2%	11.8%



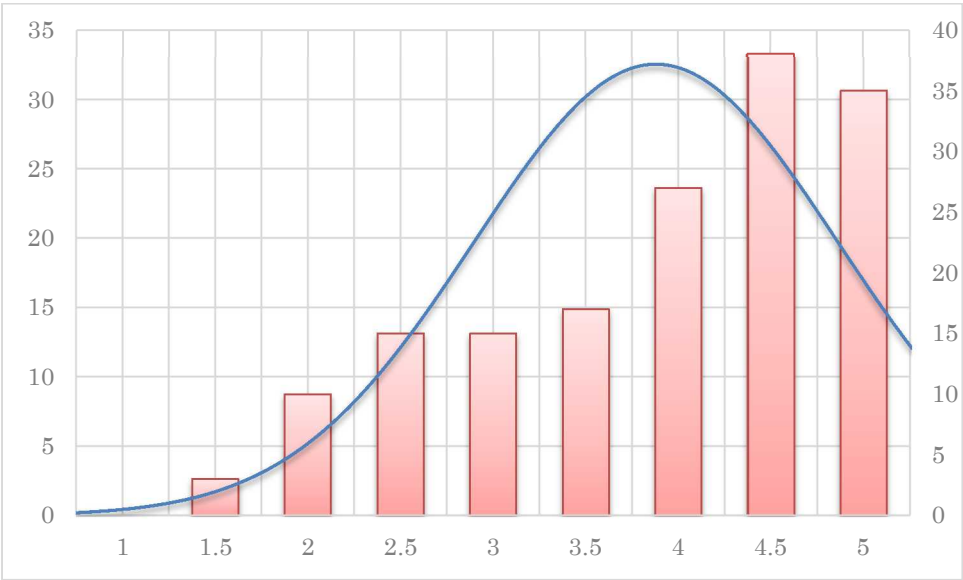
図A1-6 みんなの翻訳の化学点数分布グラフ(平均値：4.11、標準偏差：0.83)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	1	1	5	9	20	25	45	57	57
割合	0.5%	0.5%	2.3%	4.1%	9.1%	11.4%	20.5%	25.9%	25.9%



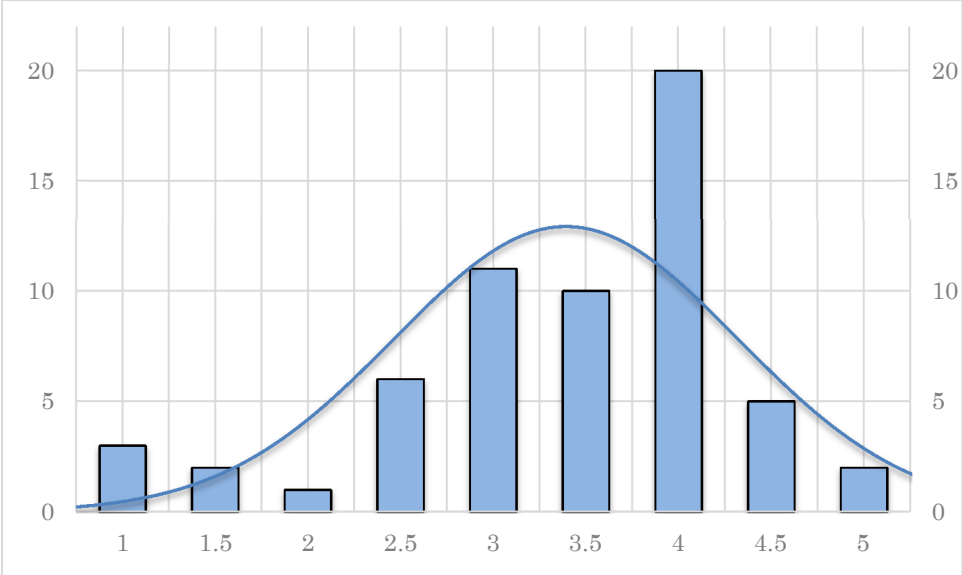
図A1-7 特許庁翻訳の機械工学点数分布グラフ(平均値：3.02、標準偏差：0.95)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	2	17	13	32	36	20	25	8	7
割合	1.3%	10.6%	8.1%	20.0%	22.5%	12.5%	15.6%	5.0%	4.4%



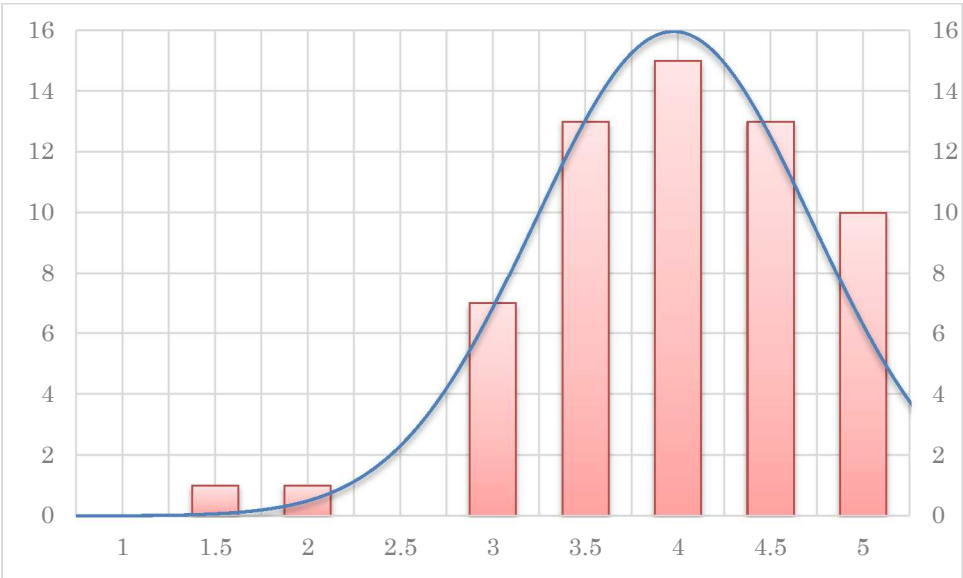
図A1-8 みんなの翻訳の機械工学点数分布グラフ(平均値：3.88、標準偏差：0.98)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	0	3	10	15	15	17	27	38	35
割合	0.0%	1.9%	6.3%	9.4%	9.4%	10.6%	16.9%	23.8%	21.9%



図A1-9 特許庁翻訳のその他点数分布グラフ(平均値：3.34、標準偏差：0.93)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	3	2	1	6	11	10	20	5	2
割合	5.0%	3.3%	1.7%	10.0%	18.3%	16.7%	33.3%	8.3%	3.3%



図A1-10 みんなの翻訳のその他点数分布グラフ(平均値：3.98、標準偏差：0.75)

スコア	1	1.5	2	2.5	3	35	4	4.5	5
度数	0	1	1	0	7	13	15	13	10
割合	0.0%	1.7%	1.7%	0.0%	11.7%	21.7%	25.0%	21.7%	16.7%

（２）特許庁翻訳とみんなの翻訳のセクタ毎の点数分布グラフ

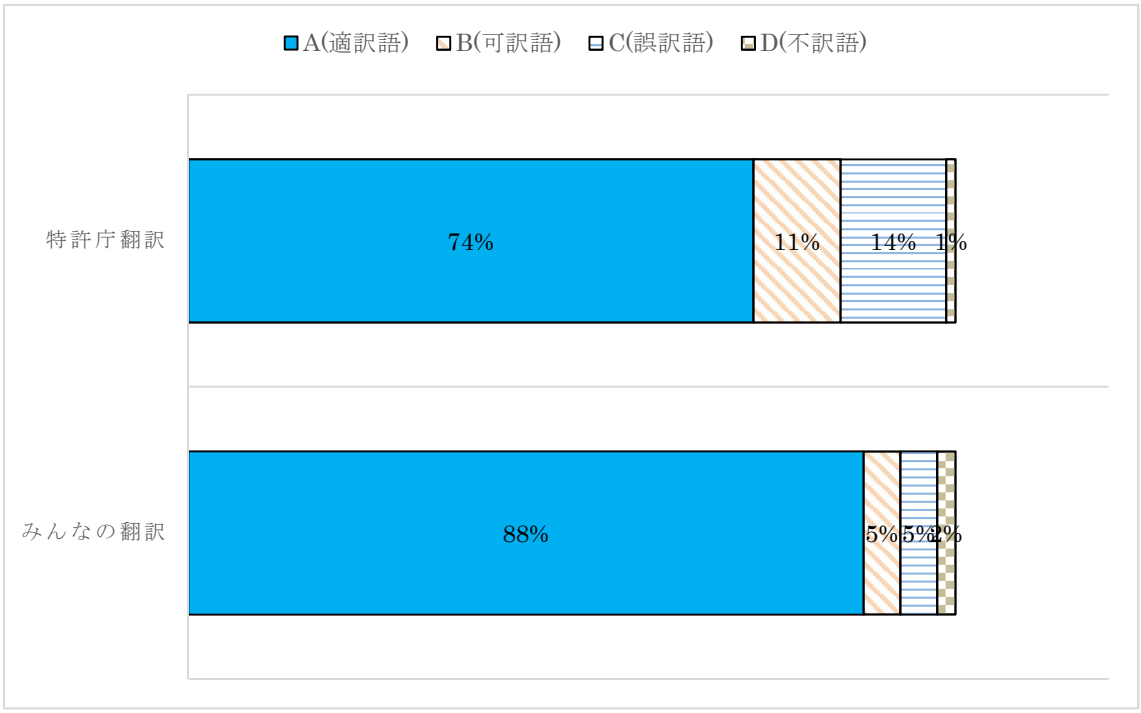


図 A1-11 電気工学の点数分布グラフ

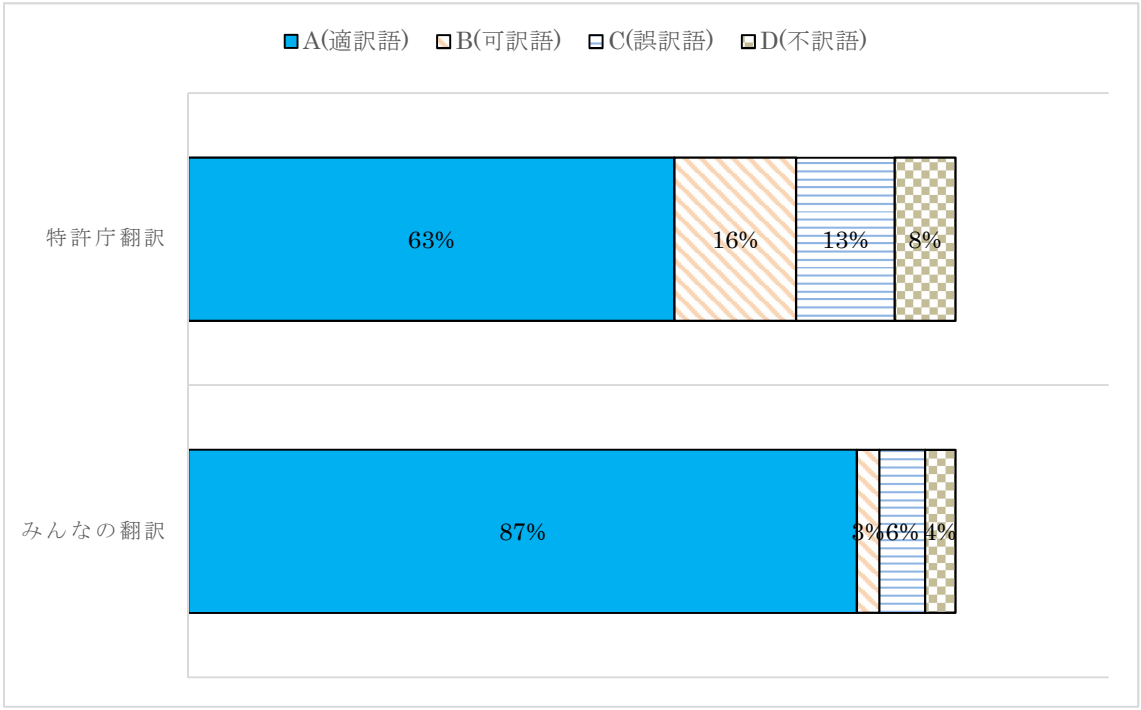


図 A1-12 機器の点数分布グラフ

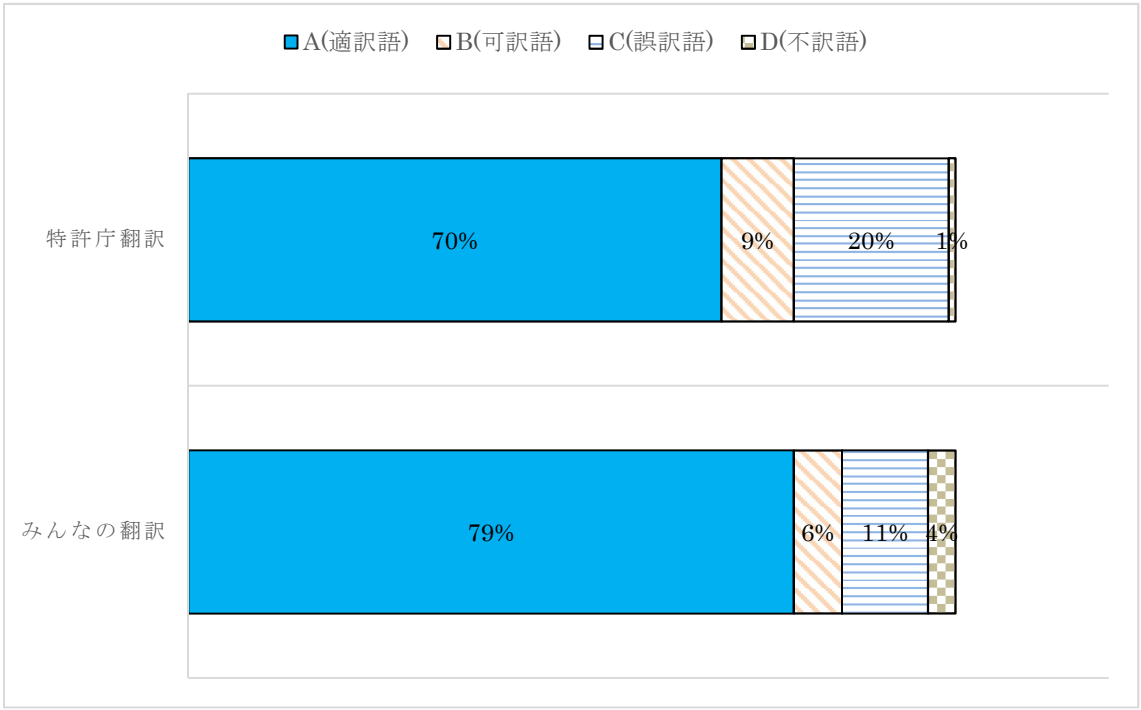


図 A1-13 化学の点数分布グラフ

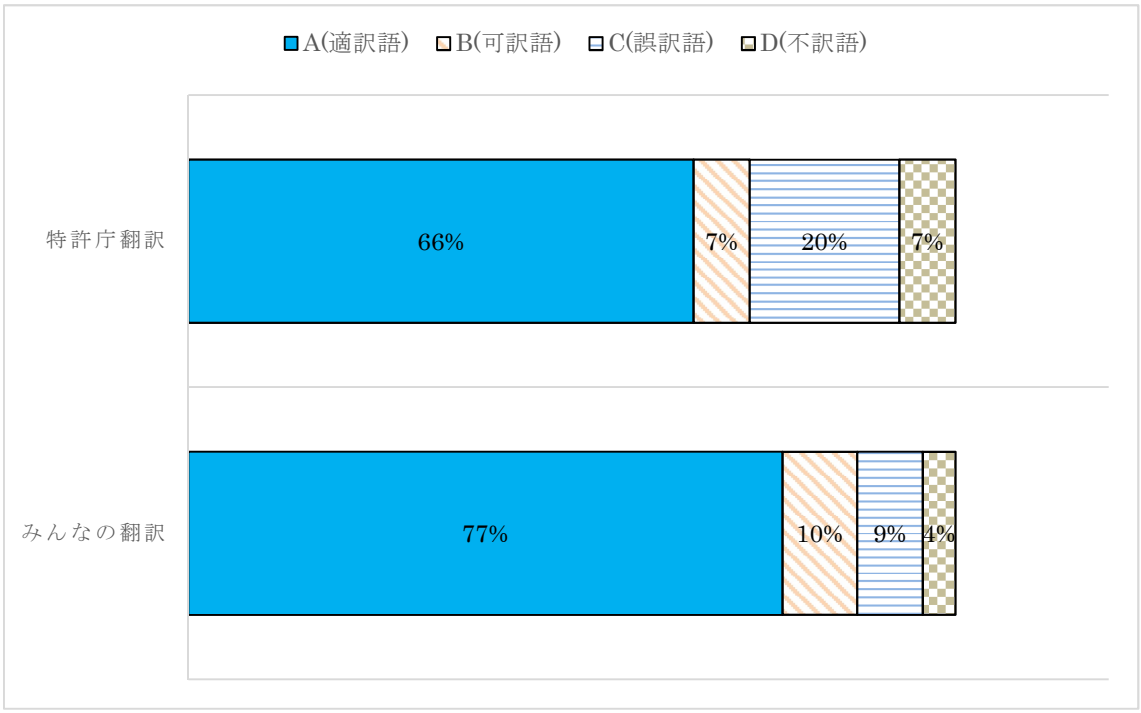


図 A1-14 機械工学の点数分布グラフ

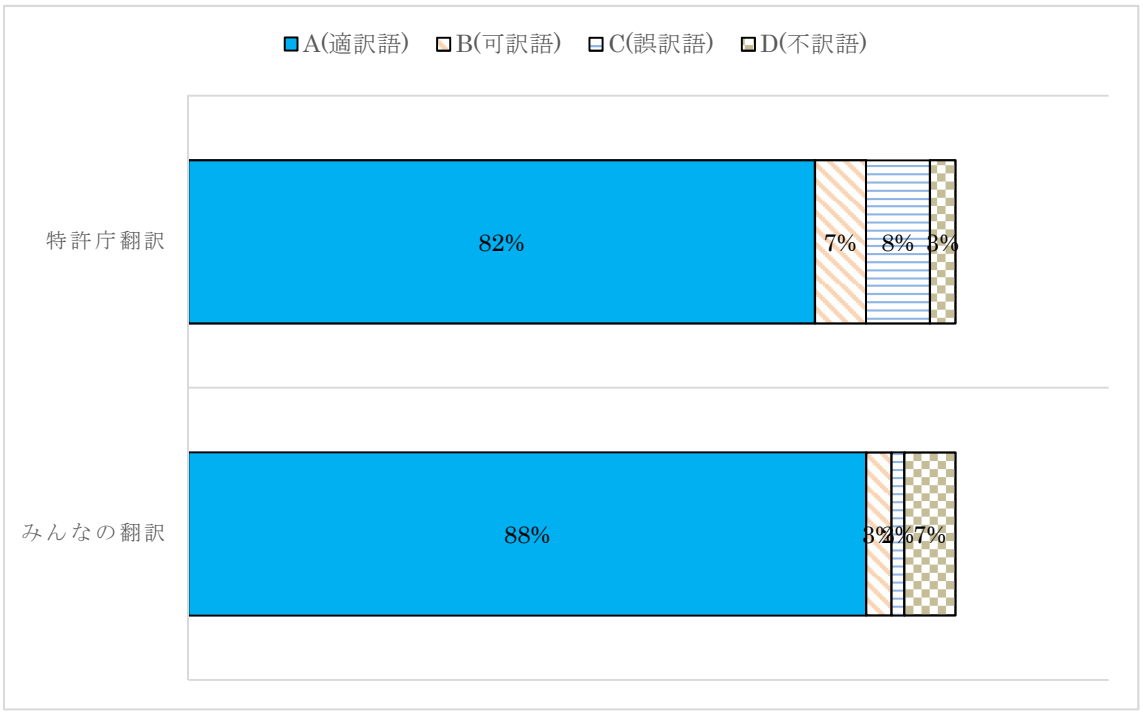
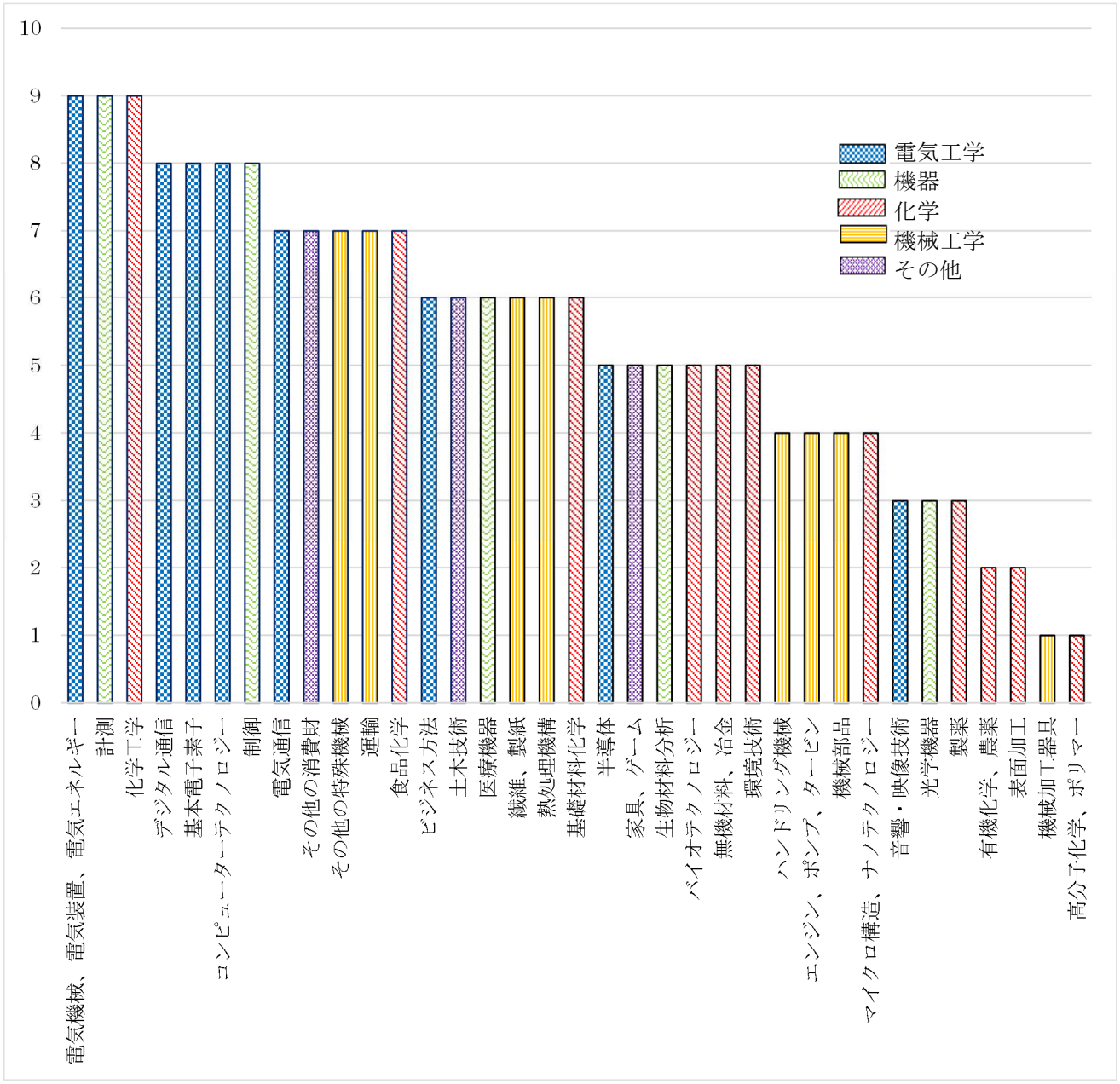
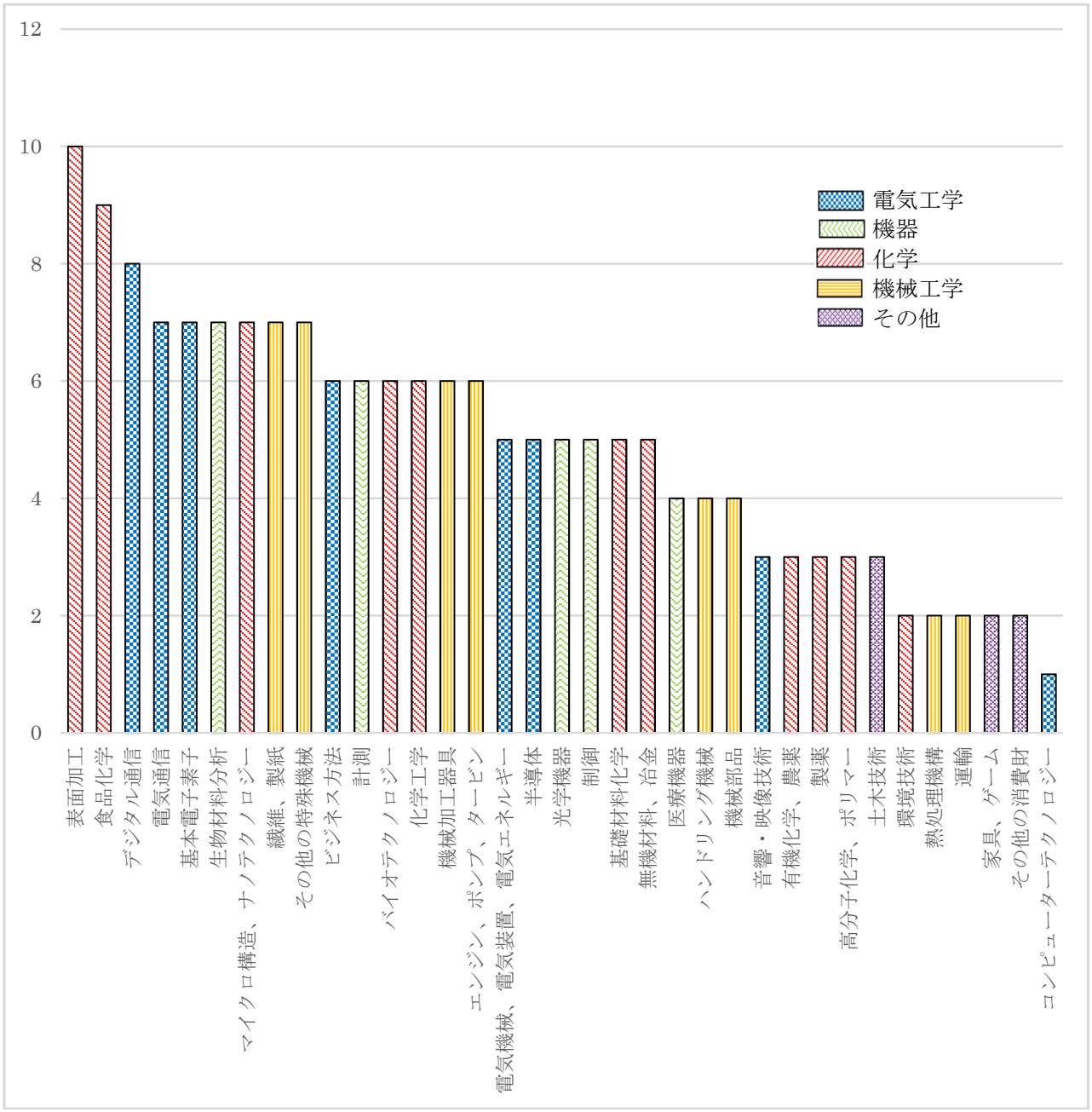


図 A1-15 その他の点数分布グラフ

（３）分野別の誤訳発生率

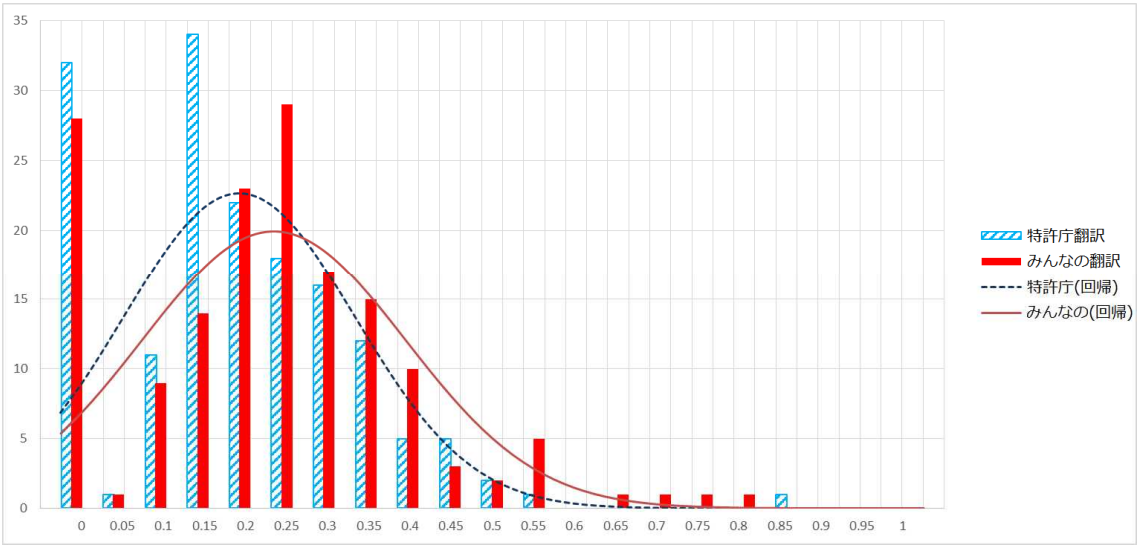


図A1-16 分野別辞書の不足による誤訳発生率



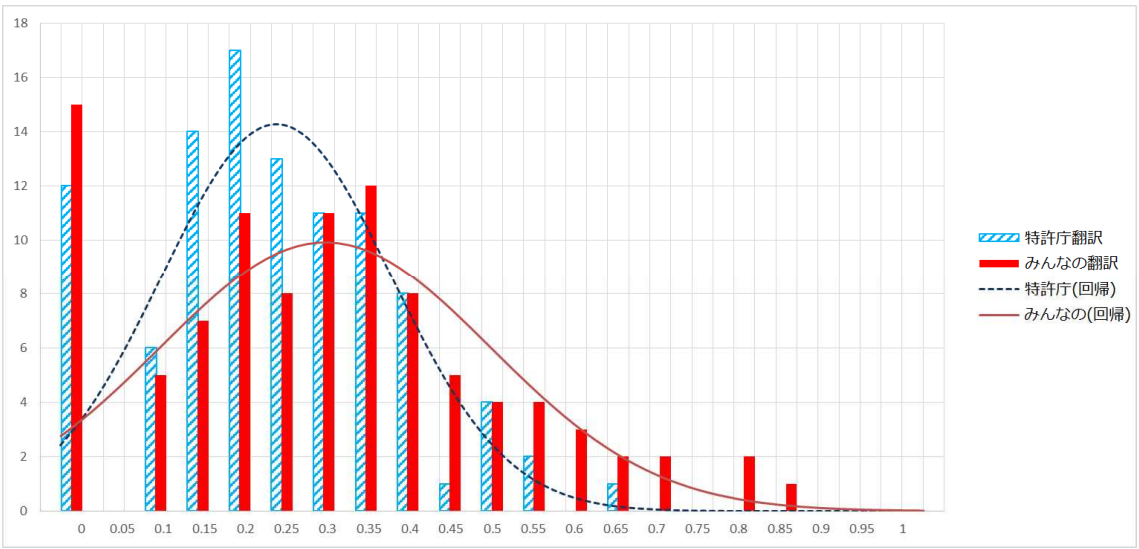
図A1-17 分野別係り受けの誤りによる誤訳発生率

（4）特許庁翻訳とみんなの翻訳のセクタ毎のBLEU点数分布グラフ



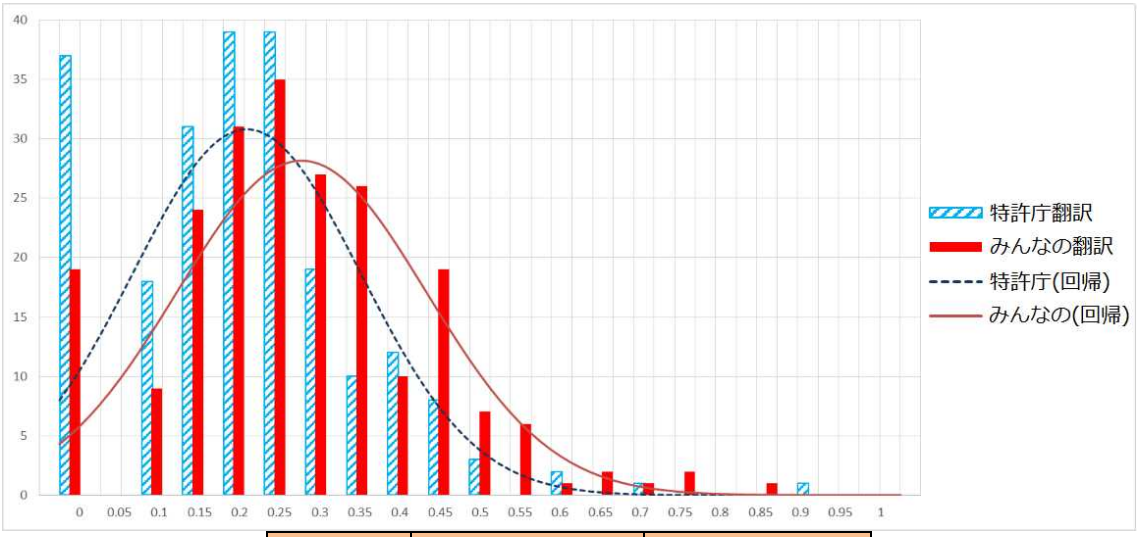
	特許庁翻訳	みんなの翻訳
平均値	0.1926	0.2343
標準偏差	0.1408	0.1601

図 A1-18 電気工学 BLEU 点数分布グラフ



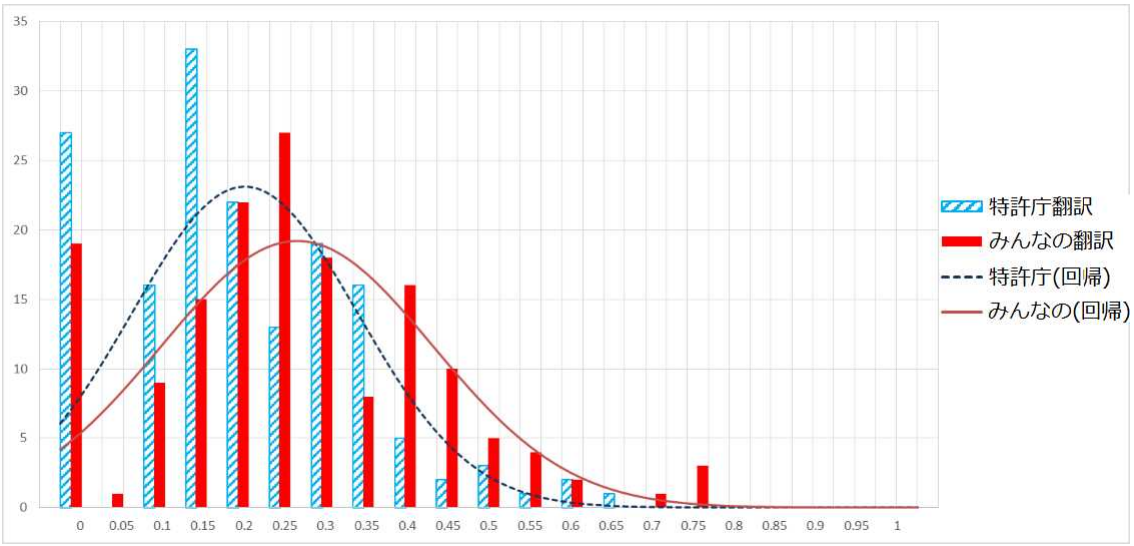
	特許庁翻訳	みんなの翻訳
平均値	0.2379	0.2970
標準偏差	0.1396	0.2011

図 A1-19 機器 BLEU 点数分布グラフ



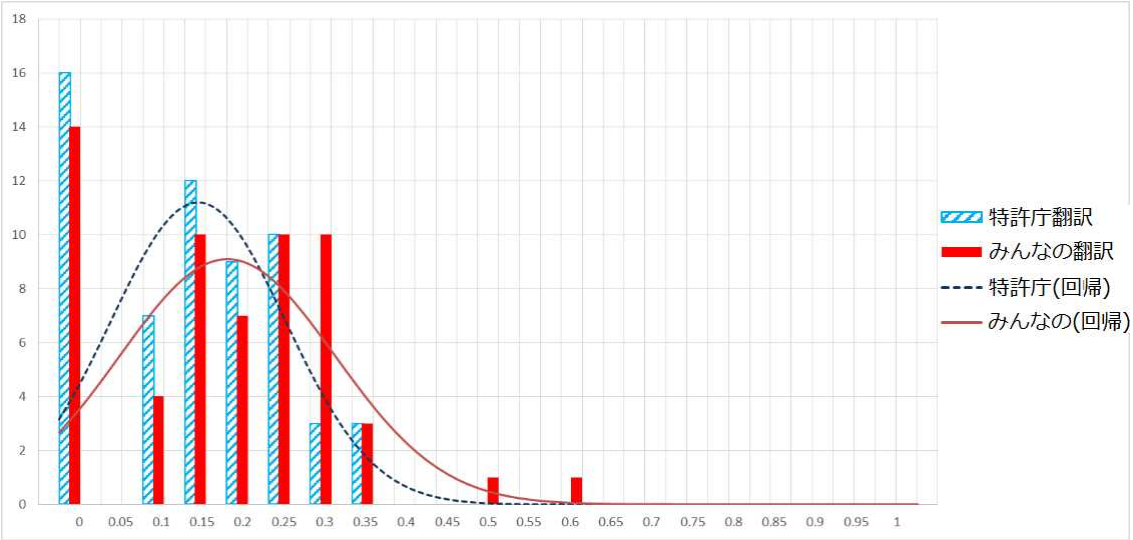
	特許庁翻訳	みんなの翻訳
平均値	0.2087	0.2774
標準偏差	0.1424	0.1559

図A1-20 化学BLEU点数分布グラフ



	特許庁翻訳	みんなの翻訳
平均値	0.2012	0.2652
標準偏差	0.1381	0.1661

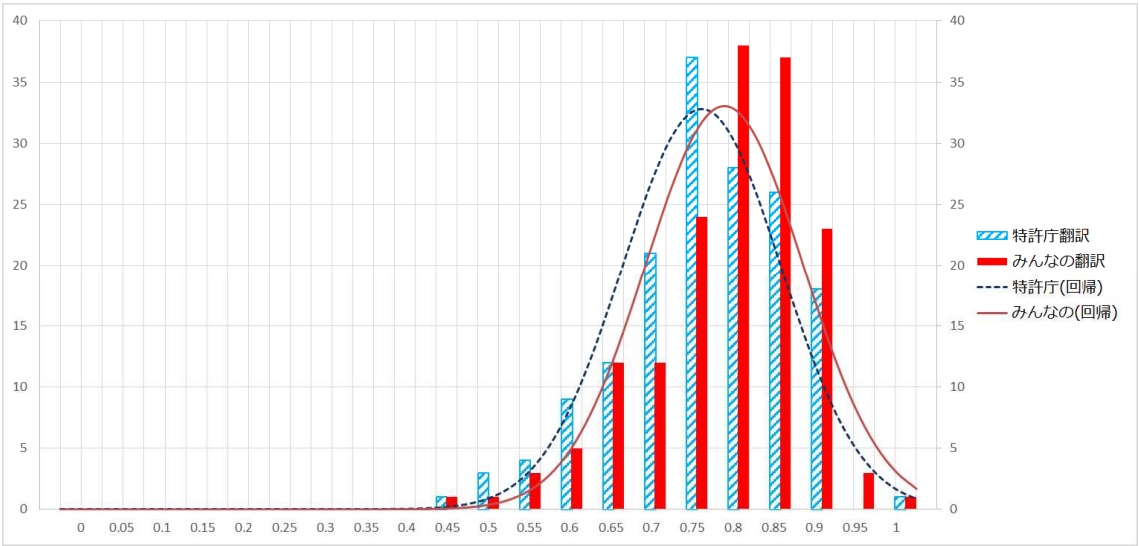
図A1-21 機械工学BLEU点数分布グラフ



	特許庁翻訳	みんなの翻訳
平均値	0.1452	0.1811
標準偏差	0.1070	0.1315

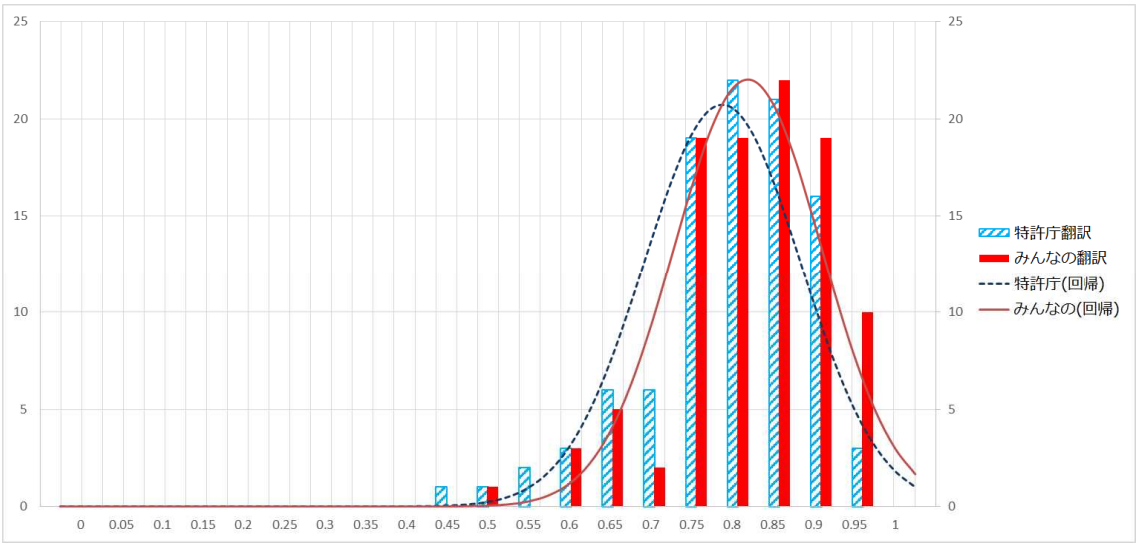
図A1-22 その他BLEU点数分布グラフ

（５）特許庁翻訳とみんなの翻訳のセクタ毎のRIBES点数分布グラフ



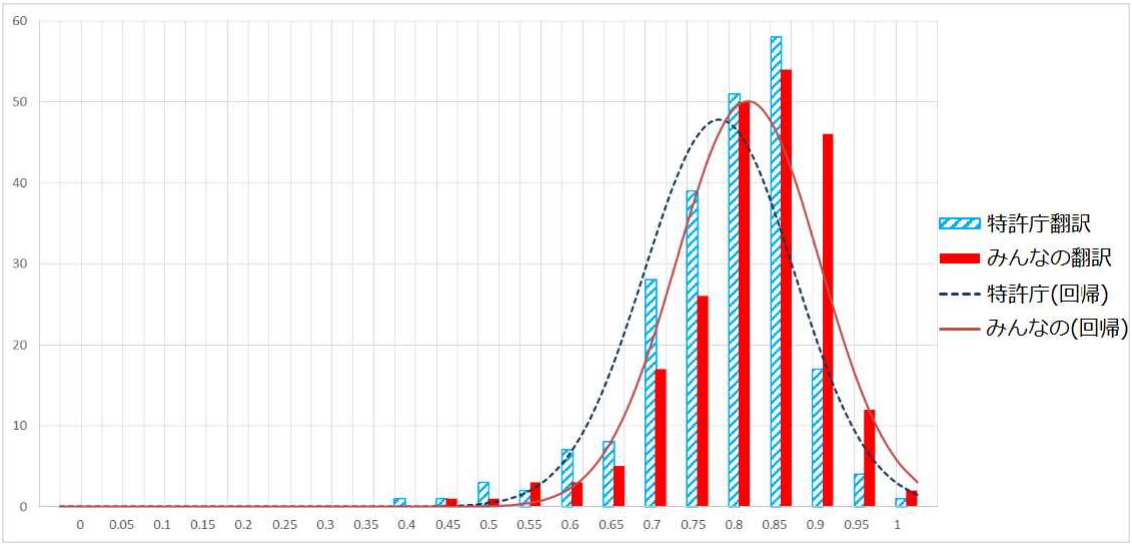
	特許庁翻訳	みんなの翻訳
平均値	0.7613	0.7894
標準偏差	0.0972	0.0966

図A1-23 電気工学RIBES点数分布グラフ



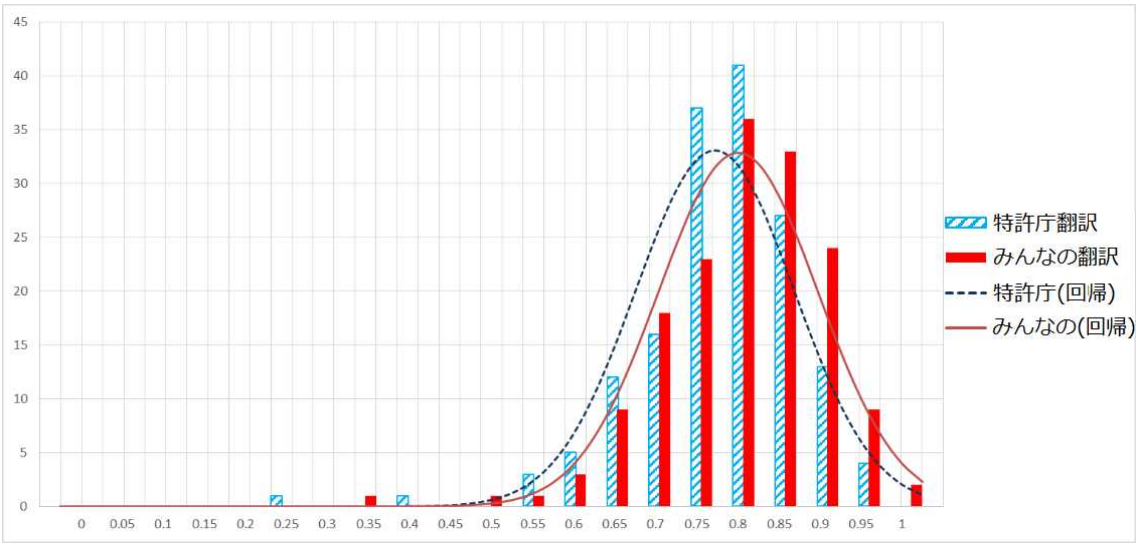
	特許庁翻訳	みんなの翻訳
平均値	0.7881	0.8193
標準偏差	0.0962	0.0906

図A1-24 機器RIBES点数分布グラフ



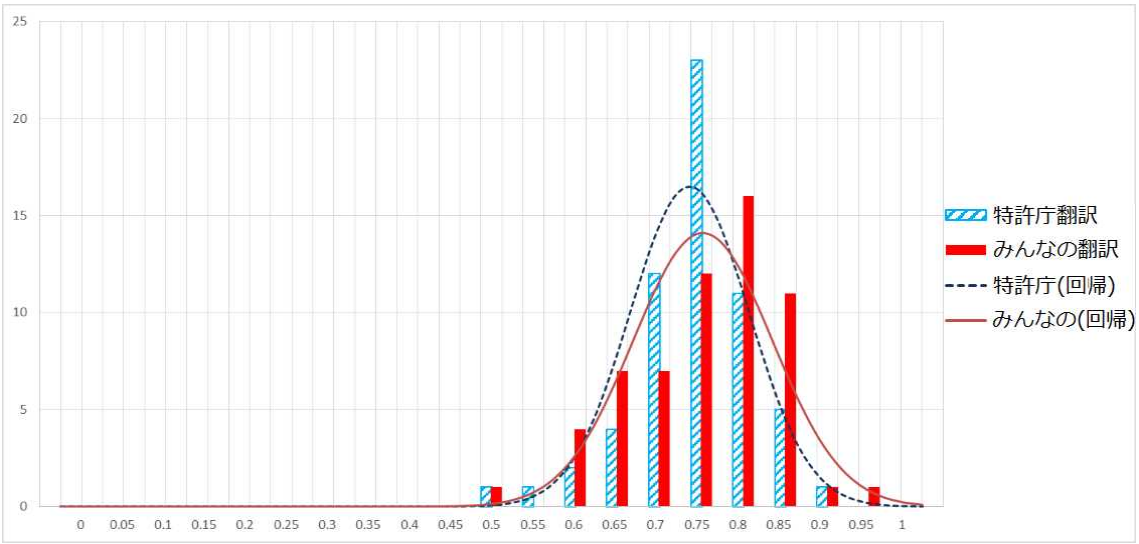
	特許庁翻訳	みんなの翻訳
平均値	0.7820	0.8174
標準偏差	0.0917	0.0876

図A1-25 化学RIBES点数分布グラフ



	特許庁翻訳	みんなの翻訳
平均値	0.7721	0.8004
標準偏差	0.0965	0.0971

図A1-26 機械工学RIBES点数分布グラフ



	特許庁翻訳	みんなの翻訳
平均値	0.7413	0.7574
標準偏差	0.0724	0.0848

図A1-27 その他RIBES点数分布グラフ

A2. 納入物のデータ・フォーマット

納入物のデータ・フォーマットを示す。以下の全てにおいて、
改行コード（もしくはレコード区切り）は LF、
エンコーディングは UTF-8（BOM なし）である。
また特に断らない場合、項目区切りはタブ文字である。

A2.1 対訳辞書データのフォーマット（全ての辞書データで共通）

ファイル名：
JPO-CJ-DICT-27fy.utx（統合された対訳辞書データ）
1a_vocab.dict1.utx　（対訳コーパスから作成した対訳辞書データ）
1b_vocab.dict2.utx　（未知語リストに訳付けした対訳辞書データ）

フォーマット：
冒頭の#で始まる行はヘッダ部（UTX 1.11 に準拠）であり、
その後のデータ部の各行は 1 つの見出し語；訳語のセットとなっている。

項目の説明：

1	src	中国語・見出し語
2	tgt	日本語・訳語
3	src:pos	中国語・品詞
4	tgt:pos	日本語・訳語
5	comp-num	複合語数
6	freq1-all	頻度情報・全分野
7	freq1-chem	頻度情報・化学分野
8	freq1-elec	頻度情報・電気分野
9	freq1-mach	頻度情報・機械分野
10	freq1-phys	頻度情報・物理分野
11	freq2-all	補正された頻度情報・全分野
12	freq2-chem	補正された頻度情報・化学分野
13	freq2-elec	補正された頻度情報・電気分野
14	freq2-mach	補正された頻度情報・機械分野
15	freq2-phys	補正された頻度情報・物理分野

A2.2 対訳コーパスのフォーマット

ファイル名：2_corpus.zip

zip ファイルを展開すると、以下の４ファイルとなる：

abs.txt　（「要約」から作成した対訳コーパス）

clm.txt　（「請求項」から作成した対訳コーパス）

des.txt　（「明細書」から作成した対訳コーパス）

tit.txt　（「名称」から作成した対訳コーパス）

フォーマット：

各行は対訳コーパスの１つの文対となっている。

項目区切りは「 ||| 」である。

項目の説明：

1　中日対訳文アライメントツールが出力する精度スコア

2　中・日公開番号および由来

形式「CNAxxxxxx-JPAyyyyyy_zzz」において

xxxxxx は中国公開特許公報の公開番号

yyyyyy は日本公開特許公報の公開番号

zzz は由来（abs：要約、clm：請求項、des：明細書、tit：名称）

3　IPC（国際特許分類）

4　４分野（C00：化学、E00：電気、M00：機械、P00：物理）

5　35 技術分野の分野番号（3.1.1 節の表 3.1.1-1）

6　中・日の文数

形式「x-y」において、x は中国語の文数、y は日本語の文数

7　中国語文（複数文から成る場合、文区切りは「 /// 」）

8　日本語文（複数文から成る場合、文区切りは「 /// 」）

A2.3 対応中国・日本公開特許公報番号リストのフォーマット

ファイル名：3_xml_list.txt

フォーマット：

冒頭の#で始まる行はヘッダ部であり、その後のデータ部の各行は

中国・日本公開特許公報の１つの対となっている。

項目の説明：

1	cntry(src)	CN	（固定：中国を表す）
2	kind(src)	A	（固定：公開特許公報を表す）
3	pub-num(src)	中国公開特許公報・公開番号	
4	pub-date(src)	中国公開特許公報・公開日付	
5	cntry(tgt)	JP	（固定：日本を表す）
6	kind(tgt)	A	（固定：公開特許公報を表す）
7	pub-num(tgt)	日本公開特許公報・公開番号	
8	pub-date(tgt)	日本公開特許公報・公開日付	
9	ipc	国際特許分類	
10	fam-id	パテントファミリーID	

データ例：

CN　A　103477788　2014.01.01　JP　A　2013255491　2013.12.26　A01D34/73　48626242

A2.4 対応中国公開特許公報・和文抄録文献番号リストのフォーマット

ファイル名：4_excerpts_list.txt

フォーマット：

冒頭の#で始まる行はヘッダ部であり、その後のデータ部の各行は

中国公開特許公報・和文抄録の１つの対となっている。

和文抄録ファイルの文献番号は中国公開特許公報の公開番号と同一なので省略。

項目の説明：

1	cntry	CN	（固定：中国を表す）
2	kind	A	（固定：公開特許公報を表す）
3	pub-num	中国公開特許公報・公開番号	
4	pub-date	中国公開特許公報・公開日付	
5	ipc	国際特許分類	

データ例：

CN　A　102301846　2012.01.04　A01B43/00

A2.5 人手確認用対訳辞書候補データのフォーマット

ファイル名：5_cands.txt

フォーマット：

各行は１つの見出し語候補；訳語候補のセットとなっている。

項目の説明：

1　　中国語・見出し語候補

2　　日本語・訳語候補

3　　35 技術分野の分野番号（3.1.1 節の表 3.1.1-1）
　　複数個ある場合は「,」区切りで並べる

データ例：

ABC 基因	ＡＢＣ遺伝子	15
ABC 现象	ＡＢＣ現象	16
AB 过程	ＡＢ過程	19
AC 电压	交流電圧	10, 15, 23, 32

A2.6 人手確認により対訳辞書から除外したデータのフォーマット

ファイル名：6_removed.txt

フォーマット：

各行は１つの見出し語候補；訳語候補のセットとなっている。

項目の説明：

1　　中国語・見出し語候補

2　　日本語・訳語候補

3　　対訳辞書から除外した理由

　　D1：　見出し語自体が不適切

　　D2：　訳語自体が不適切

　　D3：　見出し語と訳語が対応していない

　　D4：　辞書に登録する用語として不適切

4　　35 技術分野の分野番号（3.1.1 節の表 3.1.1-1）

　　複数個ある場合は「,」区切りで並べる

データ例：

AB 过程	ＡＢ過程	D1:	19
AD 模型大鼠	ＡＤモデルラット	D1:	16
ALKBH4	質	D2:	15
AP 方向	ＡＰ方向	D1:	13
AT 类型	ＡＴタイプ	D1:	15
Al 膜	膜	D2:	19
A 和 G	等	D2:	15
A 和图	図	D2:	10
A 池	渦	D2:	24

A3. 対訳辞書データ作成処理の具体例

本事業で作成した対訳辞書について、例をいくつか挙げて処理の流れを示す。

・例１「交互控制设备；インタラクティブ制御装置」

DOCDB の backfile から family-id が同一となる中・日の公開特許公報を抽出する：

【XML データの抜粋】

<exch:exchange-document country="CN" doc-number="103654720" kind="A" doc-id="416307933" date-publ="20140326" family-id="49000339" is-representative="YES" date-of-last-exchange="20150206" date-added-docdb="20140326" originating-office="EP">...

...

<exch:exchange-document country="JP" doc-number="2014045868" kind="A" doc-id="416469461" date-publ="20140317" family-id="49000339" is-representative="YES" date-of-last-exchange="20150206" date-added-docdb="20140508" originating-office="EP">...

family-id=49000339 のファミリーに、CNA103654720 と JPA2014045868 が属していることが分かる。これらの XML ファイルの一部を示す：

【CNA103654720】

<claims>...
  <claim id="cl0003" num="0003">
    <claim-text>3.根据权利要求 1 所述的交互控制设备，其中，还包括：

</claim-text>
    <claim-text>判断单元，用于根据从所述指示器装置输出的指令信号来判断是否进行

特定动作。

</claim-text>
  </claim>...
</claims>

【JPA2014045868】

<claims>...
<claim num="4">
<claim-text>
前記ポインタデバイスから出力された前記指示信号に応じて、前記第１の制御または前記第２の制御を

実行するか否かを判定するように構成された判定ユニット
をさらに備えることを特徴とする、請求項３に記載のインタラクティブ制御装置。
</claim-text>
</claim>...
</claims>

上記は、中・日が対応関係にあるにも関わらず項目番号がずれているが、アライメント処理に影響のないようにテキスト抽出の際にはいくつかの処理をしている。
CNA103654720 と JPA2014045868 から抽出されたテキストデータを入力として、中日対訳文アライメントツールにより作成された対訳コーパスの一部を示す：

0.13790167 ||| CNA103654720-JPA2014045868_des ||| A61B3/14 ||| C00 ||| 13 ||| 1-1 ||| 交互控制设备 104 可以经由有线或无线连接来与摄像设备 100 的指示器装置 107 和基座单元 103 相连接。 ||| インタラクティブ制御装置１０４は、有線または無線の接続を介して、ポインタデバイス１０７、および、撮像装置１００のベースユニット１０３と接続されてもよい。
0.08171951 ||| CNA103654720-JPA2014045868_des ||| A61B3/14 ||| C00 ||| 13 ||| 1-1 ||| 本发明涉及一种用于与对所显示的 GUI 上的虚拟指示器进行控制的指示器装置相连接的摄像设备、特别是进行光学相干断层成像的摄像设备的交互控制设备。 ||| 本開示は、表示されたＧＵＩ上で仮想ポインタを制御するポインタデバイスと接続される撮像装置、とりわけ、光干渉断層撮影を実行する撮像装置用のインタラクティブ制御装置に関する。

アライメントスコア（上記データの第一項）が 0.08 以上となる対訳コーパスを入力として、中日対訳用語抽出ツールにより「交互控制设备；インタラクティブ制御装置」が対訳辞書候補データとして抽出された。人手確認を経て対訳辞書に登録した。

・例２「微光学；マイクロ光学」

今度は DOCDB の weekly-update を使って family-id が同一となるような中・日の公開特許公報を抽出する：

【XML データの抜粋】
<exch:exchange-document country="CN" doc-number="104121851" kind="A" doc-id="423127722" date-publ="20141029" family-id="48288781" is-representative="YES" date-of-last-exchange="20150326" date-of-previous-exchange="20150212" date-added-docdb="20141029" originating-office="EP" status="A">...
...

<exch:exchange-document country="JP" doc-number="2014215300" kind="A" doc-id="424045951" date-publ="20141117" family-id="48288781" is-representative="YES" date-of-last-exchange="20150514" date-of-previous-exchange="20150326" date-added-docdb="20141128" originating-office="EP" status="A">...

family-id=48288781 のファミリーに、CNA104121851 と JPA2014215300 が属していることが分かる。それぞれの XML から抽出したテキストデータを入力して、中日対訳文アライメントツールにより作成された対訳コーパスの一部を示す：

0.10856065 ||| CNA104121851-JPA2014215300_des ||| G01B11/00 ||| P00 ||| 10 ||| 1-1 ||| 参考辐射 5 和照明辐射 6 然后在相应的输出端 44、45 处从微光学阵列 38 出现。 ||| 参照光５と照射光６は、次いで、マイクロ光学アレイ３８からそれぞれ出力４４、４５を通じて現れる。
0.10339109 ||| CNA104121851-JPA2014215300_des ||| G01B11/00 ||| P00 ||| 10 ||| 1-1 ||| 激光发射器 2 的激光辐射经由输入端 43 进入到微光学阵列 38 中。 ||| レーザーエミッタ２のレーザー光線は、入力４３を経てマイクロ光学アレイ３８に入る。

対訳コーパスを入力として、中日対訳用語抽出ツールにより「微光学；マイクロ光学」が対訳辞書候補データとして抽出された。

本用語については、上記過程とは別に中国公開特許公報要約と和文抄録の対からも抽出されている。以下に CNA102427200 の要約とその和文抄録の一部を示す：

【CNA102427200】
<abstract>
...构建与密闭操作空间相连的空气净化装置；将两片镜片与两个片状压电陶瓷分别粘结在一起形成微光学腔；计算微光学腔的腔长、精细度；...
</abstract>

【上記和文抄録】
<P>抄録文</P><P>...密閉表示空間とつながった空気清浄装置を構築する；二枚のレンズは二つの片状の圧電セラミックとそれぞれ一緒に付着させてマイクロ光学空胴を形成する；マイクロ光学空胴の長さ、精細度を計算する；...</P>

中国公開特許公報要約と和文抄録から抽出したテキストデータを入力して、中日対訳文アライメントツールにより作成された対訳コーパスの一部を示す：

0.09716271 ||| CNA102427200-excerpts_abs ||| H01S3/08 ||| E00 ||| 9 ||| 1-1 ||| 将两片镜片与两个片状压电陶瓷分别粘结在一起形成微光学腔； ||| 二枚のレンズは二つの片状の圧電セラミックとそれぞれ一緒に付着させてマイクロ光学空胴を形成する；

以上、対訳辞書候補「微光学；マイクロ光学」が別の文献から抽出される様子を述べた。この後、人手確認を経て、対訳辞書に登録した。