

日本語特許出願書類の中国語への
機械翻訳に関する調査
報告書

平成23年2月

特 許 庁

目次

第1章 概要	1
1.1. 調査目的	1
1.1.1. 調査目的	1
1.1.2. 調査内容	1
1.1.3. 調査実施方法の概要	2
1.2. 調査結果概要	4
1.3. 調査の手法	5
1.4. 調査実施体制	8
1.5. 調査スケジュール	10
第2章 調査環境整備及び対象文献の入手	12
2.1. 調査環境の整備	12
2.1.1. サーバ環境の準備	12
2.1.2. 日中翻訳ソフトの準備	12
2.2. 対象文献の入手	16
2.2.1. 文献データ	16
2.2.2. 調査対象文献	17
2.3. 分析用作業ファイルの設定	20
2.3.1. 技術分野別の分割	20
2.3.2. 明細書の各項目別の分割	20
2.3.3. 文切／文単位での分割	20
2.4. 調査に利用したツールの準備等	20
2.4.1. 形態素解析器・構文解析器の準備	20
2.4.2. パターン抽出	22
第3章 翻訳不備の原因調査	24
3.1. 調査の目的	24
3.2. 翻訳不備の要因分析	24
3.2.1. 機械翻訳の訳語の不備に基づく問題点	24
3.2.2. 助動詞に関する問題点	25
3.2.3. 訳の位置に関する問題点	27
3.2.4. 文の句切に関する問題点	28
3.2.5. 訳抜けに関する問題点	28
3.2.6. その他の問題点	29
3.3. 既存の文献から知得される翻訳不備の問題点	30
3.3.1. 「中国語機械翻訳技術に関する調査」 [5]	30
3.3.2. 特許版・産業日本語委員会報告書「産業日本語」 [6]	31
3.3.3. 「中国における特許翻訳の現状」 [7]	31
3.4. マクロ分析とミクロ分析	34
3.4.1. マクロ分析	34
3.4.2. ミクロ分析	41
3.4.3. 各不備要因に対する改善策	43
第4章 定型化可能な表現の分析	45
4.1. 調査の目的	45
4.2. 調査の対象	46
4.3. 調査の方法	47
4.3.1. 定型化が可能な場合	47

4.3.2. 定型化が困難な場合	54
4.4. 発明の名称.....	63
4.4.1. 定型化の分析及び結果.....	63
4.5. 要約	68
4.5.1. 定型化の分析及び結果.....	68
4.6. 特許請求の範囲	74
4.6.1. 定型化の分析及び結果.....	74
4.7. 明細書.....	80
4.7.1. 定型化の分析及び結果.....	80
第5章 技術分野別特性の分析.....	89
5.1. 調査の目的.....	89
5.2. 数式・化学式の特定方法.....	89
5.3. 数式・化学式・塩基配列の抽出.....	90
5.4. 特殊タグ・フォーマットの有無.....	96
5.4.1. 利用されている特殊タグ・フォーマット	96
5.5. 数式・化学式・塩基配列の技術分野別の割合	98
5.6. 特殊タグ・フォーマットの出現頻度	99
5.7. 技術分野別特性の考察	100
第6章 調査結果の検証.....	101
6.1. 調査目的	101
6.2. 改善効果の検証方法	101
6.2.1. 改善策の実装	101
6.2.2. 改善策を実施した後の分析結果の差異	102
6.2.3. 改善効果の分析.....	103
第7章 平成21年度調査結果との比較・分析.....	107
7.1. 調査の目的.....	107
7.1.1. 言語の文法的特性・言語の特性について	107
7.1.2. 定型化可能な表現について	108
7.1.3. 外来語表記について	109
7.1.4. 化学式(構造式)・塩基配列の表記法について	110
7.1.5. 句読点の表記法について	110
第8章 課題と対策	113
8.1. 機械翻訳不備の対策を実施する上での困難性.....	113
8.2. 定型可能表現の定型化	115
8.2.1. 定型化の問題点.....	115
8.2.2. 定型化の課題	118

第1章 概要

1.1. 調査目的

1.1.1. 調査目的

近年の中国の驚異的な経済的躍進に伴い、中国での特許出願が著しく増加し、特許情報に非常に注目が集まっている。この為、出願人や審査官等が中国語で記載された特許出願書類を扱う機会も著しく増えている。しかしながら、出願人や審査官が特許出願書類の検索や理解等を行う際には、中国語の有する独特の言語特性が大きな障壁として立ちふさがり、非常に労力を要するのが現実である。この問題に対する有効な手段として機械翻訳に対する期待が非常に高まっている。特に、検索の都度、日本語の検索キーワードを中国語に機械翻訳し、検索する方式（キーワード翻訳方式）等による日中機械翻訳の有効性が予想される。そこで、前記キーワード翻訳方式等による特許出願書類に対する日中機械翻訳の有効性を検討する為に分析を行い、問題点を把握し、且つ対応策の提案を行うことを目的として「日本語特許出願書類の中国語への機械翻訳に関する調査」（以下、本調査という）を実施した。また、調査に当たっては、我が国出願人の中国への特許出願書類作成に対する活用を鑑み、日本で公開された特許出願書類を中国語へ機械翻訳し、その翻訳精度を分析し、特許出願書類に係る日中機械翻訳の知見を得た。

1.1.2. 調査内容

本調査では、日本及び中国の公開特許公報データを用いて、日中機械翻訳の記述方式・内容を詳細に分析し、特性の有無を洗い出すと共に、定型化が可能な頻出表現・用語の把握、対訳データの作成、中国語の特性を考慮したうえで、機械翻訳するにあたって想定される問題点の把握、対応策の提案を行った。また、完全な定型化が困難な表現についても、類型・意味合い等によってグループ化し、共通部分について適切な登録文、語について提案し、それに対する定型文や辞書データを作成し、日中对訳を付与した。

以上の結果を踏まえて、平成 21 年度に特許庁が実施した「中国公開特許公報の機械翻訳による日本語での提供に関する調査」[4]（以下、平成 21 年度調査¹という）の結果と比較し、報告を行った。さらに、今後検討すべき課題についても報告を行った。

なお、本調査では、我が国への出願人が中国国家知識産権局（以下、SIPO という）に特許出願した中国の公開特許公報から、技術分野（国際特許分類（以下、IPC という）第 8 版 A～H セクション）毎に総計 1 万件を抽出し、その基礎出願である日本の公開特許公報を入手して行った。この際に、技術分野毎の抽出件数は、日本から SIPO へ出願された各分野の分布に従って行った（詳細は第 2 章の「2.2.対象文献の入手」を参照）。

分析は、技術分野毎に異なる特性が考えられる為、技術分野毎に行った。さらに、公開特許公報の記載項目別に異なる特性も考えられる為、各項目（発明の名称、要約、特許請求の範囲、発明の詳細な説明、発明の効果）毎に分析を行った。

¹ 平成 21 年度調査も、本調査を実施する(株)クロスランゲージが調査を行った。

1.1.3. 調査実施方法の概要

前述した本調査の実施方法の概略図が下記図 1.1.3.-1 であり、各手順を簡略的に説明したのが表 1.1.3.-1 である。なお、各手順の詳細については後述する。

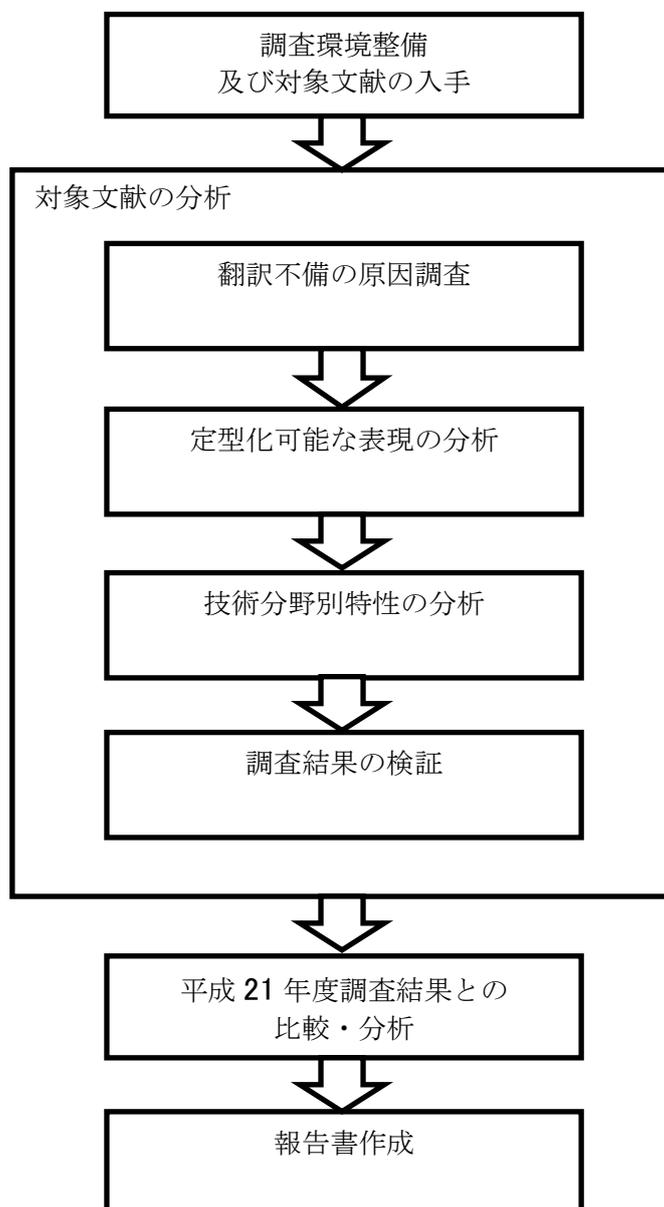


図 1.1.3.-1 本調査の実施方法

調査環境整備及び対象文献の入手		対象文献を収集し、作業環境を整備する。
対象文献の分析	翻訳不備の原因調査	翻訳不備の要因分析をして、原因を調査する。また翻訳不備案件を技術分野別に 5 件以上選択して詳細分析も行う。
	定型化可能な表現の分析	定型表現を収集して分析し、対訳を付与し、辞書登録を行う。
	技術分野別特性の分析	化学式・塩基配列、その他技術分野別の特性を分析する。
	調査結果の検証	定型化可能な表現の分析で集めた翻訳メモリや辞書を使って翻訳不備案件を再翻訳し、再度詳細分析を行って、改善効果を検証する。
平成 21 年度調査結果との比較・分析		本調査分析と平成 21 年度調査結果とを比較分析して問題点・課題を探る。
報告書作成		上記分析結果及び当該分析結果に基づく提案を検討し、報告書を作成する。

表 1.1.3.-1 本調査の手順

1.2. 調査結果概要

本調査の結果を総括する。調査に使用したツール及び調査の過程は第 2 章にて詳しく述べる。

(1) 翻訳不備の原因調査

日本語から中国語への機械翻訳の特性及び日本語と中国語の言語特性の差異に起因して考えられる問題点を考察した。次に、調査対象文献の全データをマクロ分析する事により、翻訳不備の要因について統計的な調査を行った。技術分野別の統計をとって見たところ、特に、技術分野 C(化学)のセクションで 1 文献当たりの原文の文字数が多い、未知語の出現率が高い等の明確な特性が見られた。次に、マクロ分析の結果から翻訳不備の要因を含む調査対象文献データを 40 件抽出し、詳細にミクロ分析した。その結果、名詞の訳語による不備が顕著であるという結果が得られた。

(2) 定型化可能な表現の分析

日本公開特許公報の各記載項目における定型文、定型パターン文、定型フレーズ、主文パターン文及び節パターンを収集し、対訳を付与した。全体で定型文 1,608 件、定型パターン文 540 件、定型フレーズ 758 件、主文パターン文 755 件及び節パターン 750 件の対訳を作成した。

(3) 技術分野別特性の分析

明細書本文中に化学式・塩基配列等が含まれる場合、現状の機械翻訳技術ではこの範囲をひとくくりに捉えられずに分断されて誤訳となる可能性が高い。これらを 1 つの名詞句として捉え、翻訳せずに訳出する事で翻訳結果の品質向上が期待できる。そこで数式・化学式・塩基配列の開始、中間、終端のパターンを分析して範囲の特定方法について検討を行った。

(4) 調査結果の検証

上記(1)及び(2)の分析から得られた改善策を本調査の調査対象文献データに適用し、機械翻訳の改善効果を検証する。改善策実装後に訳の各部分が実装前に比べてどう変化したか点を付けた(改善 1 点、同等 0 点、悪化-1 点)結果、全体的に「改善」(1 点)が 50%~80%と最も多く、改善効果が明確に確認できた。また、どのように改善効果が反映されたかについても具体的な例を挙げながら検証した。

(5) 平成 21 年度調査結果との比較・分析

言語の文法的特性、定型化可能な表現、外来語表記等の様々な観点から平成 21 年度調査結果と本調査との比較・分析を行った。その結果、特許文献を中国語から日本語に機械翻訳する場合と、日本語から中国語に機械翻訳する場合の差異及び課題を明確化する事ができた。例えば、特許請求の範囲に関して、中国の特許文献ではコロン及びセミコロンを多用しているが、日本の特許文献ではコロン及びセミコロンの代わりに読点を使用しているという差異が判明し、その点に関して考察した。さらに、例えば、中国語の音訳文字リストを作成する事により日本語のカタカナ表記を処理可能である等の提案も行った。

1.3. 調査の手法

本調査を効率的に行う為に、作業項目を分割して調査を行った。作業項目は、(a) 調査環境整備及び対象文献の入手、(b) 分析用作業ファイルの設定、(c) 翻訳不備の要因分析、(d) 機械翻訳の付与、(e) 人手による翻訳、(f) マクロ分析、(g) ミクロ分析、(h) 定型化可能な表現の分析、(i) 技術分野別分析 (化学式・数式)、(j) 調査結果の検証、(k) 平成 21 年度調査結果との比較・分析からなる。

(a) 調査環境整備及び対象文献の入手

中国公開特許公報のデータは、すべて XML ファイルである。これの解析を行い、後述するように、IPC 分類 (A~H セクション) に基づく調査対象文の抽出を行った。

■ 中国公開特許公報のデータの構造を以下に示す。

cn-patent-document	
cn-bibliographic-data	書誌情報: 文章としてはタイトルと要約を含む
cn-publication-reference:	
document-id	
country	国(CN)
doc-number	公開番号(e.g. 1114142)
kind	文献種別(公報の種別)(e.g. A)
date:	日付 (e.g. 19960103)
gazette-reference	番号と日付
application-reference:	
document-id	
country	国(CN)
doc-number	出願番号 ファイル名の由来 (e.g. 93111339.3)
date	日付 (e.g. 19960103)
priority-claims	優先権主張 (ないこともある)
priority-claim	
country	国(e.g. JP)
doc-number:	出願番号(e.g. 281615/93)
date	(e.g. 19931015)
classification-ipc	IPC 分類
invention-title	発明の名称
abstract	要約
cn-related-publication	
cn-related-documents	
cn-parties	
cn-applicants	cn-applicant が複数並ぶ
cn-applicant	
addressbook	名称(組織名) と住所
cn-inventors	cn-inventor が複数並ぶ
cn-inventor	名称 (個人名)
cn-agents	cn-agent が複数並ぶ
cn-agent	特許事務所と弁理士

application-body	
description	明細書
invention-title	発明の名称（書誌情報にあったものと同じものが繰り返される） ※空の事もある
technical-field	技術分野 言葉で説明されている
background-ar	背景技術（従来の技術）
disclosure	発明の開示
description-of-drawings	図面の説明
mode-for-invention:	実施例等、本文中に「本発明の実施例:」とのタイトルが別にある
claims	請求項 請求項ごとに claim タグで囲まれている
drawings	図面説明は description-of-drawings にあるため画像リンク
cn-inventor	名称（個人名）
cn-agents	cn-agent が複数並ぶ

（b）分析用作業ファイルの設定

本調査の目的に則した分析を実施する為に、前準備として調査対象文献を IPC 分類に基づく 8 種類の技術分野、明細書の各項目別及び文単位に分割する処理を行った。

（c）翻訳不備の要因分析

中国語への機械翻訳の特性及び日・中の言語特性(文法的特性、文字特性)を考慮して留意すべき要因を分析する。

（d）機械翻訳の付与

機械翻訳の問題点を探る為に人による翻訳(取得した中国の文献データ)と機械翻訳(取得した日本国の文献データを機械翻訳したもの)との両方を比較し、分析を行う。また、マイクロ分析に使用する為に、対象文の機械翻訳のみならず、形態素毎に訳振りを行う。

（e）人による翻訳

下記(h)の作業が終了した段階で定型文、定型パターン文、定型フレーズ、主文パターン文及び節フレーズに対訳を付与する。また、下記(f)の作業が終了した段階でマイクロ分析に使用する為に、8 種類の技術分野から抽出した 40 件の特許文献に翻訳者による対訳を付与する。

（f）マクロ分析

翻訳不備の要因となる要素を、統計的な情報を収集し、大局的な視点から分析する。分析の際には、1 文当たりの動詞の数や、動詞を 2 以上含む重文等にも着目する。また、技術分野別の分析も行う。詳細は第 3 章に示す。

（g）マイクロ分析

特に、機械翻訳精度の評価の低い案件を抽出し、翻訳不備の要因となる要素を個別詳細に分析する。詳細は第 3 章に示す。

(h) 定型化可能な表現の分析

パターンを統計的に集計して解析を行う事で定型的表現・フレーズを収集し、出現頻度が高い表現について、文法的特性、言い回し、慣用表現等の解析を行う。詳細は第 4 章に示す。

(i) 技術分野別分析 (化学式・数式)

技術分野別に記述内容、パターン、表現等の特性を分析する。化学式、数式を名詞句としてまとめることにより、翻訳精度向上を試みる。詳細は第 5 章に示す。

(j) 調査結果の検証

上記 (g) 及び (h) で得られた改善策を機械翻訳エンジンに反映させ、その改善効果の分析を行う。詳細は第 6 章に示す。

(k) 平成 21 年度調査結果との比較・分析

上記 (a) ~ (j) で得られた結果を平成 21 年度調査結果と比較し、例えば中国語と日本語との言語の文法的特性・言語の特性、定型化可能な表現及び外来語表記についての差異等、種々の視点から分析を行う。詳細は第 7 章に示す。

1.4. 調査実施体制

プロジェクト統括責任者のもとに、サブリーダー、調査環境整備及び対象文献の入手チーム、翻訳不備の原因調査チーム、定型化可能な表現の分析チーム、技術分野別特性の分析チーム、翻訳チーム、平成 21 年度調査結果との比較・分析チーム、報告書作成チームを設置し、効率的に調査を遂行した。また、必要に応じて調査に関するアドバイスを仰ぐ為のアドバイザーも確保した。

(1) 実施体制図

本調査の実施体制図を、下記図 1.4.-1 に示す。

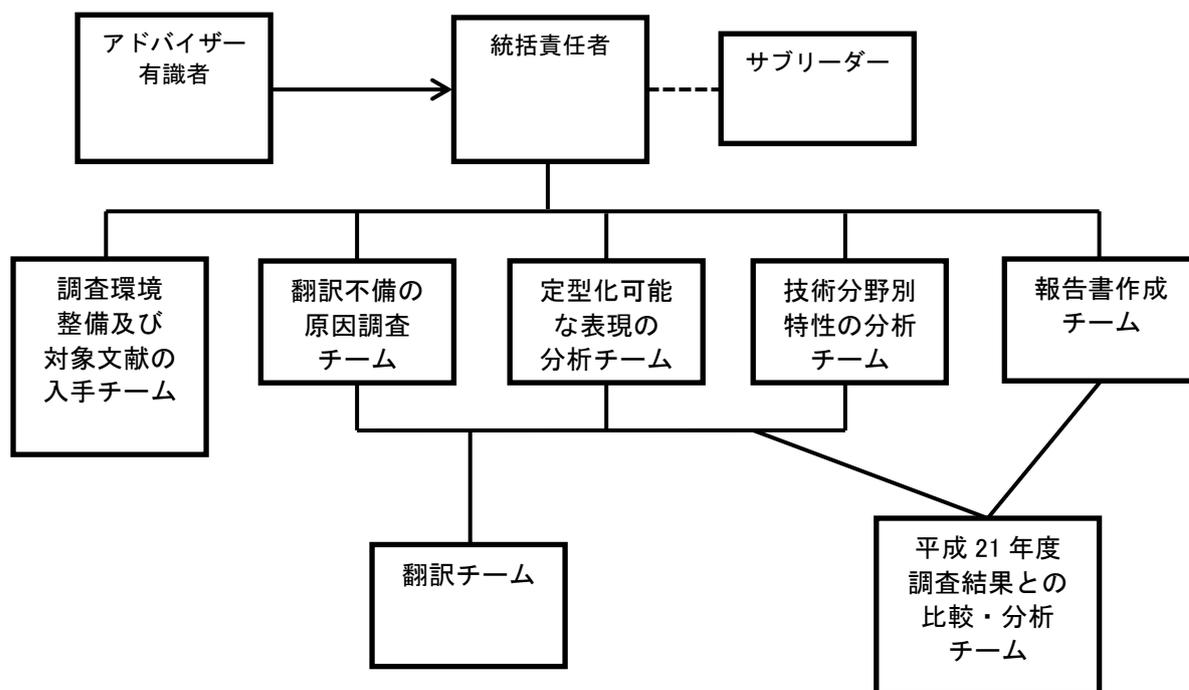


図 1.4.-1 調査実施体制

(2) 各チームの役割

- ・ 統括責任者
本調査事業の全体工程の管理、プロジェクトの推進を行う。
- ・ サブリーダー
統括責任者を補佐し、本調査事業の全体工程を管理する。
- ・ 調査環境整備及び対象文献の入手チーム
調査対象文献データの翻訳、解析を行うサーバ／マシン等を準備すると共に、調査対象文献を入手する等、作業環境の準備を行う。
- ・ 翻訳不備の原因調査チーム
翻訳不備の原因調査、及び翻訳不備案件の分析と精度向上の効果の検証を行う。
- ・ 定型化可能な表現の分析チーム
定型表現を収集して分析し、対訳を付与し、辞書登録を行う。
- ・ 技術分野別特性の分析チーム
化学式・塩基配列、その他の分野別の特性を分析する。
- ・ 翻訳チーム
定型表現に中国語訳を付与する等、必要に応じて母国語レベルの中国語スキルを調査に提供する。
- ・ 平成 21 年度年度調査結果との比較・分析チーム
平成 21 年度調査結果（「中国公開特許公報の機械翻訳による日本語での提供に関する調査」）との比較・分析を行う。
- ・ 報告書作成チーム
報告書の作成を行う。

1.5. 調査スケジュール

本調査のスケジュール概要を下記図 1.5.-1 に示す。

		10月	11月	12月	1月	2月
■計画						
◎計画の策定	予	→				
	実	→				
■準備						
◎調査環境整備及び対象文献の入手						
○調査環境の整備	予	→				
	実	→				
○対象文献の入手 (10,000 件)	予	→				
	実	→				
○分析用作業ファイルの設定	予	→				
	実	→				
■対象文献の分析						
◎翻訳不備の原因調査						
○翻訳不備の要因分析	予	→				
	実		→			
○マクロ分析	予	→				
	実		→			
○ミクロ分析	予	→	→			
	実		→	→		
◎定型化可能な表現の分析						
○定型化可能な表現の分析	予		→			
	実		→			
○定型化可能な表現の訳付	予		→	→		
	実		→	→		
○定型化が困難な表現の類別化・定型化	予		→			
	実		→			
○定型化が困難な表現の訳付	予		→	→		
	実		→	→		
◎技術分野別特性の分析						
○化学（構造）式・塩基配列の分析	予		→			
	実		→	→	→	
○その他技術分野特有表現の分析	予		→			
	実		→	→	→	
○特殊タグ等	予		→			
	実		→	→	→	

表 1.5.-1 (次頁に続く)

		10月	11月	12月	1月	2月
◎調査結果の検証						
○改良版による翻訳不備案件のマイクロ分析	予			→		
	実				→	
○改善効果の分析	予				→	
	実				→	
■分析						
○平成 21 年度調査結果との比較・分析	予				→	
	実				→	
■分析のまとめ						
◎報告書の作成						
○報告書作成	予				→	→
	実				→	→
○問題点・課題のまとめ	予					→
	実					→

表 1.5.-1 (前頁からの続き)

予：初期計画時点のスケジュールを示す。
 実：本調査の実際のスケジュールを示す。

第2章 調査環境整備及び対象文献の入手

2.1. 調査環境の整備

2.1.1. サーバ環境の準備

2.1.1.1. 利用システム

本調査は、下記に一例を示すサーバ／マシン等を含むシステムを利用して実施した。

[解析用サーバ／マシンのシステム性能]

OS:WindowsXP(SP3)
メモリ:3.0GB
CPU:Core2Duo 2.13GHz
ディスク容量:2TB

2.1.1.2. ネットワーク環境

本調査は、インターネットに常時接続可能な LAN 上に解析用のサーバ／マシンが接続されているネットワーク環境上で実施した。

2.1.2. 日中翻訳ソフトの準備

本調査は、(株)クロスランゲージの日中翻訳ソフト「NewJC」を利用して調査解析を実施した。正確には同社の WEB-Transer に実装されている NewJC のエンジンを利用した。該翻訳ソフトは、現在「Yahoo!翻訳」で利用されている等の実績がある。前述した解析用のサーバ／マシンに該翻訳ソフトをインストールした上で調査を行った。なお、該翻訳ソフトが必要とする動作環境は以下の通りである²。

[動作環境(クライアント側)]

対応 OS:WindowsXP, WindowsVista, Windows7
HD 容量:10GB 以上の空き容量
その他:InternetExplorer6/7/8/FireFox3

基本語辞書:約 363,000 語
専門語辞書:10 分野 約 537,000 語
同時使用が可能な辞書:ユーザー辞書 2 個、専門語辞書 2 個
その他の特性:1 つの辞書に登録可能な語数 10,000 語

² 上記では Windows 版の NewJC の動作環境を示したが、解析では Linux 版の NewJC のエンジン及び PAT - Transer の JE のエンジン(日本語から英語への翻訳エンジン)に実装されている形態素解析器及び構文解析器を使用した。これは、(株)クロスランゲージの既存システムを活用した為である。

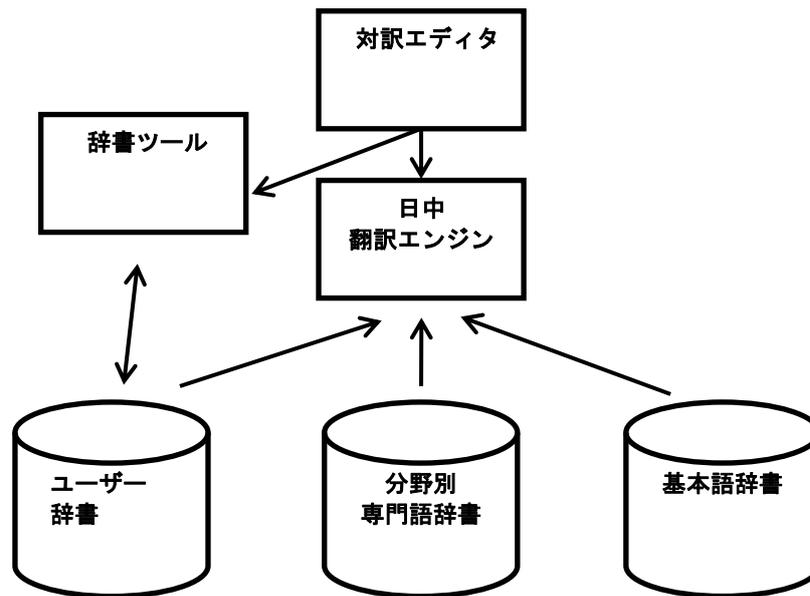


図 2.1.2.-1 翻訳システム構成図

図 2.1.2.-1 に日中翻訳システムの基本構成図を示す。対訳エディタはユーザーインターフェース画面である。ユーザーは、対訳エディタを用い日本語を入力すると共に、辞書ツールを使ってユーザー辞書に辞書登録を行うことができる。対訳エディタから呼び出される日中翻訳エンジンが、基本語辞書³・分野別専門語辞書⁴・ユーザー辞書⁵を用いて日本語を中国語に翻訳する。

表 2.1.2.-2 に基本語辞書の語数、表 2.1.2.-3 に専門語辞書の分野一覧及び語数を示す。

表 2.1.2.-2 基本語辞書の語数

分野	語数
基本語辞書	360,000

表 2.1.2.-3 専門語辞書の分野一覧と語数

分野	語数
貿易	13,000
コンピュータ	90,000
電気電子	72,000
機械工学	63,000
金属	20,000
数学物理	63,000
航空宇宙	14,000
海洋船舶	10,000
医療医学	106,000
化学	86,000

³ 翻訳エンジンに必須の辞書であり、日本語を中国語に翻訳するための基本的な語が格納された辞書

⁴ 技術分野の特有の語が格納された辞書

⁵ システムに搭載された辞書に含まれていない語などをユーザーが自分で登録・削除が可能な辞書

図 2.1.2.-4 に、対訳エディタ画面を示す。対訳エディタ画面とは、原文と訳文を左右に対比させて表示できる編集用画面である。図 2.1.2.-4 に示すように、対訳エディタ画面の下部では、原文／訳文の各語にマウスカーソルを合わせる事によってどの語がどの中国語に訳されたのかが一目でわかるような機能も利用する事が可能である。ユーザー辞書機能は対訳エディタの 1 つの機能であり、ユーザー辞書メニューから呼び出す。ユーザー辞書に単語を新規に登録、あるいは削除するための機能である。図 2.1.2.-5 に辞書登録画面を示す。

The screenshot displays the WEB-Transter website interface. At the top, there are navigation links for 'ログアウト', '設定', and 'ヘルプ'. Below this is a menu bar with options like 'テキスト翻訳', 'ウェブ翻訳', '翻訳検索', '辞書参照', 'ユーザー辞書', '翻訳メモリ', '辞書管理', and 'ユーザー管理'. A language selection bar shows '英語', '中国語', '韓国語', 'フランス語', 'ドイツ語', 'イタリア語', 'スペイン語', and 'ポルトガル語'. The main content area is titled '言語選択:' and includes radio buttons for '中国語 (簡体)→日本語', '日本語→中国語 (簡体)', '中国語 (繁体)→日本語', and '日本語→中国語 (繁体)'. The '原文:' field contains Japanese text about a fishing bag, and the '訳文:' field shows its Chinese translation. A '翻訳' button is prominently displayed between the two text areas, along with '確認翻訳' and 'クリア' buttons. Below the editor, there are checkboxes for '翻訳設定' and '訳語対応情報をマウスオーバー時に表示する。'. A '翻訳結果' section shows a table with four rows, each containing a numbered item, its original Japanese text, and its translated Chinese text. The footer includes 'WEB-Transter > テキスト翻訳 > 中国語', a 'ページトップ' link, and a copyright notice: 'Copyright (C) 2009 - 2010 Cross Language Inc. All rights reserved.'

図 2.1.2.-4 対訳エディタ画面

テキスト翻訳	ウェブ翻訳	翻訳検索	辞書参照	ユーザー辞書	翻訳メモリ	辞書管理	ユーザー管理
登録	一覧	インポート	エクスポート	新規/変更/削除			

ユーザー辞書への単語の登録を行うことで、翻訳結果をユーザーの好みにあわせて変更することができます。
(ファイルから一括登録を行う場合は[インポート]を使います)

翻訳エンジン: 日本語 → 中国語 (簡体) ⓘ

辞書: ユーザー辞書 ⓘ 情報...

見出し語: 釣人 ⓘ

訳語: 钓鱼者 ⓘ

登録 クリア

注意事項: 言語解析処理の関係上、ユーザー辞書への登録を行ってもその訳語が使用されない場合があります。

WEB-Transer > ユーザー辞書 > 登録 ページトップ

Copyright (C) 2009 - 2010 Cross Language Inc. All rights reserved.

図 2.1.2.-5 辞書登録画面

2.2. 対象文献の入手

2.2.1. 文献データ

(1)使用した文献データ

本調査で利用した文献データは、最近の傾向をより反映した分析結果が得られるようにできるだけ直近のデータを抽出して調査対象とした。具体的には、日本公開特許公報(2004年～2008年出願)及び中国公開特許公報(2004年～2006年出願)を抽出データの母体とした(表 2.2.1-1 参照)。中国公開特許公報は、特許庁より貸与された文献データである。日本公開特許公報は、(株)クロスランゲージの保有する文献データである。後述するように、日本国特許庁へ出願された特許出願の文献データ(以下、日本国の文献データという)と、当該出願を基礎出願として優先権を主張して中国国家知識産権局に出願された特許出願の文献データ(以下、中国の文献データという)とをパテントファミリーを手掛かりにして見つけ出し、見つけ出した日本語及び中国語の文献データを同一出願内容の文献データとして母体データから抽出した。

表 2.2.1-1 文献データ

内容	出願年	件数	入手先
中国公開特許公報	2004年～2006年	419,229	特許庁より貸与
日本公開特許公報	2004年～2008年	1,914,912	調査会社保有
合計		2,334,141	

2.2.2. 調査対象文献

2.2.2.1. 調査対象文献の抽出手順

(1)抽出手順の概要

データの抽出は、前述した母体データから第 1 の抽出手順でデータを抽出し、この抽出したデータから第 2 の抽出手順で最終的なデータを抽出するという 2 段階で行った。第 1 の抽出手順及び第 2 の抽出手順の詳細は次の(2)、(3)に記載の通りである。

(2)第 1 の抽出手順(日本国の文献データと、当該日本国特許庁への出願を基礎出願として優先権主張をして中国国家知識産権局に出願された中国の文献データを同一内容の文献データとして抽出)

調査対象文献を決定する第 1 の抽出手順では、日本語と中国語との両方で同一出願内容が存在するデータを母体データ(中国公開特許公報(2004 年～2006 年出願)、日本公開特許公報(2004 年～2008 年⁶出願))から抽出する。本調査では、出願人が日本国に出願し、その後この日本国特許庁への出願を基礎出願としてパリ優先権主張をして中国国家知識産権局にも出願したデータを、その優先権の対応付けに基づいて次の図 2.2.2-1 に示すように抽出した。これにより日中で対応する 35,473(×2)件のデータが抽出された。

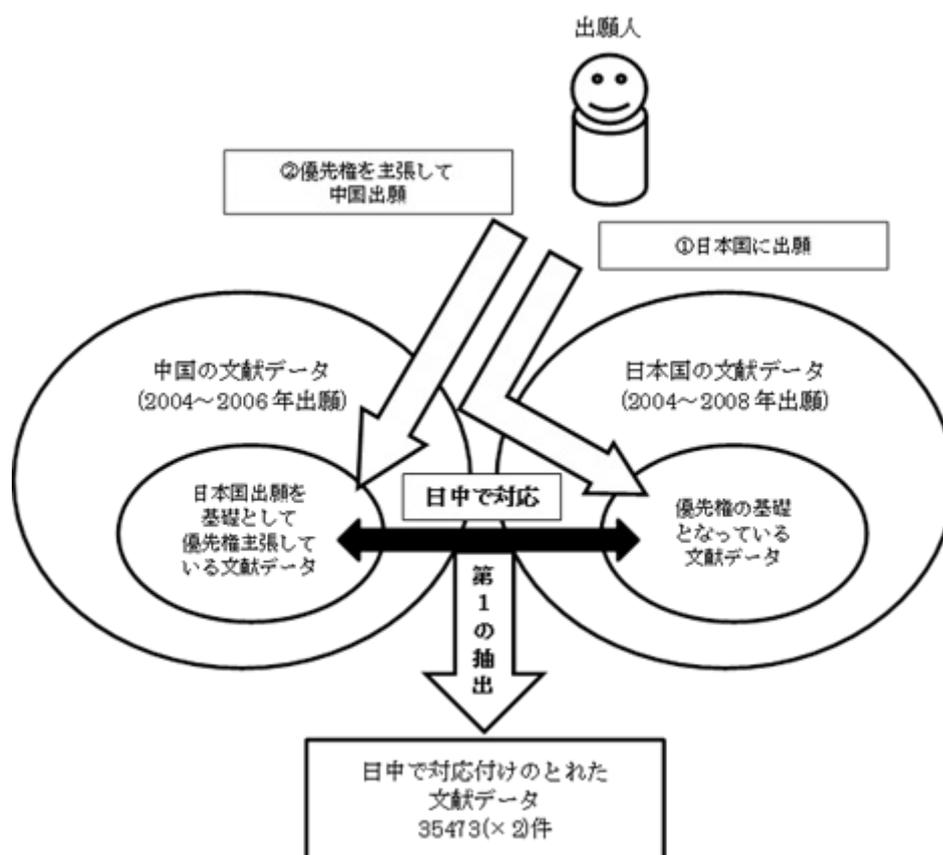


図 2.2.2.1-1 文献データ

⁶ 母体データとなる日本公開特許公報(2004 年～2008 年出願)が、中国公開特許公報(2004 年～2006 年出願)よりも公開時期の遅いデータを含むが、これは基礎出願の公開の方が遅いレアケースも含める為である。

※処理の都合上、第1の抽出手順は下記[1]~[3]の条件下で行った。

[1] 中国の文献データのうち、複数の優先権主張がされているデータは日中の対応特定が煩雑になる為、抽出対象から外した。

[2] 日本国の文献データのIPCと中国の文献データのIPCとが相違しているデータは、後述の第2の抽出の際に利用するIPCが不明確になる為、抽出対象から外した。

[3] 中国の文献データに記載の優先権主張の基礎出願の出願日と日本国の文献データの出願日とが相違している場合には誤抽出を防ぐ為、抽出対象から外した。

(3)第2の抽出手順(技術分野別の分布に基づくデータの抽出)

調査対象文献を決定する第2の抽出手順では、前述の第1の抽出手順で抽出された日中の対応付けのとれたデータから、さらに各技術分野に属するデータが「所定の比率」になるように抽出を行う。従って、抽出する為に「所定の比率」を決定する必要があるが、本調査では、1996年~2006年に中国国家知識産権局に日本国特許庁への出願を基礎出願として優先権主張をして出願された案件に対して前述の[1]~[3]の条件を付けずに抽出したデータを母体としてIPCの分布比率を求め、これを「所定の比率」として用いた。得られた結果を下記表2.2.2.1-1に示す。

なお、前述の第1の抽出手順の際には課した[1]~[3]の条件を付けなかったのは、分布比率を求める為の母集団を十分に大きくとる為である。

表 2.2.2.1-1 IPC 比率データ

IPC	日本から中国に優先権主張して出願した案件	割合	[参考] 1996年~2006年 中国公開の全案件	[参考] 割合
A:生活必需品	9,962	6.6%	127,874	16.60%
B:処理操作;運輸	20,242	13.4%	96,620	12.50%
C:化学;冶金	17,464	11.6%	145,758	18.90%
D:繊維;紙	2,839	1.9%	16,233	2.10%
E:固定構造物	1,417	0.9%	19,027	2.50%
F:機械工学;照明; 加熱;武器;爆破	11,244	7.4%	52,704	6.80%
G:物理学	40,150	26.6%	141,467	18.40%
H:電気	47,785	31.6%	170,630	22.20%
合計	151,103	100%	770,313	100%

※母集団を大きくする為に、前述した[1]~[3]の条件を付けずに第1の抽出手順で151,103件の案件を抽出し、IPC比率を求めた。

※母集団を大きくする為に、IPC比率を求める際には、より広めの1996年~2006年の文献データを母集団とした。

前述の第1の抽出手順で抽出されたデータ 35,473 件から、表 2.2.2.1.-2 に示す「所定の比率」(=IPC 比率)に基づいて第2の抽出手順でデータを抽出した結果を下記表 2.2.2.1.-2 に示す。

表 2.2.2.1.-2 最終抽出結果

IPC	第1の抽出による 日中対応出願案件	割合	最終的な 調査対象案件	最終的な 割合
A:生活必需品	1,874	5.3%	660	6.6%
B:処理操作;運輸	5,029	14.2%	1,340	13.4%
C:化学;冶金	1,691	4.8%	1,160	11.6%
D:繊維;紙	730	2.1%	190	1.9%
E:固定構造物	353	1.0%	90	0.9%
F:機械工学;照明; 加熱;武器;爆破	3,211	9.1%	740	7.4%
G:物理学	10,187	28.7%	2,660	26.6%
H:電気	12,398	35.0%	3,160	31.6%
合計	35,473	100%	10,000	100%

※処理の都合上、第2の抽出手順は下記[1]の条件下で行った。

[1] 第1の抽出手順で抽出されたデータ 35,473 件のうち、最終的な抽出案件数が 10,000 件となるように抽出した。

2.3. 分析用作業ファイルの設定

2.3.1. 技術分野別の分割

本調査の目的に則した分析を実施する為に、前準備として調査対象文献 10,000 件を IPC に基づく 8 種類の技術分野別(IPC 第 8 版 A~H セクション)にファイル分割処理を行った。

2.3.2. 明細書の各項目別の分割

特許文献固有の分析を実施する為に、ファイル分割されたデータをさらに公開特許公報の記載項目別(発明の名称、要約、特許請求の範囲、発明の詳細な説明、発明の効果)に分割処理を行った。

2.3.3. 文切／文単位での分割

データ内容を詳細に分析する為に、技術分野別／公開特許公報の記載項目別に分割処理されたデータに対して、さらに文単位での分割処理を行った。

2.4. 調査に利用したツールの準備等

2.4.1. 形態素解析器・構文解析器の準備

2.4.1.1. 目的

マクロ分析／ミクロ分析等の解析において、調査対象文献中の文構成を検出する必要がある。そのために、文を形態素に分解し、名詞などの品詞を付与することを目的とするのが形態素解析器である。さらに、形態素解析器で分解した形態素から文の係り受けを解析することを目的とするのが構文解析器である。

2.4.1.2. 手法

本調査で利用した形態素解析器・構文解析器は、前述した(株)クロスランゲージの日中翻訳ソフト「NewJC⁷」に実装されている JC エンジンの日本語形態素解析器・日本語構文解析器を利用した。さらに、これとは別に(株)クロスランゲージの日英翻訳ソフトの「PAT-Transer⁸」に実装されている JE エンジンの日本語形態素解析器・日本語構文解析器も利用し、解析条件に合わせた最適な解析を実施した。これにより、実際のエンジンに合わせた解析が好ましい場合には JC エンジンの形態素解析器・構文解析器を利用し、高精度の解析が必要となる場合には JE エンジンの形態素解析器・構文解析器を利用するという使い分けが可能になった。

具体的な JC/JE の両エンジンの形態素解析器・構文解析器の使い分け状況を下記に記載する。

<マクロ／ミクロ分析>

解析内容	解析器の種別
マクロ分析	JC エンジン形態素解析器
ミクロ分析	JC エンジン構文解析器

⁷ 日中翻訳ソフト。(株)クロスランゲージがパッケージ等で販売している製品。「2.1.2. 日中翻訳ソフトの準備」も参照。

⁸ 特許用の日英翻訳ソフト。(株)クロスランゲージがパッケージ等で販売している製品。

※マクロ分析及びマイクロ分析では実際のエンジンの出力を調査する必要があるので、JC エンジンの解析器を利用。

<定型化可能な表現の分析>

解析内容	解析器の種別
定型パターン文	JC エンジン形態素解析器
定型フレーズ	JE エンジン形態素解析器
主文パターン文	JE エンジン構文解析器
節パターン	JE エンジン構文解析器

※構文解析器を利用する際には同一エンジンの形態素解析器を用いて形態素解析を行った上で構文解析器を利用した。

※定型化可能な表現の分析ではパターンを調査する為に、より高精度な構文解析が必要となるので、基本的に JE エンジンの解析器を利用した。

※なお、特許特有のフレーズを抽出する際には、非特許用⁹の JE エンジンの形態素解析器を利用する事で特許文献特有のフレーズを浮き上がらせ、効果的な抽出を行った。

形態素解析器を用いて形態素解析を行った際の実例の例は、「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の「[1]形態素解析」を参照されたい。

構文解析器を用いて構文解析を行った際の実例の例は、「4.3.2.1.2. 主文パターン文の抽出方法の詳細」の「[2]構文解析」を参照されたい。

⁹ JE エンジンの形態素解析器としては、非特許用の PC-Transer2011(Ver.18)のものを利用し、JE エンジンの構文解析器としては特許翻訳に特化した PAT-Transer2009(Ver.10)のものを利用した。

2.4.2. パターン抽出

2.4.2.1. 目的

調査対象文献の定型化可能な表現の分析¹⁰の為、前述した形態素解析器・構文解析器¹¹を用いて形態素解析を行い、得られた解析結果から類似するパターンを抽出してまとめあげる。

2.4.2.2. 手法

2.4.2.2.1. スケルトン化

パターンの抽出に当たり、まず、対象となる文または節に対してスケルトン化を行う。スケルトン化とは、文または節の中で他の文または節との差異になると思われる部分を「@」で置換し、重複するパターンを抽出する事である。従って、「@」はパターン中の変数的な意味をもつ。

例えば、

(1)	【X】が【塩素原子】である請求項【1】記載の【マロン酸モノメチル誘導体】。
(2)	【溶剤】が【芳香族系溶剤】である請求項【1】記載の【アセト酢酸エステル類の製造方法】。

の2つの文をスケルトン化すると、

@が@である請求項@記載の@。

というパターンを抽出する事ができる。

2.4.2.2.2. パターン候補ファイルの作成

パターンを抽出した後に必要な作業は、パターンの頻度及び有効性の検証である。この頻度及び有効性を検証するに当たり、次に示すような形式のパターン候補ファイルを作成した。

- ・パターン候補ファイルデータ形式

一つのパターンにつき以下のようなデータが形成され、生成されたパターンの個数だけ繰り返される。

1 カラム目	2 カラム目
頻度	パターン候補
	パターンにマッチした原文 1
	パターンにマッチした原文 2
	...
	パターンにマッチした原文 n

¹⁰ ここではパターン抽出の概略的な説明のみを行う。具体的な解析内容については後述の「第4章 定型化可能な表現の分析」を参照。

¹¹ 形態素解析器・構文解析器については、「2.4.1.1. 目的」等を参照。

1. 頻度
対象とした特許文のうち、このパターン候補にマッチした原文の数
2. パターン候補
生成されたパターンそのもの。変数の部分を意味する「@」を含んでいる。
3. パターン候補にマッチした原文
パターン候補を生成する元になった文を意味している。目視チェックが容易に行えるように、差異の部分(原文パターンの@に対応する部分)を【】で囲む処理を施している。

・パターン候補ファイルデータ例

1 カラム目	2 カラム目
53	ことを特徴とする請求項@に記載の@。
	ことを特徴とする請求項【2】に記載の【電動ドライバ】。
	ことを特徴とする請求項【25】に記載の【ヒータプレートの製造方法】。
	ことを特徴とする請求項【1】に記載の【記録装置】。
	ことを特徴とする請求項【6】に記載の【撮像装置】。
	ことを特徴とする請求項【12】に記載の【再生装置】。
	……(省略)
--	
52	@が@である請求項@記載の@。
	【X】が【塩素原子】である請求項【1】記載の【マロン酸モノメチル誘導体】。
	【溶剤】が【芳香族系溶剤】である請求項【1】記載の【アセト酢酸エステル類の製造方法】。
	【蒸留】が【単蒸留】である請求項【1】記載の【製造方法】。
	【粉体状ポリマーの平均粒径】が【0.05~5mm】である請求項【1】記載の【方法】。
	【酸化チタンの平均粒子径】が【40nm以下】である請求項【1】記載の【酸化チタン分散液】。
	……(省略)
--	
	……(以下省略)

第3章 翻訳不備の原因調査

3.1. 調査の目的

日本語から中国語への機械翻訳を行う際の翻訳不備の原因を調査し、その改善策を考察する。最初に翻訳不備の要因を分析する。次に、調査対象文献の全データに対してのマクロ分析と、翻訳不備の要因を有する(又は有する可能性が高い)データに対してのミクロ分析を実施し、実際の翻訳不備要因を調査する。そして、各要因に対して改善策を検討する。

3.2. 翻訳不備の要因分析

翻訳不備の原因調査を行うに当たって、まず、日本語から中国語への機械翻訳の特性及び日本語及び中国語の言語特性(文法的特性、文字特性)の差異に起因して考えられる問題点を考察する。各問題点を説明する際には、理解の容易さを目的として後述の「3.2.2. ミクロ分析」で実際に分析した結果、検出された実例を提示して説明する。また、各問題点に関する改善策は、後述の「3.4.3. 各不備要因に対する改善策」で考察をする。

3.2.1. 機械翻訳の訳語の不備に基づく問題点

日本語を中国語に機械翻訳する際に非常に重要な問題点として訳語の不備が挙げられる。この訳語の不備に関して、以下に3つの代表的な例を挙げる。

(1)一般的な訳語の他に特許文にふさわしい訳語がある場合

一般的な訳語ではなく特許文にふさわしい訳語がある場合には、特許文にふさわしい訳語を用いるべきである。しかしながら、機械翻訳では、一般的な訳語が用いられてしまう場合がある。下記に実例を挙げる。

例(ミクロ分析結果の実例データ A0164)

日本語 原文	本発明は、魚を釣り上げた後に、この魚を収納しておく バッグに関する 。
中国語 機械翻訳文	钓鱼, 并且放了之后本发明 关于 收藏这条鱼的手提包。
対応する 中国特許文	本发明 涉及 一种在钓到鱼之后可以存放钓到的鱼的鱼护。

上記例では、「～に関する」が「关于～」と機械翻訳されている。しかしながら、特許文では「关于～」よりも「涉及～」の訳語を用いる方が適切である。

(2)一般的な訳語の他に専門的な訳語がある場合

専門的な訳語がある場合には、ニュアンスの異なる訳語ではなく、内容に即した専門的な訳語を用いるべきである。しかしながら、機械翻訳では、ニュアンスの異なる訳語が用いられてしまう場合がある。下記に実例を挙げる。

例(ミクロ分析結果の実例データ B0392)

日本語 原文	2個のクランプ31を備える ハンガー 30は、案内装置32により上下方向に位置決め自在である。
-----------	--

中国語 機械翻訳文	至于拥有 2 个的防滑钉片 31 的 衣架 30, 定位被向导装置 32 在上下方向自在。
対応する 中国特許文	具有两个夹钳 31 的 悬挂件 30 能够由引导装置 32 垂直地定位。

上記例では「ハンガー」が機械翻訳では「衣架」と訳されているが、「衣架」は「衣」の字が入っていることから分かるように「洋服掛け」の意味であり、この場合の機械装置の一部としての「ハンガー」の訳としては不適切である。対応する中国特許文のように「悬挂件」（「悬挂」は「掛ける」の意味で「件」は「部品」の意味）等の訳語を用いる方が適切である。

(3) 訳し分けの問題がある場合

日本語の助詞に対応する中国語の訳し方には複数の種類が存在する。その為、日本語の助詞に正しく対応する中国語が適切に訳し分けされる必要がある。しかしながら、機械翻訳では、用法の異なる中国語を用いてしまう場合がある。下記に実例を挙げる。

例(マイクロ分析結果の実例データ A0164)

日本語 原文	そして、釣人は魚を釣り上げるとバッグの袋部内に魚を収納する。
中国語 機械翻訳文	以及当钓鱼人提高鱼钓鱼的时候 在 手提包的袋子部里收藏鱼。
対応する 中国特許文	钓鱼者钓到鱼后, 将鱼 放 到鱼护的袋部分内存放。

上記例では、「袋部内に魚を収納する。」の助詞「に」が介詞「在」と訳されているが、中国語の介詞「在」は、状態を表す動詞とともに時間や場所を表すのに使われるものであって、上記の例のように動作の帰着点を表すのには用いられない。例えば、中国語で「收藏在～」は「～に収納されている」の意であるが、この場合は「收藏（収納されている）」が状態（存在）を表す動詞であるため、「在」の使用は正しい。一方で、上記の例の「収納する」は動作を表す動詞であり、その帰着点を表すには、対応する中国特許文のように「到」等の訳語を用いる方が適切である。

3.2.2. 助動詞に関する問題点

日本語の助動詞は、「時制」（現在形、過去形など）、「相」（完了形、進行形など）、「態」（能動態、受動態など）及び「法」（直説法、仮定法など）等の種々の重要な意味を表す品詞である。機械翻訳をする際に日本語のこの助動詞が適切に処理されない場合が存在する。以下に 4 つの代表的な例を挙げる。

(1) 日本語の「～した」を、中国語で誤って「了」に訳している場合

日本語で用いられる「～した」という表現は多くの場合は完了の意味だが、必ずしも完了を意味するわけではない。しかしながら、機械翻訳では完了の意味でないものも完了の意味で処理してしまう場合がある。下記に実例を挙げる。

例(マイクロ分析結果の実例データ A0164)

日本語 原文	魚を <u>活かした</u> まま収納可能なコンパクトなバッグであって、且つ、魚釣りの際の場所移動も容易な、携帯性に優れる魚収納バッグを提供する。
中国語 機械翻訳文	<u>发挥了</u> 鱼提供收藏是能被收藏的小型手提包以及在钓鱼的时候的地方移动容易的携带性杰出鱼手提包。
対応する 中国特許文	本发明提供一种携带性能优良的鱼护，它结构紧凑，是一种在钓鱼时可容易地移动场所，并且可以将鱼在 <u>存活</u> 状态存放的小型鱼护。

上記例では、「した」を含む「活かした」が「发挥了」と訳されているが、この「した」は完了を意味しておらず、中国語で完了を表す「了」を用いるのは不適切である。

(2)日本語の「～される」を、中国語で誤って受身に訳している場合

日本語で用いられる「～される」という表現は、必ずしも受身を意味するわけではない。しかしながら、機械翻訳では受身の意味でないものも受身の意味で処理してしまう場合がある。下記に実例を挙げる。

例(マイクロ分析結果の実例データ A0568)

日本語 原文	この工程で用いる海藻の煮汁としては、食用とされる褐藻類の煮汁であれば特に限定されないが、例えばひじき、わかめ、昆布、あらめ、もずく等の煮汁が <u>挙げられ</u> 、好ましくはひじき、あらめの煮汁である。
中国語 機械翻訳文	如果是褐藻類被认为是食用作为到这道工程使用的海藻的煮出来的汤的煮出来的汤的话，不被限定特别，但是肘比方说来，并且裙带菜，海带，毛病め，伯劳く等の煮出来的汤 <u>被列举</u> ，并且肘希望来，并且是毛病めの煮出来的汤。
対応する 中国特許文	作为该工序中使用的海藻煎汁，只要是能食用的褐藻类煎汁，没有特别的限定，例如 <u>可以举出</u> 羊栖菜、裙带菜、海带、茶色海藻、海蕴等煎汁，优选羊栖菜、茶色海藻的煎汁。

上記例では、「られ」を含む「例えば・・・が挙げられ、」が「比方说・・・被列举」と訳されているが、この「挙げられ」は可能の意味で用いられている為、中国語で受身として訳されているのは不適切である。対応する中国特許文のように「可以」等の訳語を用いる方が適切である。

なお、上記「中国語機械翻訳文」中に「毛病め」「伯劳く」とあるのは、日本語原文中の「あらめ」「もずく」がいずれも辞書に登録されていないため、機械翻訳によって「あら＝毛病（欠点、の意）の訳が出て、残った「め」が未知語として日本語のまま出力されており、同様に、「もず＝伯劳（鳥の「百舌）」と残った「く」が日本語のまま出力されているものである。

(3)日本語の「～させる」を、中国語で使役に訳して不自然になっている場合

日本語で用いられる「～させる」という表現は、そのまま中国語で使役の表現にすると不自然な文章になってしまう事がある。しかしながら、機械翻訳では使役の意味で処理してしまう場合がある。下記に実例を挙げる。

例(マイクロ分析結果の実例データ B0392)

日本語 原文	回転プレート 20 を垂直方向に <u>回転させ</u> 、図示を省略するクランプによりワーク w を保持させる。
中国語 機械翻訳文	<u>使</u> 旋转铭牌 20 在垂直方向旋转, 并且让被省略图示的防滑钉片保持工作 w。
対応する 中国特許文	旋转板 20 垂直地 <u>旋转</u> , 工件 w 被夹钳(未示)所夹持。

上記例では、「させ」を含む「回転させ、」が、そのまま「使・・・旋转」と使役で訳されているが、中国語で使役として訳されているのは不自然である。

(4)その他の助動詞が誤って訳されている場合

上記(1)～(3)に挙げた例以外のその他の助動詞表現(例えば、「～している」等)が、機械翻訳で誤って処理されている場合がある。下記に実例を挙げる。

例(マイクロ分析結果の実例データ E0007)

日本語 原文	内袋の周囲を縫製糸で縫製された形態の土嚢袋を <u>示しており</u> 、(イ)は斜視図で、(ロ)はB-B端面図である。
中国語 機械翻訳文	<u>正表示</u> 被用缝助线在内袋子的周围缝制了的形态的土囊袋(イ), 并且在斜视图(ロ)是 B-B 端面图。
対応する 中国特許文	<u>表示了</u> 用缝线缝上内袋四周的沙袋, (a)为斜视图, (b)为 B-B 端面图。

上記例では、「～しており」を含む「示しており、」が、「正表示」と訳されている。しかしながら、中国語の「正」は、進行中を表す副詞であり、この場合の状態を表す「～しており」の訳としては誤っている。正しい訳は、対応する中国特許文では「了」となっており、また、何も訳を出さないなど文脈によって異なる。

3.2.3. 訳の位置に関する問題点

機械翻訳による処理結果で中国文と日本文との語順が変わってしまい、不都合な場合がある。以下にその代表的な例を挙げる。

例(マイクロ分析結果の実例データ H2965)

日本語 原文	<u>本発明によれば</u> 、配線基板の孔に透光性蓋部の少なくとも一部が嵌入されているので、従来に比較して薄い固体撮像装置を得ることが出来る。
中国語 機械翻訳文	因为至少透光性盖子部的 1 部被 <u>据本发发明说</u> 电线敷设基础的洞嵌入さ所以在从来比较, 能得到薄固体撮像装置。
対応する 中国特許文	<u>根据本发发明的</u> 固态图像传感装置的特征在于透光盖部分的至少一部分安装于印刷电路板的孔中, 可以得到与常规固态图像传感装置相比更薄的固态图像传感装置。

上記例では、日本語では文頭にある「本発明によれば、」が、中国語では文の中ほどで「据本发发明说」と訳されている。しかしながら、これは、対応する中国特許文のように文頭で「根据本发明」と訳される方が望ましい。

3.2.4. 文の句切に関する問題点

機械翻訳による処理結果で中国文に不適切な句切が発生してしまう場合がある。以下にその代表的な例を挙げる。

例(マイクロ分析結果の実例データ B0456)

日本語 原文	ステープル1は、例えば、真っ直ぐな針金状のものを接着剤等で多数連結したシート状のものがマガジン（図示せず）内に積層状態で収納されており、その最上位に位置したシート状のものの最先に位置するステープル1を略コ字状に成形した後、その成形後のステープル1をドライバ2にて打ち込む際に、ドライバ2と 同期して下降する 成形板（図示せず）により次位のステープル1がコ字状に成形されているものである。
中国語 機械翻訳文	下次品位的ステープル1被形成ステープル1使在把笔直铁丝状东西多数用粘合剂连接起来了的座席状东西被用把层状态在杂志(不用图来解释)里收藏,并且使最先位于位于了那个最高品位的座席状ステープル1象形地诡计コ字形成了之后把那个成型之后的ステープル1在司机2敲进去的时候象形地コ字司机2和 同步,下跌的 成型板(不用图来解释)比方说。
対応する 中国特許文	订书钉1形成为例如用粘合剂将笔直的针状的订书钉连接成板状。该板状的连接的订书钉1以上下方向上堆积的状态被收容在料台(图中未示出)内。位于其最上方(或最下方)的板状连接的订书钉之中,将位于前后方向的最前沿的订书钉1形成大致为C字状之后,通过驱动器2将该成形成大致为C字状的订书钉1打入。此时,通过与驱动器2 同步下降的 成形板(图中未示出),将下一个的订书钉1形成大致为C字状。

上記例では、日本語の「同期して下降する」が、中国語では「同步,下跌」と訳されている。しかしながら、中国語でのカンマは不要である。ここでカンマが入る理由は、「～して～する」の「～して」と「～する」の間に更に別の動詞句が入ることを想定しているためと思われる。この例の「同期して下降する」のように間に何も入らない場合に対応する必要があるものと思われる。

3.2.5. 訳抜けに関する問題点

日本語を中国語に機械翻訳すると、日本語の原文の一部が中国語では抜け落ちてしまっている(以下、訳抜けという)場合がある。「訳抜け」は、開発者が意図しないもの(システムのバグ)である場合と、意図的に訳していない場合があつて、後者はどんな機械翻訳エンジンでも起こりうる。後述の3.2.6(1)で挙げる「もの」のように、形式的または冗長な表現と判断される場合には意図的に訳さないという手法が用いられるが、その判断を誤ると訳抜けとなってしまうことが起こりうるためである。

以下に2つの代表的な例を挙げる。

例1(マイクロ分析結果の実例データ E0059)

日本語 原文	屋外配電盤収納ボックス等の固定枠体側から扉を開閉するために、下記特許文献に見られる、図5と図6で示すL型のハンドルが扉側に取り付けられている。
中国語 機械翻訳文	由于图5和为开关门在下列专利文献能够被看见的图6显示型L车把被在门一侧屋外配电盘收藏箱的固定范围身体一侧安装。

対応する中国特許文	为了从屋外配电盘收纳箱等的固定框体侧对门进行开闭，有一种在门侧安装有下述专利文献中记载的、如图 5 和图 6 所示的 L 形手柄。
-----------	---

上記例 1 では、日本語での「収納ボックス等の」が、中国語では「收藏箱的」と訳されており、「等の」の訳が抜けている。対応する中国特許文のように、「等」と訳すべきである。

例 2(マイクロ分析結果の実例データ H2965)

日本語原文	配線基板 2 と、該配線基板 2 に固定された固体撮像素子 1 3 と、該固体撮像素子 1 3 に撮像面（有効画素領域面） 1 4 を覆うようにして固定された透光性蓋部 1 7 とを有する固体撮像装置において、前記配線基板 2 は厚さ方向に貫通する孔 3 を有し、該孔 3 に前記透光性蓋部 1 7 の少なくとも一部が嵌入され、前記配線基板 2 の接続端子 9 と前記固体撮像素子 1 3 の接続端子 1 5 とが接続されている。
中国語機械翻訳文	有上述电线敷设基础 2 在固体摄像装置有遮盖摄像面(有效像素领域表面) 14, 被被电线敷设基础 2 和那电线敷设基础 2 固定了的固体撮像素子 13 和該固体撮像素子 13 固定了的透光性盖子部 17 在厚度方向贯通的洞 3 盖子部 17 上述透光性在那洞 3 的被一部分至少嵌入, 并且端子连接端子 9 和上述电线敷设基础 2 的前記固体撮像素子 13 的连接 15 被连接。
対応する中国特許文	一种固态图像传感装置具有印刷电路板(2); 固定于印刷电路板(2)上的固态图像传感元件(13); 以及固定于固态图像传感元件(13)上以便覆盖着图像传感区(有效像素区)(14)的透光盖部分(17)。 印刷电路板(2)具有沿其厚度方向穿透印刷电路板(2)的孔(3), 透光盖部分(17)的至少一部分安装于所述孔(3)中, 并且印刷电路板(2)的连接端子(9)连接于固态图像传感元件(13)的连接端子(15)上。

上記例 2 では、日本語での「覆うようにして」が、中国語機械翻訳文では「遮盖」と訳されている。これは、日本語の「ようにして」に対応する中国語の訳語が存在していないからである。この場合の「ようにして」は対応する中国特許文のように「以便」などと訳される方が望ましい。

3.2.6. その他の問題点

上記 3.2.1.~3.2.5.で説明した例の他にも、日本語を中国語に機械翻訳する際の問題は多数存在する。以下にそのうちの 3 つの代表的な例を挙げる。

(1)原文にはあるが、訳さない方がよい場合

例えば、日本語の「～するものである」を機械翻訳すると、「もの」が「东西」と訳されてしまう場合がある。中国語の「东西」は、「物」、「件」等を意味し、無理に訳さない方がよい。

(2)原文にはないが、訳文には補った方がよい場合

日本語での特許文献の定型表現である「本発明は、・・・に関する。」は、中国語では「本發明涉及一种・・・。」と記載される事が多い。この「一种」は、日本語では「一種」を意味するが、日本語の「本発明は、・・・に関する。」の表現中に該当する語は存在しない。

しかしながら、中国語に機械翻訳する際には、この「一种」を含む慣用的な表現に訳される事が好ましい。

(3) 単語の切れ目で誤りが発生する場合

機械翻訳する際に、日本語の単語の切れ目が誤って解釈され、間違った中国語に訳されてしまう場合がある。以下に代表的な例を挙げる。

例(マイクロ分析結果の実例データ A0164)

日本語 原文	発明3のバッグは、発明1のバッグであって、袋部の上面及び下面の開口側縁にはロープ穴が 複数形成され 、ロープがロープ穴を挿通することで開口が開閉自在に綴じられており、閉塞具をロープが兼ねている。
中国語 機械翻訳文	发明3的手提包是发明1的手提包, 被绳索洞孔袋子部的外表以及开口一侧预先方面的 缘形成复数形式 , 并且因为绳索在绳索洞孔插通所以被把开闭自在地开口订起来, 并且绳索正兼任闭塞工具。
対応する 中国特許文	发明3中的鱼护是发明1中的鱼护, 其特征在于: 上述袋部分的上表面及下表面的开口边缘处 形成有多个绳孔 , 上述绳子穿过上述绳孔, 将上述开口可自由开闭地束起, 其中, 所述绳子兼用作上述封闭构件。

上記例では、日本語での「複数形成され」が、「複数」と「形成され」という単語の正しい切れ目で解釈されず、「複数形」と「成され」という誤った切れ目で解釈され、日本語の「複数形」に対応する「复数形式」と「成され」に対応する「形成」とに訳されてしまっている。

3.3. 既存の文献から知得される翻訳不備の問題点

3.3.1. 「中国語機械翻訳技術に関する調査」 [5]

中国語に関連する機械翻訳の問題点は、既存の文献にも記載されている。前記「3.2. 翻訳不備の要因分析」では、日本語から中国語への機械翻訳する際の問題点について列挙したが、ここでは、平成20年度に特許庁が実施した「中国語機械翻訳技術に関する調査」 [5] から知得される中国語から日本語への機械翻訳する際の種々の問題点を列挙してみる。

- (a) 中国語の「涉及～」が日本語では「～が触れて」と機械翻訳されている。「涉及～」は特許文献の常套句であり「～に関する」と訳すべきである。
- (b) 中国語の「本発明中」が日本語では「本発明に」と機械翻訳されている。特許文では「本発明において」と訳すべき。
- (c) 中国語の「其特征在于」が日本語では「その特徴によってあって」と機械翻訳されている。「其特征在于」は、特許文献の常套句であり「その特徴は以下のとおり:」と訳すべきである。
- (d) 中国語の「包含以下步骤」が日本語では「以下の手順を含んで」と機械翻訳されてしまう。「包含以下步骤」は、特許文献の常套句であり「以下の手順を含む:」と訳すべきである。

(e)中国語の「块」が日本語では「塊」と不自然な訳語に機械翻訳されている。技術分野にふさわしい訳語(例えば、グロック)が辞書に登録されていない。

(f)中国語の「一种～」が日本語では「一種～」と機械翻訳されている。「一种～」は、特許文献の常套句であるが、日本語ではあえて訳出する必要がない。

上記(a)～(f)までに挙げた例を考察してみると、翻訳の方向こそ、中国語から日本語へと本調査の場合(日本語から中国語への機械翻訳)と反対ではあるものの、前記「3.2. 翻訳不備の要因分析」で列挙した問題点と共通している事が分かる。

即ち、上記(a)～(d)の問題点は、前記「3.2.1. 機械翻訳の訳語の不備に基づく問題点」の「(1)一般的な訳語の他に特許文にふさわしい訳語がある場合」に該当し、上記(e)の問題点は、前記「3.2.1. 機械翻訳の訳語の不備に基づく問題点」の「(2) 一般的な訳語の他に専門的な訳語がある場合」に該当し、上記(f)は、「3.2.6. その他の問題点」の「(2)原文にはないが、訳文には補った方がよい場合」に該当している。

従って、本調査では、前記「3.2.1. 機械翻訳の訳語の不備に基づく問題点」で列挙した各問題点に関して、後述の「3.4.2. ミクロ分析」で実際の要因分析を実施した。

3.3.2. 特許版・産業日本語委員会報告書「産業日本語」[6]

また、平成 22 年度の特許版・産業日本語委員会報告書「産業日本語」[6]では、機械翻訳システムの改善だけでは限界があり、翻訳の対象となる日本語の特許関連文書の改善からのアプローチが不可欠となってきているとして、「特許明細書ライティングマニュアル第 0 版<準備編>」と「日英機械翻訳産業日本語<第 0.1 版>」を策定している。

「特許明細書ライティングマニュアル」は、海外への特許出願に取り組む企業が、翻訳コスト減少のため機械翻訳を活用するにあたり経験則として蓄積してきた工夫や留意点を、言語学的な裏づけに基づき体系的に整理し、手順だった作業の流れにまとめ上げる目的で作成された。

「日英機械翻訳産業日本語」は、日本語を正確に機械翻訳するための日本語表現に関する規約である。ここでは英語に翻訳するためとなっているが、機械翻訳の正確さを損なう大きな要因として挙げられている「日本語側の曖昧性」は、中国語に機械翻訳する際にも共通する問題である。

「曖昧性」についてここでは、「英語の例であるが」として次の英文で説明している。

I saw a tree with the telescope.

① 構文的曖昧性

“with the telescope” の係り先が “saw” であるか “tree” であるか曖昧。

② 語彙的曖昧性

“saw” の意味が「見る」の過去形であるか、「のこぎり引きする」という動詞の現在形であるか曖昧。

このような曖昧性を機械翻訳システムが誤解析した場合に誤訳につながるため、この規約では曖昧性の生じない日本語表現を規定する規約を最終目標としている。

3.3.3. 「中国における特許翻訳の現状」[7]

また、AAMT/Japio 特許翻訳研究会 シンポジウム発表資料「中国における特許翻訳の現状」[7]では、「誤訳率の高い問題」として、以下の 7 つを挙げている。

一. 多義語の選択ミスによる誤訳

日本語原稿：ローラ軸中心が移動することによりスライド移動する

元訳：由于其轴向中心运动而滑动

修正案：通过滚筒的轴心移动而滑动

※「により」は「由于」（原因・理由）ではなく「通过」（手段）が正しい。

二. 因果関係抜けによる誤訳

日本語原稿：（～繰り返し行う）ことで、前記第1の位置情報を取得する

元訳：获得所述第1位置信息

修正案：以次获得所述第1位置信息

※「以次」（順序通り、順番に）を補うことで文意を明らかにする。

三. 原文にない情報の付け加えによる誤訳

日本語原稿：画面を表示するための画面情報

元訳：显示一屏幕的屏幕信息

（一つの画面を表示するための画面情報）

修正案：显示屏幕的屏幕信息

※原文にない「一」を削除することにより不要な限定を避ける。

四. 修飾関係の乱れによる誤訳

日本語原稿：前記塗工層が帯電制御剤、抗菌剤、紫外線吸収剤および酸化防止剤のうち少なくとも一つを含有する

元訳：涂层包括至少一种电荷控制剂，抗菌剂紫外线吸收剂和抗氧化剂

（塗工層が少なくとも一種類の帯電制御剤、抗菌剤、紫外線吸収剤および酸化防止剤を含有する）

修正案：上述涂层至少包括电荷控制剂，抗菌剂，紫外线吸收剂和抗氧化剂中之一

※「少なくとも一つ」が「帯電制御剤、抗菌剤、紫外線吸収剤および酸化防止剤」のそれぞれを修飾するのではなく、それらの中の「少なくとも一つ」であることを明確にする。

五. 特許技術用語に対する理解の不十分さによる誤訳

日本語原稿：着脱

元訳：拆卸（取り外す）

修正案：装卸

※「着脱」は「拆卸」（取り外す）ではなく「装卸」が正しい。

六. 「同形詞¹²」をそのまま使用することによる誤訳

¹² 「同形詞」とは、日本語と中国語で漢字が同じ語のことである。意味が同じ場合もあれば、異なる場合もある。

日本語原稿：手段

元訳：手段

修正案：装置/设备/机构/单元/部

※日本特許の「手段」は中国特許では「装置/设备/机构/单元/部」等の「手段」以外の適切な語に翻訳されなければならない。

七. クレームの記載形式が違うことによる誤訳

日本語原稿：～ことを特徴とする半導体発光素子

元訳：一种半导体发光元件，其特征在于，具有：

修正案：一种半导体发光元件，具有：

※クレームは中国特許で定められたクレーム記載形式に則らなければならない。

こちらは人間の翻訳者による翻訳について述べているものと思われるが、人間の翻訳者が誤訳するポイントは、機械翻訳システムにおいてもルール化しづらいと考えられる。実際に、本章「3.2. 翻訳不備の要因分析」で述べた機械翻訳の問題といくつかの共通点が見られる。

例えば、「一. 多義語の選択ミスによる誤訳」は本章「3.2.1.(3) 訳し分けの問題がある場合」と、「五. 特許技術用語に対する理解の不十分さによる誤訳」は、「3.2.1.(1) 一般的な訳語の他に特許文にふさわしい訳語がある場合」と同種の問題である。

また、「二. 因果関係抜けによる誤訳」は、本章「3.2.5 訳抜け」で述べたように、意図的に訳さないという機械翻訳システムの判断が誤りであるという点が、翻訳者の原文に対する理解の浅さに共通するものがある。

3.4. マクロ分析とミクロ分析

本調査で実施した「マクロ分析」及び「ミクロ分析」について説明する。まず、「マクロ分析」は、全調査対象文献に関する統計的データを取得して行う概要分析である。本調査では、特に、IPCに基づく技術分野での統計的な差異に着目して「マクロ分析」を行った。これに対して「ミクロ分析」は、調査対象文献の各データを個別に調査していく詳細分析である。本調査では、「マクロ分析」によって全体の傾向を把握した上で、特に翻訳精度が著しく低いと思われる案件を抽出し、それらの特定の文献に対して「ミクロ分析」を行った。

3.4.1. マクロ分析

本調査では、調査対象文献の全データ(10,000件)の日本語文に対して形態素解析器¹³を用いて形態素解析を行い、「マクロ分析」を実施した。以下、その結果及び考察を翻訳不備の要因(以下、誤り指標という)に言及して述べていく。

3.4.1.1. 平均文字数

まず、下記図 3.4.1.1.-1 は日本国の文献データ 1 件当たりの文字数を技術分野別に示した表とグラフである。なお、文字数は、要約等を含めた公報全体の文字数を数えている(化学式又は数式等のイメージは対象外)。

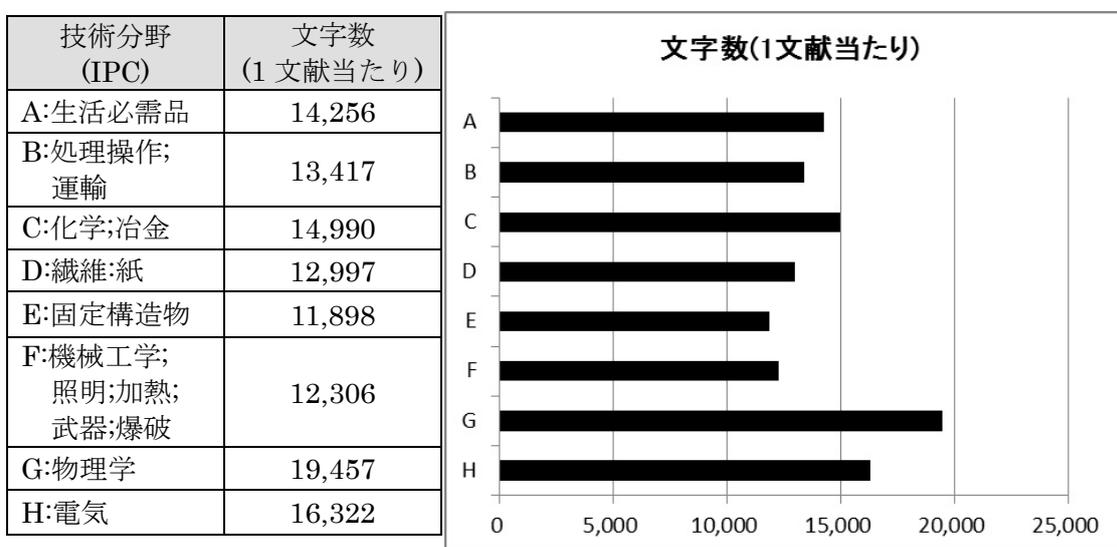


図 3.4.1.1.-1 文献当たり文字数

平均文字数が多い上位 3 技術分野は、順に G(物理学)、H(電気)、C(化学)となっており、中国の文献データに関する平成 21 年度調査の結果(例えば、明細書の平均文字数の調査結果等)と、若干の順位変動があるものの同じ結果が得られている。

これは中国への出願に特有の結果というよりは、特許文献の一般的な技術分野別の傾向を示していると考えられる。即ち、C(化学)に関する特許文献では、長い化学物質名が繰返

¹³ マクロ分析で利用した形態素解析器の詳細については「2.4.1.2. 手法」を参照。

して記述されている場合や、発明の詳細な説明に記載された実施例のデータ等が長い場合が多い。また、G(物理学)及びH(電気)の技術分野では、最近の傾向としてIT系/ビジネスモデル系の特許文献が多数含まれており、これらの説明は、カタカナ表記の外来語が用いられる事が多い為、その結果として文字数が比較的多くなっていることが一因として考えられる。

文字数が多い場合には翻訳不備が発生する可能性が高くなる為、上記3技術分野の翻訳内容に関しては特に注意する必要があると言える。

3.4.1.2. 動詞の数と複文出現率

翻訳をする場合、1文中に含まれる動詞の数が多ければ多いほど、その文はより複雑で翻訳の難易度も高くなっていると言える。同様に、文の構造が重文・複文¹⁴になっても翻訳の難易度は上がる。本調査では、複雑度を示す有効な指標として、1文当たりの動詞の数と複文の出現率(複文の出現率は、該当技術分野の複文の数/該当技術分野の全文数により算出した)とをIPC別に取得した。なお、本調査では、1文に動詞が2つ以上あるものを複文とみなしている。その結果を図3.4.1.2.-1及び図3.4.1.2.-2に示す。

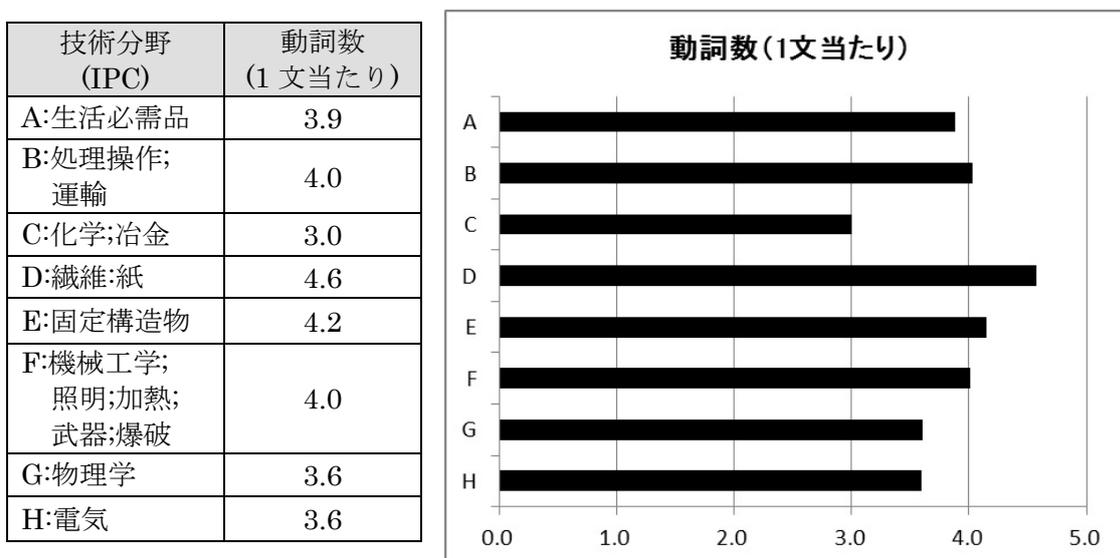


図 3.4.1.2.-1 文当たり動詞数

動詞の数に関してはいずれの技術分野においても1文当たりで3~4動詞とほぼ同じ個数であり、技術分野別の差異はさほど見られなかった。技術分野D(繊維;紙)は、1文当たりの動詞数が4.6と最も多くなっている。これは、マシン等、動作を説明する出願案件が多かった為とも考えられる。

¹⁴ 一文中で複数の述語が対等の関係になっているものが重文、主従関係があるものが複文であるが、文の複雑さの観点では両者に差異はないので区別せず、以下複文と表記する。

技術分野 (IPC)	複文出現率
A:生活必需品	79.50%
B:処理操作; 運輸	81.90%
C:化学;冶金	71.20%
D:繊維;紙	83.40%
E:固定構造物	82.00%
F:機械工学; 照明;加熱; 武器;爆破	82.00%
G:物理学	79.00%
H:電気	78.90%

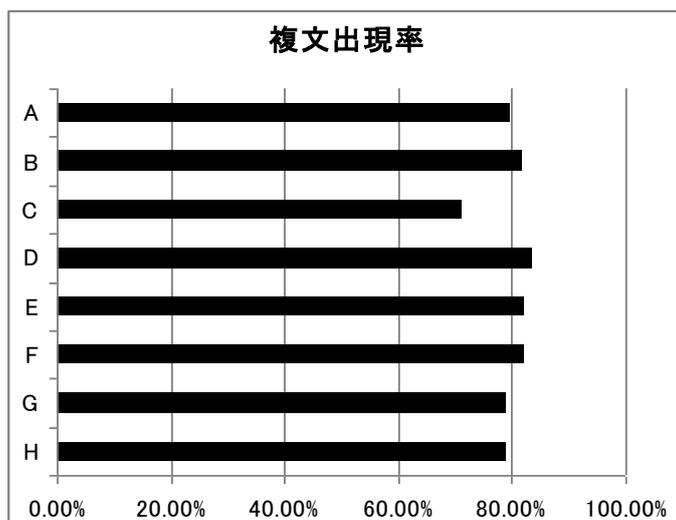


図 3.4.1.2.-2 複文出現率

上記図 3.4.1.2.-2 を見ると、複文出現率は、C(化学)で約 70%とやや低くなっている以外は、概ね 80%でほぼ同じ値になっている。上述した動詞の数の結果と同様に、複文出現率の結果からも C(化学)では文の構造自体は他の技術分野よりも単純である事が分かる。

3.4.1.3. 未知語の出現率

未知語に関しては技術分野による傾向がはっきりと出現した。下記図 3.4.1.3.-1において、未知語出現率は、(未知語を含む文節数/全文節数)である。技術分野 C(化学)と、化学に密接に関係があると思われる技術分野 D(繊維;紙)では、未知語の出現率が他の技術分野よりも高くなっていた。反対に技術分野 G(物理学)及び H(電気)では、出現率が低くなっていた。化学での説明では物質の固有名称等を多用する為である。残りの 4 技術分野では、概ね同じ出現率となっていた。

技術分野 (IPC)	未知語出現率
A:生活必需品	2.82%
B:処理操作; 運輸	2.79%
C:化学;冶金	4.20%
D:繊維;紙	3.93%
E:固定構造物	2.64%
F:機械工学; 照明;加熱; 武器;爆破	2.93%
G:物理学	1.86%
H:電気	1.98%

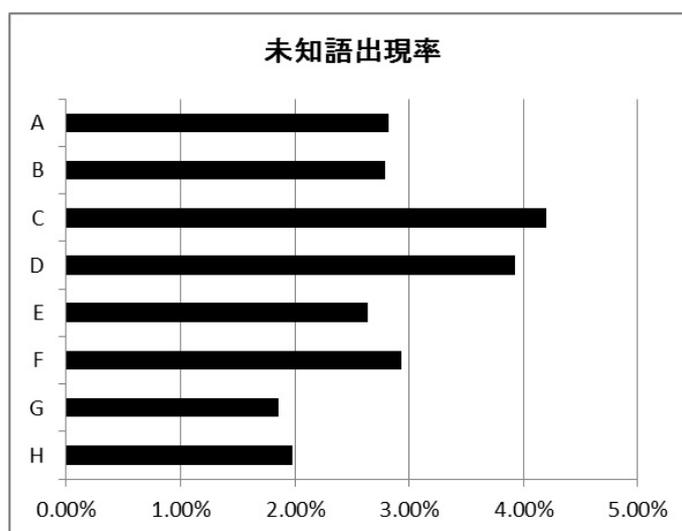


図 3.4.1.3.-1 未知語出現率

次に、ユニーク未知語出現率を調査した。ユニーク未知語出現率とは、同じ未知語の複数回の出現を1回としてカウントした場合の未知語出現率であり、これにより未知語の種類が多さがわかる。結果を下記図 3.4.1.3-2 に示す。これにより興味深い結果が得られた。下記表 3.4.1.3.-2 において、ユニーク未知語出現率は、(重複を排除した未知語を含む文節数/文節数)である。技術分野 C(化学)は他技術分野に比べて未知語の出現率が高いままであるが、化学に密接に関係があると思われる技術分野 D(繊維;紙)は、他の技術分野と同じ程度にまで値が下がる。即ち、技術分野 D(繊維;紙)では、同じ未知語を多数回用いている事が分かる。技術分野 G(物理学)及び H(電気)が、他技術分野に比べて未知語の出現率が低い状況には変化がなく、出現する未知語の種類も使用頻度も少ない事が分かる。

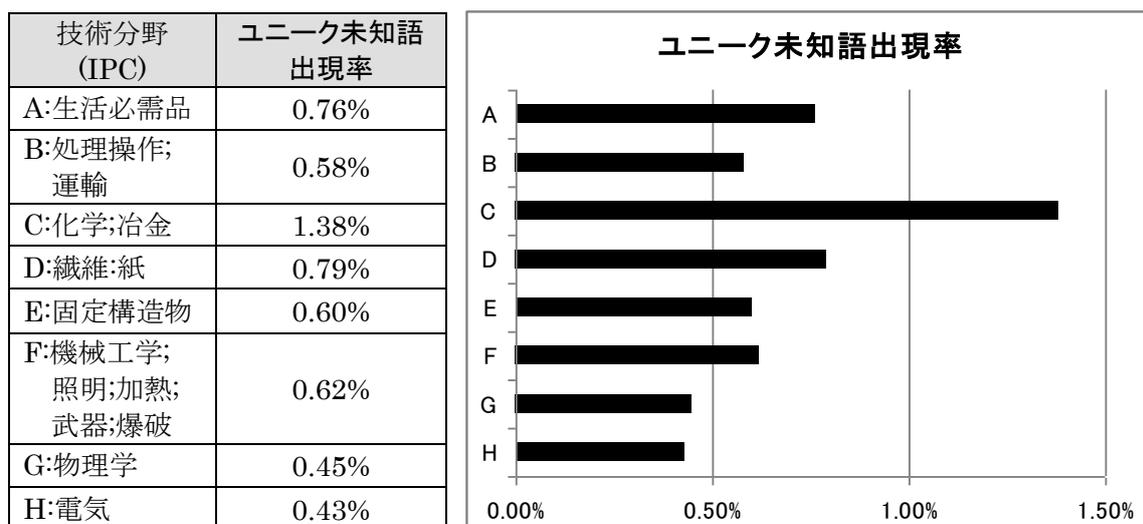


図 3.4.1.3.-2 ユニーク未知語出現率

3.4.1.4. ミクロ分析¹⁵用の文献データ抽出

「マクロ分析」の最後の作業として、調査対象文献の全データ(10,000 件)から、機械翻訳精度の評価の低い文献データ及び低くなる要因を有する文献データを各技術分野別(IPC 第 8 版 A~H セクション)に 5 件ずつ、翻訳不備案件として抽出し、ミクロ分析用の文献データとした。機械翻訳精度の評価が低くなる要因としては、上述した誤り指標の複文出現率及びユニーク未知語出現率を採用した。5 件を選択する際の選択基準の詳細は下記の通りである。

[翻訳不備案件の選択基準]

(基準 1)調査の都合上、原則として総文字数 5,000 文字以下の文献データより選択。

[注 1]

(基準 2)自動評価手法 BLEU¹⁶による評価値が小さい(即ち、機械翻訳精度が悪いと評価される)方から 2 件を選択。[注 2]

(基準 3)複文出現率の大きいものから 2 件を選択。(3.4.1.2 節を参照)

(基準 4)ユニーク未知語出現率の大きいものから 1 件を選択。(3.4.1.3 節を参照)

(基準 5)上記各基準の文献データが重複した場合には、一方の順位をずらして選択。

[注 3]

[注 1]

技術分野 E(固定構造物)に関しては、5,000 文字以下の条件では文献データ数が不足してしまう為、6,000 文字以下の条件で選択を行った。

[注 2]

自動評価手法 BLEU では、機械翻訳文の翻訳結果を評価する際に正解とみなされる基準翻訳文が必要になる。本調査では、この基準翻訳文として中国の文献データを用いたが、マクロ的な調査では個別の確認が難しい為、正解とみなされる中国の文献データが原文の日本の文献データと対応して同一内容であるかの判断基準は文の数の一致度合で設定した。具体的には、日本及び中国の文献データの文の数が概ね同程度(±2%以内)である場合に同一内容と判断した。

[注 3]

実際には、技術分野 D(繊維;紙)の文献データで(基準 5)に該当した。即ち、(基準 2)の BLEU の 2 番目に評価が悪い文献データが(基準 4)で選択される文献データと重複する為、(基準 2)では BLEU の 3 番目に小さい文献データを選択。また、(基準 3)の複文比率の大きさが 2 番目の文献データは、(基準 2)の文献データの選択と重複していた為、(基準 3)は複文比率の大きさが 3 番目の文献データを選択。

上記選択基準により得られた技術分野別の各 5 件、合計 40 件の文献データの一覧は次頁表 3.2.1.4-1 の通りである。これらの 40 件の文献データに対して、次の「3.4.2. ミクロ分析」にて詳細な分析を行った。

¹⁵ ミクロ分析の詳細は「3.4.2. ミクロ分析」を参照。

¹⁶ 代表的な自動評価手法であり、機械翻訳結果をツールを利用して 0~1.0 の値で自動評価する(1.0 に近い程、機械翻訳の精度の評価が高い)事が可能である。通常、4-gram で計算されるのが一般的である。詳細については、以下を参照。Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu : BLEU : a Method for Automatic Evaluation of Machine Translation, Proc. of ACL2002, 2002

表 3.4.1.4.-1 ミクロ分析用の文献データ

通し 番号	IPC	文献 データ名	日本出願番号	中国出願番号	日本出願の 発明の名称	選択 基準
1	A	A0164	2002-172356	03140746.3	魚収納バッグ	基準 2
2	A	A0369	2003-356425	200480030210.X	歯周病治療薬	基準 3
3	A	A0451	2004-380090	200510135228.2	串春巻き	基準 3
4	A	A0554	2004-095842	200510056266.9	食品用素材	基準 4
5	A	A0568	2004-215539	200510087451.4	乾燥ひじきの製造方法	基準 2
6	B	B0044	2003-061362	03107732.3	解繊繊維表面層を 有する発泡チップ製品 の製法	基準 2
7	B	B0392	2003-392535	200410077188.6	加工装置	基準 2
8	B	B0433	2003-033802	200480004121.8	管状脆性材料の内表面 研磨方法および 該研磨方法で得られた 管状脆性材料	基準 4
9	B	B0456	2003-285300	200480021776.6	ステープラ	基準 3
10	B	B1312	2005-154826	200610084663.1	自動車の前部車体構造	基準 3
11	C	C0322	2002-371866	200380107520.2	ミルペマイシン類の 精製法	基準 4
12	C	C0453	2003-186051	200410061786.4	無機塗料組成物	基準 3
13	C	C0886	2004-271743	200510103000.5	溶射前処理方法 およびエンジンの シリンダブロック	基準 3
14	C	C0942	2004-019399	200510006322.8	粘着剤層の貼合方法。	基準 2
15	C	C0978	2004-093425	200510006764.2	誘電体セラミック材料 及び積層セラミック基板	基準 2
16	D	D0032	2002-296048	03122290.0	ミシンの天びん	基準 3
17	D	D0044	2002-345587	200310116484.8	ミシンの針棒用 油排出装置	基準 2
18	D	D0121	2004-330936	200510116281.8	ミシン	基準 2,5
19	D	D0165	2004-214107	200510072625.X	水噴射式織機の 機上脱水装置	基準 4,5
20	D	D0171	2005-175883	200610091749.7	二重環縫いミシンに おけるスパンコール 縫付け方法	基準 3,5
21	E	E0007	2003-162658	200310101386.7	土嚢袋	基準 3
22	E	E0028	2003-327391	200410078653.8	ハンドル錠	基準 4
23	E	E0056	2003-290016	200410055843.8	桁橋の構築方法	基準 2
24	E	E0059	2004-193925	200510053937.6	扉用ハンドルの ロック装置	基準 2
25	E	E0090	2005-124670	200610005003.X	オーニング装置	基準 3
26	F	F0216	2003-320621	200410078513.0	ボールベアリング	基準 4
27	F	F0183	2003-322045	200480025943.4	内燃機関の 燃料供給装置	基準 3
28	F	F0514	2004-218925	200510089531.3	クロスフローファン	基準 2
29	F	F0629	2004-189593	200510082393.6	ビルトイン冷蔵庫	基準 3
30	F	F0650	2005-154866	200610077832.9	電動弁	基準 2

31	G	G0292	2002-327068	200310103560.1	自動視準機能と測距機能を有する測量機	基準 4
32	G	G0948	2003-325490	200410039387.8	自動販売機	基準 3
33	G	G1993	2004-126715	200510066373.X	映像表示装置及びその光源ユニット	基準 2
34	G	G2277	2004-124489	200510065093.7	撮像用光学ユニット	基準 2
35	G	G2581	2005-055108	200610008016.2	防水ハウジング	基準 3
36	H	H0276	2002-325036	200310103478.9	エキシマランプ	基準 4
37	H	H0469	2002-360894	200310124605.3	バンク巻線方法	基準 3
38	H	H0491	2002-251279	03153369.8	分電盤	基準 2
39	H	H1748	2003-291455	200480019175.1	ガラスチップ管の封止切断方法	基準 3
40	H	H2965	2005-058128	200610059404.3	固体撮像装置	基準 2

3.4.2. ミクロ分析

前節「3.4.1.4. ミクロ分析用の文献データ抽出」で準備したミクロ分析用の文献データに対して、個別にミクロ分析を行い、翻訳精度を低下させる要因を詳細に分析した。ミクロ分析の手法としては、平成 21 年度調査の経験を含む(株)クロスランゲージのこれまでの機械翻訳の問題点に関する知見を活かし、まず、前記「3.2. 翻訳不備の要因分析」で列挙した機械翻訳の問題点に留意し、日中機械翻訳の不備要因を以下のように分類した。

表 3.4.2.-1 機械翻訳不備の分類

不備要因の分類		日中機械翻訳の不備要因
大分類	小分類	
訳語	名詞	名詞の訳語が誤っている場合
	動詞	動詞の訳語が誤っている場合
	形容詞	形容詞の訳語が誤っている場合
	副詞	副詞の訳語が誤っている場合
	その他	その他の訳語が誤っている場合
助動詞	した	日本語の「～した」を誤って完了の意味に訳している場合
	される	日本語の「～される」を誤って受身の意味に訳している場合
	させる	日本語の「～させる」を誤って使役の意味に訳している場合
	その他	その他の助動詞を誤って訳している場合
位置	—	訳文中における訳出位置に問題がある場合
句切	—	カンマの位置が不適切な場合
訳抜け	—	訳されていない日本語がある場合
その他	—	上記以外の不備がある場合。 例) ・動詞の連用形がすべて「并且」「また」「かつ」の意の接続詞)と訳されているが不要 ・中国語で「～の」を意味する「的」が、「的的」のように連続している

前述した表 3.4.1.4.-1 に示すミクロ分析用の文献データを詳細に分析し、上記分類に基づいて分類した¹⁷。全 40 件の文献データの分類結果を集計し、各不備の割合を分析したところ、以下の図 3.4.2.-2～3.4.2.-4 に示す結果が得られた。

下記の図 3.4.2.-2 全不備要因の割合に、ミクロ分析用の文献データの全ての不備要因を不備要因の分類の大分類で分類した結果を示す。一方、下記の図 3.4.2.-3 不備要因「訳語」の内訳に、不備要因の分類の大分類が「訳語」に属する不備要因だけを小分類で分類した結果を示す。同様に、下記の図 3.4.2.-4 不備要因「助動詞」の内訳は、不備要因の分類の大分類が「助動詞」に属する不備要因だけを小分類で分類した結果を示す。

¹⁷ ミクロ分析の分類結果のサンプルは添付資料 1「ミクロ分析サンプル」を参照。

不備要因の分類 (大分類)	不備要因 の件数
訳語	2,962
助動詞	752
位置	538
句切	235
訳抜け	108
その他	516
合計	5,111

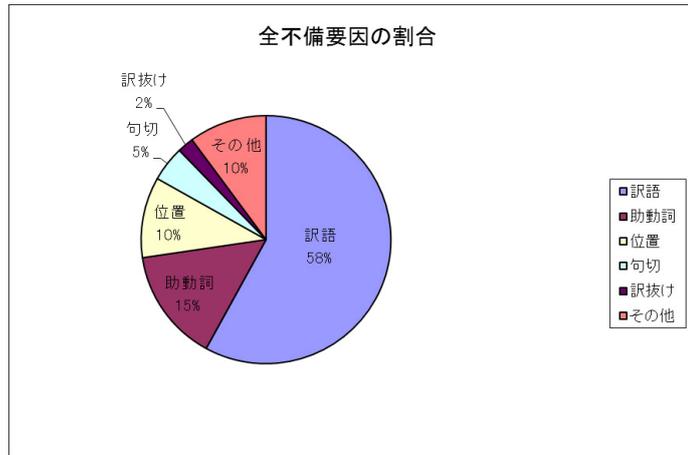


図 3.4.2.-2 全不備要因の割合

不備要因の分類 (「訳語」の小分類)	不備要因 の件数
名詞	1,804
動詞	772
形容詞	66
副詞	114
その他	206

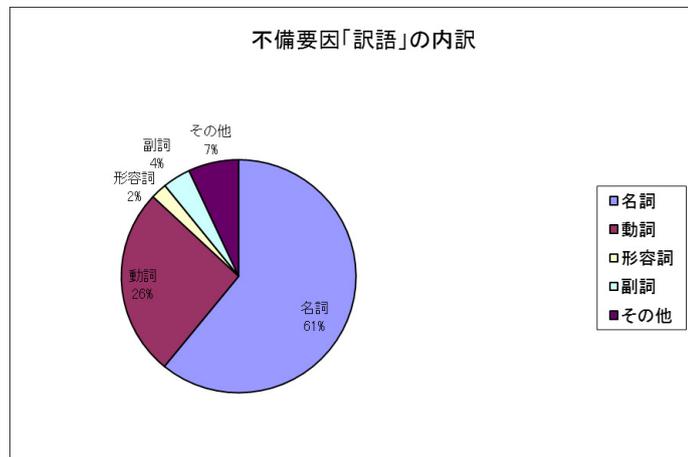


図 3.4.2.-3 不備要因「訳語」の内訳

不備要因の分類 (「助動詞」の小分類)	不備要因 の件数
した	155
される	311
させる	55
その他の助動詞	231

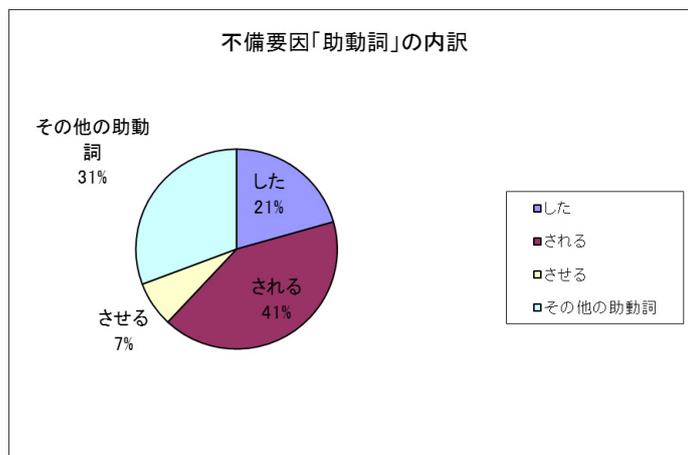


図 3.4.2.-4 不備要因「助動詞」の内訳

3.4.3. 各不備要因に対する改善策

前節「3.4.2. ミクロ分析」の表 3.4.2-1「機械翻訳不備の分類」で示した各不備要因に対しては、次のような改善策が考えられる。

不備要因の分類		改善策
大分類	小分類	
訳語	名詞	適切な訳語を名詞として辞書に登録する。
	動詞	適切な訳語を動詞として辞書に登録する。
	形容詞	適切な訳語を形容詞として辞書に登録する。
	副詞	適切な訳語を副詞として辞書に登録する。
	その他	適切な訳語を該当する品詞で辞書に登録する等。【注 1】
助動詞	した	機械翻訳エンジンで制御する。例えば「3.2.2. 助動詞に関する問題点」の(1)の例では、「活かしたまま」と語が続くので、「まま」が後に続く「～た」の場合は完了の意に訳さないようにする。「～したまま」の他に「～した場合」「～したら」とそのバリエーションが考えられるが、このようにある程度有限な数の助詞、助動詞に対応することでほぼ対応可能である。
	される	辞書にフラグを設けて機械翻訳エンジンで受身訳を出さないように対応する。例えば、「3.2.2. 助動詞に関する問題点」の(2)の例では、「挙げられる」の動詞「挙げる」にフラグを設けて機械翻訳エンジンで受身に訳さないようにする。同じタイプの動詞も同様に処理する。
	させる	機械翻訳エンジンで制御する。例えば、「3.2.2. 助動詞に関する問題点」の(3)の例では、使役表現で訳す必要がある表現（例えば「…によって」など）が原文にない場合に、使役を使わずに訳すようにする方法が考えられる。
	その他	機械翻訳エンジンで制御する。例えば、「3.2.2. 助動詞に関する問題点」の(4)の例では、特許文である事を考慮し、機械翻訳エンジンで訳す際に、「進行中」ではなく、「現在」で表現する。即ち、訳を出さないようにする。
位置	—	辞書に登録する。例えば、「3.2.3. 訳の位置に関する問題点」の例では、「本発明によれば」を副詞として登録し、文頭への訳出を指示する情報を付与する。
句切	—	機械翻訳エンジンで制御する。例えば、「3.2.4. 文の句切に関する問題点」の例では、「～して～する」のように動詞が密接している場合にはカンマを出さないようにする。
訳抜け	—	辞書に登録するか機械翻訳エンジンで制御する。例えば「3.2.5. 訳抜けに関する問題点」の例 1 では、「等の」を連体助詞として辞書に登録する。また、「3.2.5. 訳抜けに関する問題点」の例 2 では、「～するようにする」という表現のうち、「毎日朝食をとるようにする」のような意志を示す場合と「機械を動くようにする」のような使役を示す場合のうち、特許文であれば、使役の訳を出すようにする。
その他	—	状況に応じて対応する。【注 2】

表 3.4.3-1 機械翻訳不備の改善策

[注 1] 「3.2.1. 機械翻訳の不備の訳語に基づく問題点」の「(3)訳し分けの問題がある場合」に示す例の場合の改善策としては、対応する動詞「収納する」の辞書に「…に」の格助詞情報を追加する事等が考えられる。

[注 2] 一例として、「3.2.6. その他の問題点」の(1)～(3)の場合の改善策を各々、以下に挙げる。

- (1) 「ものである」を助動詞で扱うように機械翻訳エンジンを改良する。
- (2) 翻訳メモリのパターン文に登録する。これにより、原文にない語を訳出し、辞書よりも安全且つ容易に対処する事が可能になる。
- (3) 「複数形成する」を動詞として辞書に追加する。

第4章 定型化可能な表現の分析

4.1. 調査の目的

日本公開特許公報の各記載項目において、定型的に用いられる表現・フレーズを収集し、出現頻度の高い表現について分析を行い、定型化が可能な表現については対訳を付与する。本調査では、定型化が可能である場合と、定型化が困難である場合との両方の表現について分析する。定型化が可能である場合には、「定型文」、「定型パターン文」及び「定型フレーズ」を対訳で登録する事により、一方、定型化が困難である場合には、文の構造を解析して得られる「主文パターン文」及び「節パターン」を対訳で登録する事により、機械翻訳結果の品質向上が期待できる。「定型文」、「定型パターン文」、「定型フレーズ」、「主文パターン文」及び「節パターン」については下記の説明を参照されたい。

<定型化が可能な場合>

「**定型文**」は、複数の文献中に出現する文を指す。機械翻訳システムにおいては、翻訳メモリと呼ばれる機能を使用し、原文と訳文のペアを登録して、完全に一致する文が出現した場合に訳文を再利用する事で、翻訳結果の品質向上を行う事ができる。

「**定型パターン文**」は、頻度の高いパターン文で、パターン文とは文の一部が変数となる文を指す。例えば、飲食店でよく使われる言い回しに「本日のメインディッシュは～です。」という文があった場合、「～」に該当する部分が変数という事になる。なお、変数部分以外を固定部分と呼ぶ。機械翻訳システムにおいては、翻訳メモリ機能を使用し、その原文のパターンと訳文のパターンを登録して、パターンに一致する文が出現した場合には、その訳文パターンを再利用し、変数部分だけを機械翻訳する事で、翻訳結果の品質向上を行う事ができる。

「**定型フレーズ**」は、頻度の高いフレーズ(句)を指す。機械翻訳システムにおいては、名詞句・形容詞句・副詞句・動詞句等、単語より広い単位の句として辞書に登録する事により、機械翻訳の品質向上が期待できる。

<定型化が困難な場合>

「**主文パターン文¹⁸**」は、上記「定型パターン文」に似ているが、「定型パターン文」が単純に文全体の一部を変数として得られるパターンであるのに対し、「主文パターン文」は文を構成する柱となる主文のみに着目して得られるパターンである。

「**節パターン¹⁹**」は、上記「定型フレーズ」がフレーズ(句)に着目しているのに対し、主語と述語で構成される節に着目して得られるパターンである。

¹⁸ 主文パターンについての本調査での詳細な定義については「4.3.2.1.2. 主文パターンの定義」を参照。

¹⁹ 節パターンについての本調査での詳細な定義については「4.3.2.2.2. 節パターンの定義」を参照。

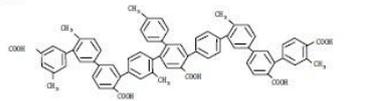
4.2. 調査の対象

日本公開特許公報(調査対象文献²⁰10,000件)の発明の名称、要約、特許請求の範囲及び明細書に対して分析を行う。以下に、日本公開特許公報イメージのうち、調査対象の部分を示す。

特許請求の範囲

【請求項1】
電圧回線を用いて相互通信を行うファクシミリ端末等により、相手端末に自端末の端末パラメータを通知し、通信時の端末パラメータを識別する方法において、端末パラメータを含む制御信号の送信端末は該制御信号のファクシミリ情報フィールドを、複数のサブフィールドに分断し、各サブフィールドの情報を送受信するファクシミリ情報フィールドのデータ中には現れない特定の識別コードを挿入してファクシミリ情報フィールド内の上記特定の識別コードを検出し、該ファクシミリ情報フィールドの情報の内容を解析し相手端末の端末パラメータの内容を抽出することを特徴とするファクシミリ端末パラメータ識別方式。

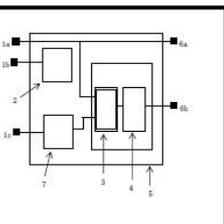
【請求項2】
請求項1の装置を用いた方法。

【発明の詳細な説明】
【技術分野】
本発明は簡単にして、装置機能のパラメータの記憶が容易なファクシミリ端末パラメータ識別方式に関するものである。
【化1】
 20
【背景技術】
【0002】
従来、電圧回線を用いて通信を行うファクシミリ装置においては、相互通信を可能とするため、国際電信電話諮問委員会(以下CCITTと記述する)において、電圧回線本線の複信化が行われ、一般電話設備網における互換型送受信ファクシミリ伝送手段として標準化(以後勧告T.30と記述する)されている。この勧告T.30の中で、アクロシンのデータモジュールを用い、制御信号の送受信を行うバイナリー半線は冗長度低圧化学処理を行うグループ3ファクシミリ装置に広く適用されている。
【特許文献1】特開2003-123456(P2003-123456A)
【特許文献2】特開2003-123456(P2003-123456A)
【0003】
このバイナリー半線を用いるファクシミリ装置には、デジタル識別信号(DIS信号と称されている)を用いて、国際的な標準化されているファクシミリ端末パラメータ、例えば伝送速度、解像度、符号化方式、送受信モード等を相手端末に通知する方法が提案されている。さらに、T.30には標準的な端末パラメータの他に、非標準の端末パラメータについても非標準ファクシミリ制御信号(NSF信号と呼ぶ)を用いて通知することも規定されている。
【発明の要旨】
【発明が解決しようとする課題】
【0004】
一方、端末技術の向上により、ファクシミリの高機能化、多機能化が行われ新しく開発されるファクシミリ装置には新しい端末パラメータを付加することが要求される。さらに高機能な装置に備わる装置との相互通信も要求される。

発明の名称

【発明の名称】 ファクシミリ装置装置

【要約】 (修正有)
【課題】 ファクシミリ端末パラメータ識別方法に關し、ファクシミリ装置機能のパラメータを容易にする。
【解決手段】 通信時の端末パラメータを識別する方法において、端末パラメータを含む制御信号の送信端末1a、1bは該制御信号のファクシミリ情報フィールドを、複数のサブフィールドに分断し、各サブフィールドの情報を送受信するファクシミリ情報フィールドのデータ中には現れない特定の識別コードを挿入してファクシミリ情報フィールドを分断する。制御信号の受信端末2はファクシミリ情報フィールド内の上記特定の識別コードを検出し、ファクシミリ情報フィールドの情報の内容を解析し相手端末の端末パラメータの内容を抽出する。装置機能のパラメータを記憶する場合はエコーコードを挿入して可変長の端末パラメータを分断する。送受信のユニコータローは該端末装置の製造会社または特許に該装置の制御信号の一致として読み出し専用メモリにインプリメントされるので、ハードウェア上の負担にはならない。
【権利】 図1



-43-

要約

明細書

20 「2.2. 対象文献の入手」を参照。

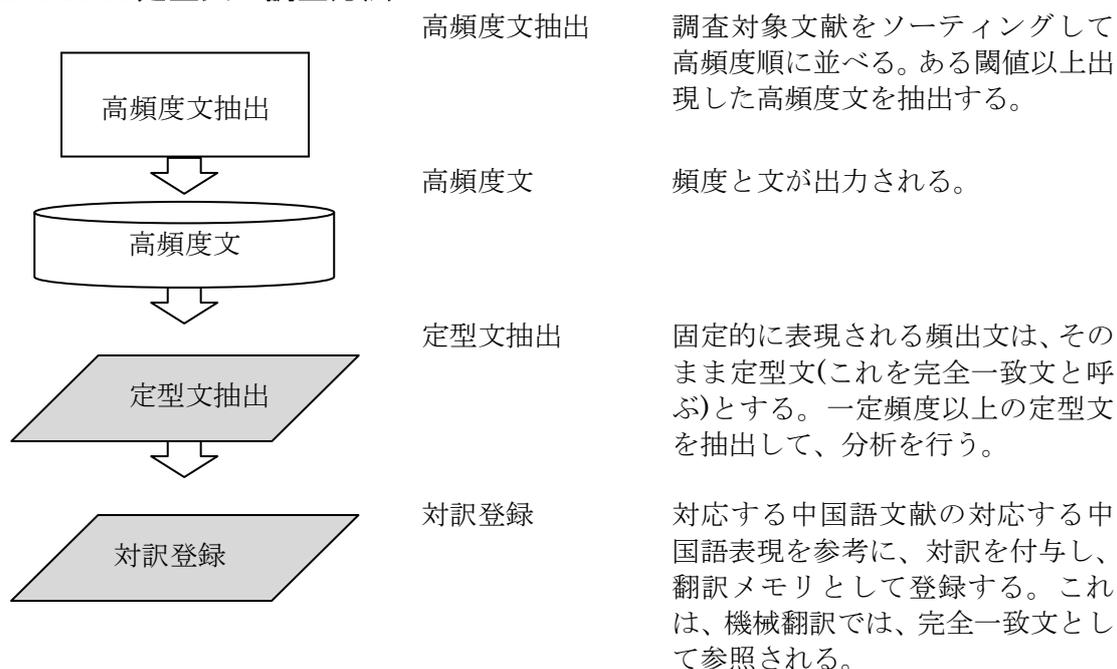
4.3. 調査の方法

4.3.1. 定型化が可能な場合

調査対象文献に含まれる文の定型化が可能であると仮定して、その出現頻度等から定型化を試みる。具体的には下記「4.3.1.1. 定型文」、「4.3.1.2. 定型パターン文」及び「4.3.1.3. 定型フレーズ」に記載の手法を用いて定型化を行った。

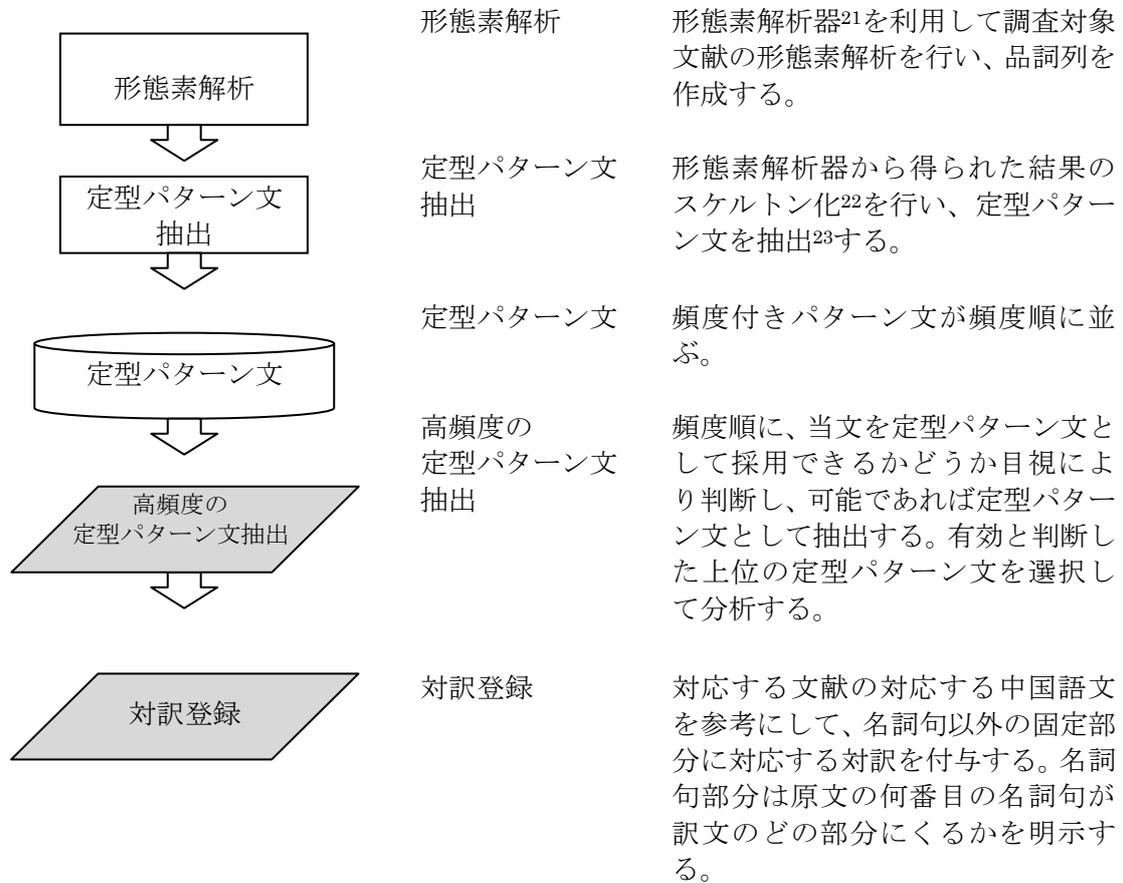
4.3.1.1. 定型文

4.3.1.1.1. 定型文の調査方法



4.3.1.2. 定型パターン文

4.3.1.2.1. 定型パターン文の調査方法



²¹ 形態素解析器の詳細については「2.4.1.2. 手法」を参照。

²² スケルトン化の詳細については「2.4.2.2.1. スケルトン化」を参照。

²³ 定型パターン文の抽出方法の詳細については「4.3.1.2.2. 定型パターン文の抽出方法の詳細」を参照。

4.3.1.2.2. 定型パターン文の抽出方法の詳細

[1]形態素解析

まず、大量の調査対象文献に対して、形態素解析器を用いて形態素解析を行う。以下に、形態素解析を実行した例を示す。

例文	図 1 は、電動油圧式パワーステアリング装置を示す要部断面図である。
形態素解析結果	図 1 (名詞)/は(助詞:係助詞ハ)/、(助詞:特殊助詞カンマ):連用 電動(名詞):活用形無し 電動(動詞):活用形無し 油圧(名詞):活用形無し 式(名詞):活用形無し 式(形容詞):活用形無し パワーステアリング(名詞):活用形無し 装置(名詞)/を(助詞:格助詞ヲ):連用 示す(動詞):連体 要(形容詞):活用形無し 要(名詞):活用形無し 部(名詞):活用形無し 断面図(名詞)/である(助動詞:断定助動詞デス):終止 。(記号):1

上記の例から分かるように、形態素解析結果は、文節ごとに各行に表示される。文節内の単語の区切りは「/」で示される。括弧の中には品詞が示される。スペースで区切られて複数の候補が並んでいる場合があるが、形態素解析段階で最も有力だと思われる候補順に左側から並べられるため、本調査では最左端の候補のみを使用する事とする。末尾には「:」の後に活用形が示される。

[2]定型パターン文抽出

このようにして得られた形態素解析結果に対してスケルトン化²⁴を行い、定型パターン文を抽出する。以下で「定型化パターン文」を抽出する為の具体的なパターン抽出手順を下記に示す。

- (手順 1)数字(数詞)を「#」で置換する。【注 1】
- (手順 2)名詞を「@」で置換する。【注 2】
- (手順 3)名詞を修飾する形容詞を削除する。
- (手順 4)以下のルール((手順 4-1)~(手順 4-5))で「@」をまとめあげる。
 - (手順 4-1)「@#」を「@」に置換する。【注 3】
 - (手順 4-2)「@の@」を「@」に置換する。
 - (手順 4-3)「#」を「@」に置換する。
 - (手順 4-4)「@・@」を「@」に置換する。
 - (手順 4-5)2個以上の「@」の連続を1つにまとめる。

<【注 1】~【注 3】の説明 ※説明の便宜上、【注 2】から先に説明する。>

【注 2】

文の骨格として残したい名詞(例えば、特許文献特有の語等)は、置換しない。以下にそのような名詞の例を挙げる。

本発明、本願、実施例、比較例、実施の形態、本実施形態、実施形態、発明の効果、本発明の効果、請求項、図、いずれか、記載、こと、もの、工程、結果、その結果、この結果、上記、前記、下記、以下、前述、上述、後述、ここ
※単なる例示列举で他にも多数存在

【注 1】

手順(1)で数字を一旦「@」以外の記号「#」に置換しているのは、例えばパターン例「**実施形態 1 のヘッド群を示す模式図。**」におけるパターン抽出結果が(a)「**実施形態@を示す@。**」ではなく、(b)「**実施形態@の@を示す@。**」となるようにする為である。記号「#」を利用する場合と記号「#」を利用しない場合の手順の差異を下記に示す。

【「#」を利用しない場合】

実施形態 1 のヘッド群を示す模式図。

↓(手順 2)

実施形態@の@を示す@。

↓(手順 4-2)

実施形態@を示す@。

・・・ (a)

²⁴ スケルトン化の詳細については「2.4.2.2.1. スケルトン化」を参照。

【「#」を利用する場合】

実施形態 1 のヘッド群を示す模式図。

↓(手順 1)

実施形態 # のヘッド群を示す模式図。

↓(手順 2)

実施形態 # の @ を示す @。

↓(手順 4-3)

実施形態 @ の @ を示す @。

・・・ (b)

【注 3】

(手順 4-1)は、数字を記号「#」に置き換えた副作用を抑制するもので、「@の@」や「@・@」を「@」にするのと同様に、「@#の@」や「@#@」を「@」にするためのものである。例えばパターン例「図 12 に、第 1 回転部材 46 の構成を示す。」におけるパターン抽出結果が(a)「図@に、@の@を示す。」ではなく、(b)「図@に、@を示す。」となるようにする為である。(手順 4-1)を実行する場合と(手順 4-1)を実行しない場合の手順の差異を下記に示す。

【手順(4-1)を実行しない場合】

図 12 に、第 1 回転部材 46 の構成を示す。

↓(手順 1)

図 # に、第 # 回転部材 # の構成を示す。

↓(手順 2)

図 # に、@ # @ # の @ を示す。

↓(手順 4-3)

図 @ に、@ @ @ @ の @ を示す。

↓(手順 4-4)

図 @ に、@ の @ を示す。

・・・ (b)

【手順(4-1)を実行する場合】

図 12 に、第 1 回転部材 46 の構成を示す。

↓(手順 1)

図 # に、第 # 回転部材 # の構成を示す。

↓(手順 2)

図 # に、@ # @ # の @ を示す。

↓(手順 4-1)

図 # に、@ @ の @ を示す。

↓(手順 4-2)

図 # に、@ @ を示す。

↓(手順 4-3)

図 @ に、@ @ を示す。

↓(手順 4-4)

図 @ に、@ を示す。

・・・ (a)

[3]高頻度の定型パターン文の抽出

上記「[2]定型パターン文抽出」の抽出処理によって、大量の定型パターン文が機械的に生成されるが、その中から高頻度で有効なもの²⁵を抽出する。最終的に、目視チェック及び訳付を経て正式に定型パターン文として認められる。

この際に、「2.4.2.2.2. パターン候補ファイルの作成」の手順によりパターン候補ファイルを生成し、このパターン候補ファイルを利用して定型目視チェック及び訳付作業を行った。

パターン文の訳付けでは、原文の何番目の変数が訳文のどこに対応するかが分かる形にする。ここでは、以下のような形式で、原文パターンと訳文パターンを作成する。

原文パターン：<\$1>が<\$2>であることを特徴とする請求項<\$3>に記載の<\$4>。

訳文パターン：如权利要求<\$3>所述的<\$4>，其特征在于，<\$1>为<\$2>。

これは、原文パターン側の<\$1>～<\$4>の部分に来る語は、それぞれ訳文パターン側の同じ数字の<\$1>～<\$4>の位置に来る語と対応することを意味している。

なお、この訳付けは、以降の各種パターン文のすべての訳付け作業で上記の形式を使用して行う。

²⁵ 有効でないものには、機械翻訳で問題なく訳せるものや、ある一文献の中だけで高頻度のものや、変数以外の部分が少なすぎるもの（例：「@を図る。」）などがあるので、それ以外を抽出する。

4.3.1.3. 定型フレーズ

4.3.1.3.1. 定型フレーズの調査方法

形態素解析器²⁶を利用して調査対象文献の形態素解析を行い、得られた形態素解析結果から、以下の手順で生成した。

(手順 1) 名詞の連続を 1 つの名詞にまとめる。【注 1】

(手順 2) 各語から始まる 2 語～10 語のフレーズを出力する。【注 2】

【注 1】

これは、名詞句を途中で分割してしまうフレーズを作らない為である。

【注 2】

但し、不要な要素を極力減らすため、以下のフレーズを出力しないように設定した。

(a)以下で始まるフレーズ

1. 助詞・助動詞
2. 記号

(b)以下を含むフレーズ

1. 句読点
2. 副詞
3. 主格(助詞「は」「が」)
4. 目的格(助詞「を」)
5. 並列(助詞「と」)
6. 記号

(c)以下で終わるフレーズ

1. 助動詞・助詞
2. 形容詞
3. 数字または数字+アルファベット
4. 助詞「を」+動詞で始まり、動詞以外で終わるフレーズ

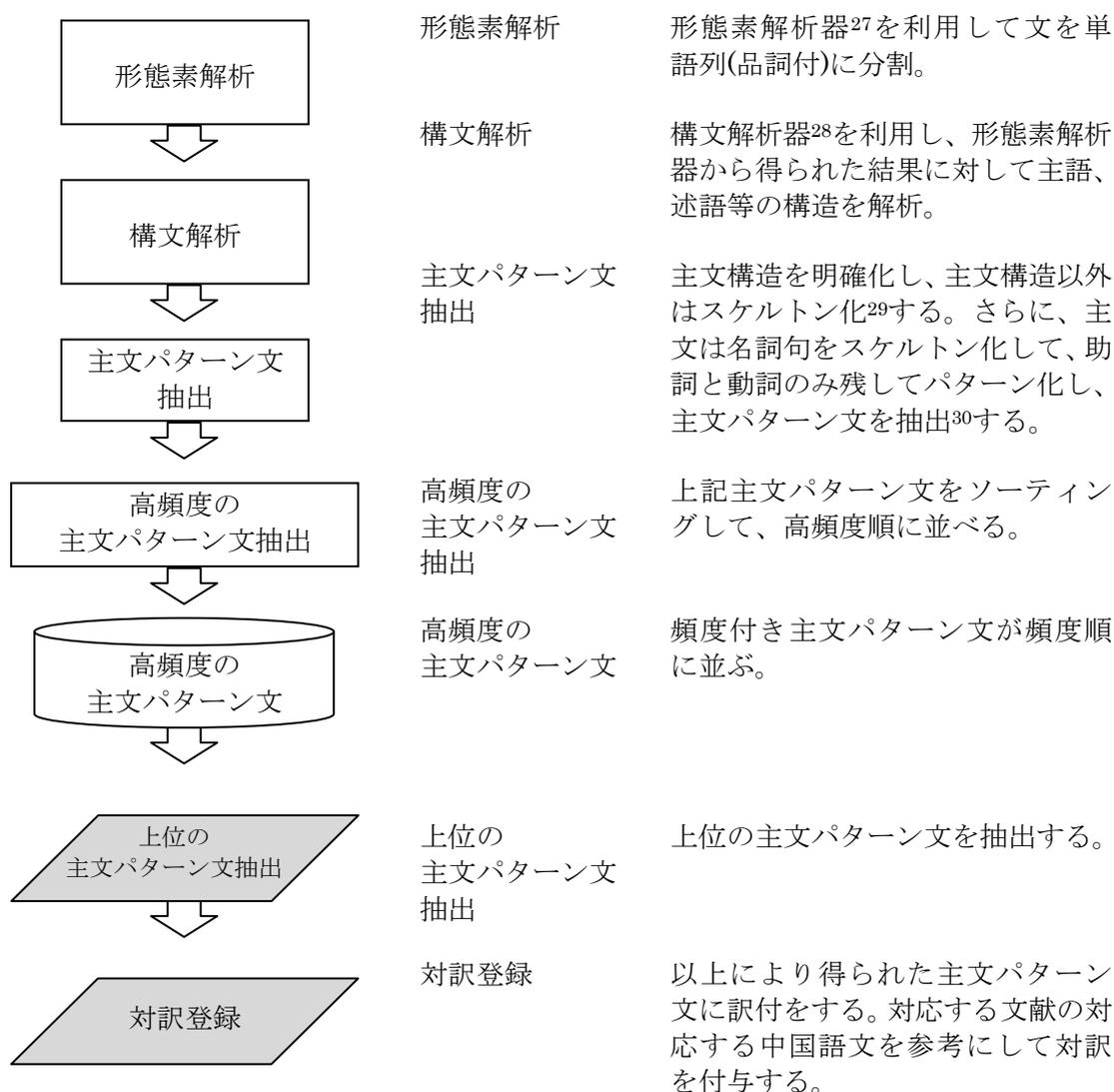
²⁶ 形態素解析器の詳細については「2.4.1.2. 手法」を参照。

4.3.2. 定型化が困難な場合

調査対象文献に含まれる文には、1文に複数の節が含まれる等、その構造が複雑で上述した「4.3.1. 定型化が可能な場合」の4.3.1.1.定型文～4.3.1.3.定型フレーズの手法では、定型化が困難な場合がある。そこで、そのように定型化が困難な場合を考慮して下記の手法でも定型化を試みた。

4.3.2.1. 主文パターン文

4.3.2.1.1. 主文パターン文の調査方法



²⁷ 形態素解析器の詳細については「2.4.1.2. 手法」を参照。

²⁸ 構文解析器の詳細については「2.4.1.2. 手法」を参照。

²⁹ スケルトン化の詳細については「2.4.2.2.1. スケルトン化」を参照。

³⁰ 主文パターン文の抽出方法の詳細については「4.3.2.1.3. 主文パターン文の抽出方法の詳細」を参照。

4.3.2.1.2. 主文パターン文の定義

ここでは、文の構造を考えた場合に、修飾語をすべて取り除いた骨格の構造を主文構造と呼ぶことにする。たとえば、「私は彼から借りた本をなくした。」であれば、「本」を修飾する「彼から借りた」を除いた、「私は本をなくした。」が主文ということになる。つまり、主節の文の動詞ひとつと、その動詞に直接かかる主語や目的語などのみを含む文が主文ということになる。

そして、ここでは、この主文構造を抽出し、パターンを作成する。これを主文パターン文と呼ぶことにする。

この主文パターン文を使用して翻訳することにより、骨格部分が固定訳となり、変数部分だけ翻訳して固定訳にはめ込んでいくことになる。

したがって、修飾語がたくさん付いた長い文でも、訳文の構造を破たんさせずに訳せるようになることが期待される。

なお、主文パターン文を使用した翻訳を行うことができる翻訳エンジンは、まだ実用化されていない。したがって、ここでは、上記のように主文パターン文の目的を述べ、以下にデータの作成方法を述べ、実際にデータを作成しておくことで、今後の翻訳エンジンの発展の一助としたい。

4.3.2.1.3. 主文パターン文の抽出方法の詳細

[1]形態素解析

前述の「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の場合と同様にして大量の調査対象文献に対して、形態素解析器を用いて形態素解析を行う。

[2]構文解析

上記「[1]形態素解析」により得られた形態素解析結果に対し、構文解析器を用いて構文解析を行う。以下に、構文解析を実行した例を示す。

例文	本発明は、クロック信号の間引き処理にともなう特性劣化を防止したタイミングリカバリ回路及び間引きクロック生成方法を提供することを目的とする。
構文解析結果	<pre> └1>"本発明 は、"(名詞:1)(3) └"クロック信号 の"(名詞)(11) └"間引き処理 にともなう"(名詞)(10) └"特性"(形容詞[限定形容詞])(10) └2>"劣化 を"(名詞)(9) └"防止し た"(動詞)(8) └"タイミングリカバリ回路"(名詞)(7) └"及び"(並列詞オヨビ)(6) └"間引き"<強連接格>名詞(6) └"クロック"<強連接格>名詞(6) └"生成"<強連接格>名詞(6) └2>"方法 を"(名詞)(5) └"提供する"(動詞)(4) └2>"こと を"(名詞)(3) └"目的とする"(動詞)(2) └"。(各種記号)(2) [文末](1) </pre>

上記の例において、末尾に書かれているカッコ内の数字は、何階層目であるか(階層レベルと呼ぶ)を示す。語は文節ごとに分けられており、たとえば「こと|を」のように、1行に複数の語が含まれる場合がある。また、文節は例えば「こと|を」であれば「名詞+助詞」に「名詞」というラベルがつけられているが、ここで名詞は先頭の要素のみを意味する。

[3]主文パターン文抽出

前述のようにして得られた構文解析結果に対して主文構造を明確化するように主文構造以外をスケルトン化³¹し、主文パターン文を抽出する。

以下で「主文パターン文」を抽出する為の具体的なパターン抽出手順を下記に示す。

(手順 1)階層レベルの最大値を NMAX とする。

(手順 2)変数 n に対し、n=3 から n=NMAX まで n を 1 ずつ増やしながら(手順 3)を繰り返す。

(手順 3)以下のルール((手順 3-1)~(手順 3-9))でパターンをまとめあげる。

(手順 3-1)全レベルの名詞と副詞を「@」に置換する。**[注 1]**

(手順 3-2)全レベルの形容詞で、名詞を修飾する形容詞を全て「@」に置換する。

(手順 3-3)レベル n+2 以上の語は品詞に関わらず「@」に置換する。

(手順 3-4)レベル n+1 では、項番号がついていない場合は、品詞に関わらず「@」に置換する。

(手順 3-5)全ての語を、原文の順番通りに結合する。

(手順 3-6)「@の@」を「@」に置換する。

(手順 3-7)「@・@」を「@」に置き換える。

(手順 3-8)2 個以上の「@」の連続を 1 つにまとめる。

(手順 3-9)前回出力した主文パターン文と異なる場合、主文パターン文を出力。

[注 1]

文の骨格として残したい名詞(例えば、特許文献特有の語等)は置換しない。具体例については、「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の「[2]定型パターン文抽出」の**[注 2]**を参照されたい。

以下に、上記パターン抽出手順を前記[2]構文解析の例文に対して実行した結果を示す。

変数 n の値	出力結果
n=3	本発明は、@ことを目的とする。
n=4	本発明は、@を提供することを目的とする。
n=5	-----※n=4 の場合と同じなので出力なし
n=6	本発明は、@及び@を提供することを目的とする。
n=7	-----※n=6 の場合と同じなので出力なし
n=8	本発明は、@を防止した@及び@を提供することを目的とする。
n=9	-----※n=8 の場合と同じなので出力なし
n=10	本発明は、@にともなう@を防止した@及び@を提供することを目的とする。
n=11	-----※n=10 の場合と同じなので出力なし

³¹ スケルトン化に関しては「2.4.2.2.1. 定型パターン文」の「[2]定型パターン文」を参照。

なお、通常は主文の動詞が階層レベル 2 に現れ、その動詞がとる主語や目的語などが階層レベル 3 に現れるため、最上位の述語だけを持つという本来の意味での主文パターン文であれば、階層レベル 3 だけの処理で済むのだが、ここでは全レベルの処理を行い、複数のパターンを抽出している。これは、例えば「本発明は、@を提供することを目的とする。」のように、厳密には主文ではないが頻度の高いパターンも抽出する為である。

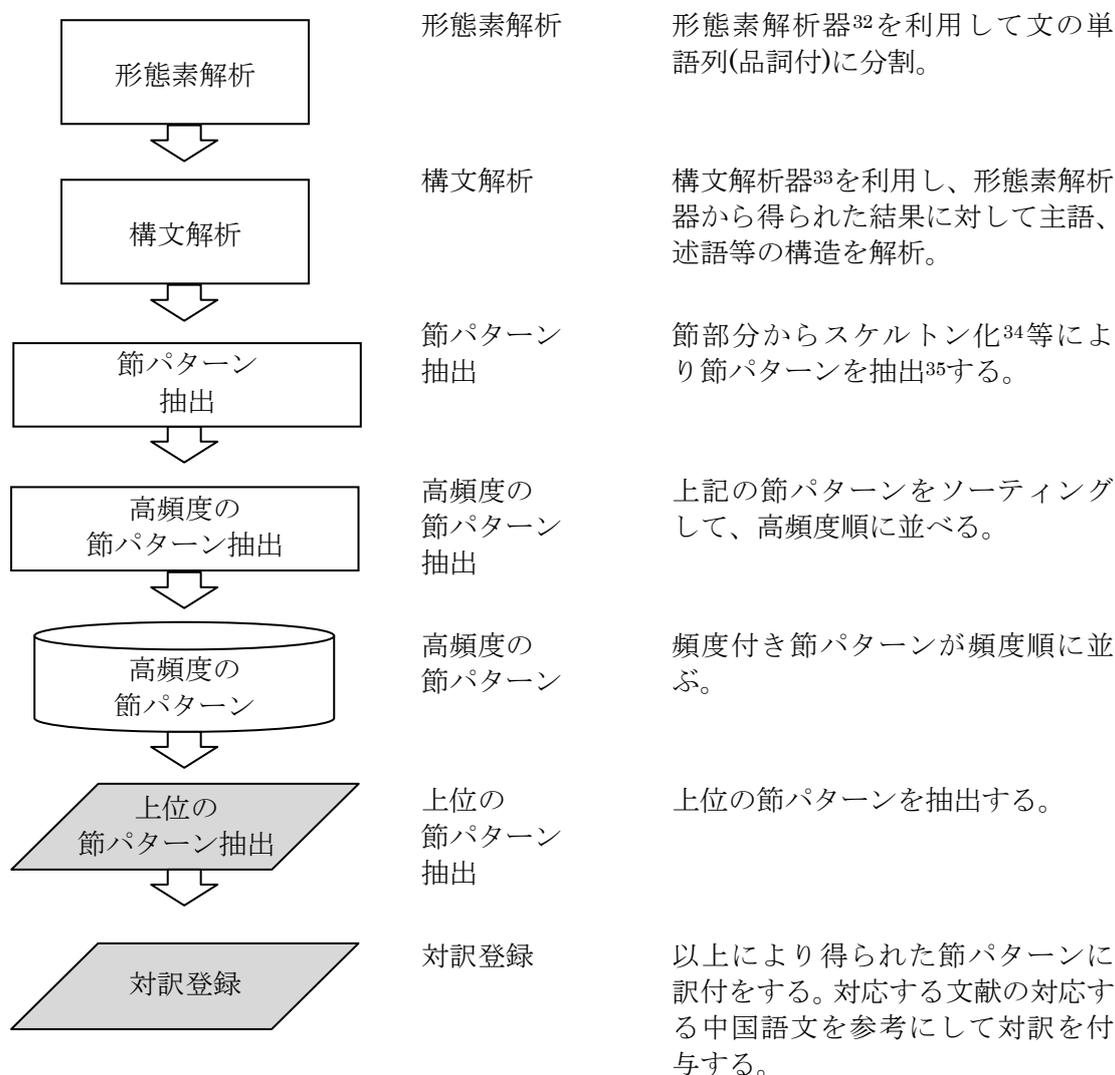
[4]高頻度の定型パターン文の抽出

上記「[2]主文パターン文抽出」の抽出処理によって、大量の主文パターン文が機械的に生成されるが、その中から高頻度で有効なものを抽出する。最終的に、目視チェック及び訳付を経て正式に主文パターン文として認められる。

この際に、「2.4.2.2.2. パターン候補ファイルの作成」の手順によりパターン候補ファイルを生成し、このパターン候補ファイルを利用して定型化目視チェック及び訳付作業を行った。

4.3.2.2. 節パターン

4.3.2.2.1. 節パターンの調査方法



※節パターンの定義は種々あるが、本調査では前述の「4.3.2.1. 主文パターン文」の変数部分に該当するものを節パターンと想定して解析を行った。次の「4.3.2.2.2. 節パターンの定義」にこの点に関しての詳細を記載した。

³² 形態素解析器の詳細については「2.4.1.2. 手法」を参照。

³³ 構文解析器の詳細については「2.4.1.2. 手法」を参照。

³⁴ スケルトン化の詳細については「2.4.2.2.1. スケルトン化」を参照。

³⁵ 節パターン文の抽出の詳細については「4.3.2.2.3. 節パターンの抽出方法の詳細」を参照。

4.3.2.2.2. 節パターンの定義

本調査での節パターンでは、名詞節と連用中止節と連体修飾節と副詞節を取り上げる。なお、節の定義には諸説あるのと、あらゆる節を解析するのは不可能であるため、ここでは、各節は以下の構造をとるものと定義する。

節の種類	節の構造
名詞節	「述語の項」 + 「動詞/形容詞(+助動詞)」 + 「名詞」
連用中止節	「述語の項」 + 「動詞/形容詞(+助動詞)」 + 「読点」
連体修飾節	「述語の項」 + 「動詞/形容詞(+助動詞)」 + 「読点」
副詞節	「述語の項」 + 「動詞/形容詞(+助動詞)」 + 「助詞」 + 「読点」

※ 「述語の項」は動詞や形容詞がとる主語や目的語のことを意味する。

※ ここでの連体修飾節は、動詞の直後に名詞を伴うものは除く。これは直後の名詞とともに名詞節として扱う。

例

- ・名詞節 : 断面が台形の溝を有するスペーサ
- ・連用中止節 : 優れた電荷輸送能力を有し、
- ・連体修飾節 : 脱水素酵素がグルコース脱水素酵素である、
- ・副詞節 : 硫酸はメッキ液の支持電解質であるので、

本調査では、節パターンを主文パターン文³⁶のパターンの変数部分にマッチするサブパターンとして利用することを想定している。

例えば、以下の原文を訳文に翻訳することを考えてみる。

原文	薄形化を図ることができる回転電機の固定子を提供する。
訳文	提供一种实现减薄的旋转电机的定子。

上記の原文の主文と、原文に含まれる節は以下のとおりである。

主文	原文	固定子を提供する。
	訳文	提供一种定子。
名詞節	原文	薄形化を図ることができる回転電機の固定子
	訳文	实现减薄的旋转电机的定子

これから抽出されるパターンは、以下のようになる。

主文 パターン文	原文	@を提供する。
	訳文	提供一种@。
節 パターン	原文	@を図ることができる@
	訳文	实现@的@

³⁶ 主文パターン文については「4.3.2.1. 主文パターン文」を参照。

原文に上記の主文パターン文を適用し、変数に対応する部分を墨付き括弧で示すと以下のようになる。

原文	【薄形化を図ることができる回転電機の固定子】を提供する。
訳文	提供一种【薄形化を図ることができる回転電機の固定子】。

ここで、さらに、変数内に上記の節パターンを適用すると以下のようになる。

原文	【薄形化】を図ることができる【回転電機の固定子】を提供する。
訳文	提供一种实现【薄形化】的【回転電機の固定子】。

最後に、残った変数内だけを機械翻訳し、翻訳結果とする。

このようにして、パターンの中のサブパターンとして節パターンを利用することにより、機械翻訳の精度向上を期待することができる。

なお、節パターン文を使用した翻訳を行うことができる翻訳エンジンは、まだ実用化されていない。したがって、ここでは、上記のように節パターンの目的を述べ、以下にデータの作成方法を述べ、実際にデータを作成しておくことで、今後の翻訳エンジンの発展の一助としたい。

4.3.2.2.3. 節パターンの抽出方法の詳細

[1]形態素解析

前述の「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の場合と同様にして大量の調査対象文献に対して、形態素解析器を用いて形態素解析を行う。

[2]構文解析

前述の「4.3.2.1.3. 主文パターン文の抽出方法の詳細」の場合と同様にして、得られた形態素解析結果に対し、構文解析器を用いて構文解析を行う。以下に、構文解析を実行した例を示す。

例文	記録層のキュリー温度を高く設定することにより繰り返し記録再生特性が劣化することなく、信号量が増大した光磁気記録媒体を提供する。
構文解析結果	<pre> └"記録層 "の"(名詞)(7) └2>"キュリー温度 "を"(名詞)(6) └"高く"(副詞)(6) └"設定する"(動詞)(5) └"こと "により"(名詞)(4) └"繰り返し"(副詞)(4) └"記録"<強連接格>(名詞)(5) └"再生"<強連接格>(名詞)(5) └1>"特性 "が"(名詞)(4) └"劣化する ことなく、"(動詞)(3) └"信号"<弱連接格>(名詞)(6) └1>"量 "が"(名詞<接尾>)(5) └"増大し "た"(動詞)(4) └"光磁気記録"<強連接格>(名詞)(4) └2>"媒体 "を"(名詞)(3) └"提供する"(動詞)(2) └"。"(各種記号)(2) [文末](1) </pre>

上記の構文解析結果の見方については、「4.3.2.1.3. 主文パターン文の抽出方法の詳細」の「[2]構文解析」を参照。

[3]節パターン抽出

前述のようにして得られた構文解析結果に対して節パターンを明確化するようにスケルトン化³⁷し、節パターンを抽出する。

以下で「節パターン」を抽出する為の具体的なパターン抽出手順を下記に示す。

(手順 1)構文解析結果の一番下の行から上に向かって走査していく。

(手順 2)レベル 3 以上で動詞または叙述用法の形容詞を見つけたら、以下のルール((手順 2-1)~(手順 2-10))で節パターンにする。

(手順 2-1)全レベルの名詞と副詞を「@」に置換する。**[注 1]**

(手順 2-2)全レベルの形容詞で、名詞を修飾する形容詞を全て「@」に置換する。

(手順 2-3)見つけた動詞又は形容詞のレベルを N とし、いったんその語句が節の末尾と仮定し、その行番号を Le(節の終了位置の語の行番号)とする。

(手順 2-4)動詞が修飾する名詞句の末尾を得るため、動詞が読点で終わらず、直下に名詞がある場合、品詞が名詞である間、下に走査していき、Le を最後の名詞の行番号に変更する。

(手順 2-5)動詞や形容詞がとる主語や目的語などを得るため、レベル N から始め、レベルが N 以上である間、上に走査していき、最後のレベルの行番号を Ls(節の開始位置の語の行番号)とする。

(手順 2-6)節の開始位置と終了位置が判明したので、その節を得るため、Ls 行目から Le 行目までの語を原文の順番通りに結合する。

(手順 2-7)「@の@」を「@」に置換する。

(手順 2-8)「@・@」を「@」に置換する。

(手順 2-9)2 個以上の「@」の連続を 1 つにまとめる。

(手順 2-10)節パターンを出力。

[注 1]

文の骨格として残したい名詞(例えば、特許文献特有の語等)は置換しない。具体例については、「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の「[2]定型パターン文抽出」の**[注 2]**を参照されたい。

[4]高頻度の節パターンの抽出

上記「[2]節パターン抽出」の抽出処理によって、大量の節パターンが機械的に生成されるが、その中から高頻度で有効なものを抽出する。最終的に、目視チェック及び訳付を経て正式に節パターンとして認められる。

この際に、「2.4.2.2.2. パターン候補ファイルの作成」の手順によりパターン候補ファイルを生成し、このパターン候補ファイルを利用して定型化目視チェック及び訳付作業を行った。

³⁷ スケルトン化に関しては「2.4.2.2.1. 定型パターン文」の「[2]定型パターン文」を参照。

4.4. 発明の名称

4.4.1. 定型化の分析及び結果

4.4.1.1. 分析

(1) 定型文についての分析

発明の名称については、同じ発明の名称を用いている日本国出願の文献データが多い為に、高頻度文(通常、**名詞句**である)が多数、存在する。また、技術分野別に異なる発明の名称が高頻度になる。例えば、技術分野 G(物理学)及び H(電気)に非常に高い頻度で出現する「半導体装置」は、他の技術分野には出現していない。

(2) 定型パターン文についての分析

通常、発明の名称は**名詞句**である為、抽出されたパターンは、大きく分けて以下の 3 種類に分類される。

(a) 名詞句の並列表現型

定型パターン文	@、@及び@
実例	【下地パネル】、【外壁化粧方法】及び【PCカーテンウォール】
模範訳	【基底面板】、【外壁装饰方法】和【预制幕墙】

定型パターン文	@と@、及び@
実例	【半導体素子】と【その製造方法】、及び【電子部品ユニット】
模範訳	【半导体元件】及【其制造方法】和【电子部件单元】

(b) 形容詞節で修飾された名詞句型

定型パターン文	@における@
実例	【エンジン】における【オイル通路構造】
模範訳	【发动机】中的【机油通路结构】

定型パターン文	@を有する@
実例	【キャスター】を有する【鞆】
模範訳	具有【脚轮】的【箱包】

(c) (a)と(b)の複合型

定型パターン文	@及びそれを用いた@
実例	【プラズマディスプレイパネル】及びそれを用いた【画像表示システム】
模範訳	【等离子显示板】以及使用它的【图像显示系统】

定型パターン文	@及び@を有する@
実例	【ドライバ回路】及び【ドライバ回路】を有する【システム】
模範訳	【驱动器电路】和具有【驱动器电路】的【系统】

(3) 定型フレーズについての分析

通常、発明の名称は**名詞句**なので、抽出されたフレーズも、ほぼ**複合名詞**に限られた。複合名詞の構造は、以下の形式が大半を占めた。

(a) (名詞+名詞)型

例：撮像装置（「撮像」＋「装置」）

(b) (名詞+名詞+名詞)型

例：固体撮像装置（「固体」＋「撮像」＋「装置」）

(c) (名詞+「の」+名詞)型

例：半導体装置の製造方法（「半導体装置」＋「の」＋「製造装置」）

ここでの**名詞**には**複合名詞**も含む。例えば、**(名詞+名詞)**は、さらに細分化すれば**((名詞+名詞)+名詞)**や**((形容詞+名詞)+名詞)**の場合もある。これは、機械翻訳用の形態素解析の辞書には、翻訳の精度向上のために、一般的に**名詞**として**複合名詞**も登録されている事に起因する。

技術分野ごとの特徴については、定型文同様、各分野に特徴的なフレーズが高頻度となった。たとえば、「遊技機」は技術分野 A(生活必需品)にのみ出現し、「ドラム式洗濯機」は技術分野 D(繊維;紙)にのみ出現していた。一方、分野を超えて共通して使用されるフレーズは少なく、「制御方法」や「製造方法」といったフレーズであった。

(4) 主文パターン文についての分析

発明の名称については、定型パターン文の 3 種類でほぼ網羅しており、それ以外の種類のパターンで有用なものは抽出されなかった。これは、発明の名称が構文的に複雑でなく、形態素解析結果から十分にパターンを作成できたことを意味している。

(5) 節パターンについての分析

発明の名称については、構文的に複雑でないため、抽出された節パターンは非常に少なかった。また、抽出されたのは、すべて**名詞節**で、**連用中止節**と**連体修飾節**と**副詞節**は抽出されなかった。その中でも高頻度のものを見ると、**動詞**としては「有する」、「用いる」及び「含有する」を用いたものが多く、また、「及び」を使用した並列表現が多かった。

節パターン	@を有する@
実例	【ケーブル部】を有する【回路基板の製造方法】
模範訳	设有【电缆部】的【电路基板的制造方法】

節パターン	@及び@に用いる@
実例	【カード接続構造】及び【それ】に用いる【カードコネクタ】
模範訳	【卡连接结构】及【其】使用的【卡连接器】

発明の名称について作成した定型文・定型パターン文・定型フレーズ・主文パターン・節パターンの結果の中から特に上位のものを抜粋して頻度順に以下に示す。

なお、発明の名称については、定型文と定型フレーズは、技術分野ごとに異なる表現が多いため技術分野のセクションごとに示すことにする。なお、セクション別の表の頻度はセクション内での頻度を示し、それ以外の表の頻度は全セクションの累計頻度を示す。

4.4.1.2. 定型文

A セクション：

日本語	中国語	頻度
遊技機	游戏机	21
食器洗い機	餐具清洗机	13
電気掃除機	电动吸尘器	13

B セクション：

日本語	中国語	頻度
空気入りタイヤ	充气轮胎	12
エアバッグ装置	气囊装置	9
記録装置	记录装置	8

C セクション：

日本語	中国語	頻度
研磨用組成物	抛光组合物	5
アクリル酸の製造方法	丙烯酸的制作方法	5
(メタ) アクロレイン又は (メタ) アクリル酸の製造方法	(甲基) 丙烯酸或 (甲基) 丙烯酸的生产方法	4

D セクション：

日本語	中国語	頻度
ミシン	缝纫机	16
洗濯機	洗衣机	11
ドラム式洗濯機	滚筒式洗衣机	10

E セクション：

日本語	中国語	頻度
建設機械	建筑机械	3
立体駐車装置	立体停车装置	2
建設機械の表示装置	工程机械的显示装置	2

F セクション：

日本語	中国語	頻度
空気調和機	空调器	17
冷蔵庫	冰箱	13
圧縮機	压缩机	10

G セクション：

日本語	中国語	頻度
画像形成装置	图像形成装置	46
液晶表示装置	液晶显示装置	40
表示装置	显示装置	37

H セクション：

日本語	中国語	頻度
半導体装置	半导体器件	75
半導体装置及びその製造方法	半导体器件及其制造方法	57
半導体装置およびその製造方法	半导体器件及其制造方法	41

4.4.1.3. 定型パターン文

日本語	中国語	頻度
<\$1>及び<\$2>	<\$1>以及<\$2>	1019
<\$1>および<\$2>	<\$1>以及<\$2>	759
<\$1>、<\$2>及び<\$3>	<\$1>、<\$2>以及<\$3>	202
<\$1>、<\$2>および<\$3>	<\$1>、<\$2>以及<\$3>	158
<\$1>、<\$2>、及び<\$3>	<\$1>、<\$2>、以及<\$3>	63
<\$1>と<\$2>	<\$1>和<\$2>	63
<\$1>、<\$2>、<\$3>および<\$4>	<\$1>、<\$2>、<\$3>以及<\$4>	62
<\$1>、<\$2>、および<\$3>	<\$1>、<\$2>、以及<\$3>	60
<\$1>及びそれをういた<\$2>	<\$1>以及使用它的<\$2>	58
<\$1>、<\$2>、<\$3>及び<\$4>	<\$1>、<\$2>、<\$3>以及<\$4>	57

4.4.1.4. 定型フレーズ

各セクション共通：

日本語	中国語	品詞	頻度
制御方法	控制方法	名詞	140
製造装置	制造装置	名詞	58
形成方法	形成方法	名詞	49

A セクション：

日本語	中国語	品詞	頻度
遊技機	游戏机	名詞	30
吸収性物品	吸收用物品	名詞	15
X線CT装置	X线CT装置	名詞	10

B セクション：

日本語	中国語	品詞	頻度
記録装置	记录装置	名詞	25
インクジェット記録装置	喷墨记录装置	名詞	24
画像形成装置	图像形成装置	名詞	21

C セクション：

日本語	中国語	品詞	頻度
プレス成形	冲压成型	名詞	14
アクリル酸の製造方法	丙烯酸的制作方法	名詞	12
プレス成形用プリフォーム	加压成型用预制体	名詞	10

D セクション：

日本語	中国語	品詞	頻度
ドラム式洗濯機	滚筒式洗衣机	名詞	10
洗濯乾燥機	洗衣干燥机	名詞	9
縫製装置	缝制装置	名詞	6

E セクション：

日本語	中国語	品詞	頻度
開閉装置	开关装置	名詞	3

F セクション：

日本語	中国語	品詞	頻度
空気調和機	空调机	名詞	31
燃料供給装置	燃料供给装置	名詞	13
内燃機関の制御装置	内燃机的控制装置	名詞	12

G セクション：

日本語	中国語	品詞	頻度
画像形成装置	图像形成装置	名詞	138
液晶表示装置	液晶表示装置	名詞	112
駆動方法	驱动方法	名詞	67

H セクション：

日本語	中国語	品詞	頻度
半導体装置の製造方法	半导体装置的制造方法	名詞	76
撮像装置	摄像装置	名詞	62
固体撮像装置	固体摄像装置	名詞	48

4.4.1.5. 主文パターン文

日本語	中国語	頻度
<\$1>及び<\$2>	<\$1>以及<\$2>	1780
<\$1>および<\$2>	<\$1>以及<\$2>	1298
<\$1>、<\$2>及び<\$3>	<\$1>、<\$2>以及<\$3>	268
<\$1>、<\$2>および<\$3>	<\$1>、<\$2>以及<\$3>	213
<\$1>並びに<\$2>	<\$1>以及<\$2>	204
<\$1>を用いた<\$2>	使用<\$1>的<\$2>	90
<\$1>、<\$2>、<\$3>及び<\$4>	<\$1>、<\$2>、<\$3>以及<\$4>	78
<\$1>及び<\$2>並びに<\$3>	<\$1>及<\$2>以及<\$3>	73
<\$1>、<\$2>、<\$3>および<\$4>	<\$1>、<\$2>、<\$3>以及<\$4>	69
<\$1>と<\$2>	<\$1>和<\$2>	68

4.4.1.6. 節パターン

日本語	中国語	頻度
<\$1>を有する<\$2>	具有<\$1>的<\$2>	40
<\$1>及び<\$2>に用いる<\$3>	<\$1>以及用于<\$2>的<\$3>	13
<\$1>、<\$2>、<\$3>及び<\$4>に用いる<\$5>	<\$1>、<\$2>、<\$3>以及用于<\$4>的<\$5>	8
<\$1>に用いる<\$2>	用于<\$1>的<\$2>	8
<\$1>を含有する<\$2>	含有<\$1>的<\$2>	8
<\$1>、<\$2>及び<\$3>に用いる<\$4>	<\$1>、<\$2>和用于<\$3>的<\$4>	7
<\$1>を用いる<\$2>	使用<\$1>的<\$2>	7
<\$1>及び<\$2>を有する<\$3>	<\$1>以及具有<\$2>的<\$3>	7
<\$1>、<\$2>、<\$3>、および、<\$4>を記録した<\$5>	<\$1>、<\$2>、<\$3>以及记录了<\$4>的<\$5>	4
<\$1>および<\$2>を用いる<\$3>	<\$1>和使用<\$2>的<\$3>	4

4.5. 要約

4.5.1. 定型化の分析及び結果

4.5.1.1. 分析

(1) 定型文についての分析

要約については、見出し以外の高頻度文がほとんど存在しない結果となった。これは、要約が要旨を記述する部分であり、その発明特有の内容を簡潔に記述するので、他の発明と共通して使用される文が書かれることが少ないためと考えられる。その結果、見出しに用いられる「【選択図】」、「【課題】」、「【解決手段】」、「図1」及び「図2」のような用語が高頻出の上位を占める結果になった。これらはどの技術分野でも共通して用いられる。

(2) 定型パターン文についての分析

要約は、大きく分けて以下の2種類が抽出された。

(a) 【課題】に頻出する表現型

定型パターン文	@を向上させることができる@を提供する。
実例	【サイクル特性】を向上させることができる【電解液】および【電池】を提供する。
模範訳	提供一种能够提高【循环特性】的【电解质溶液】和【电池】。

定型パターン文	@に優れた@を提供する。
実例	【押出成形加工性】に優れた【エチレン- α -オレフィン共重合体】を提供する。
模範訳	本发明提供【挤出成型加工性】优异的【乙烯- α -烯烃共聚物】。

定型パターン文	@を有する@を提供する。
---------	--------------

実例	【ゲルマニウム治療効果】を有する【眼鏡用鼻当て及び眼鏡】を提供する。
模範訳	提供一种具有【锗治疗效果】的【眼镜用鼻托及眼镜】。

(b) 【解決手段】に頻出する表現型

定型パターン文	本発明は、@に適用できる。
実例	本発明は、【画像表示装置】に適用できる。
模範訳	本发明可以应用于【图像显示装置】。

定型パターン文	@は、@と@とを備える。
実例	【この柄杓10】は、【カップ11】と【柄部12】とを備える。
模範訳	【该长柄勺(10)】具有【杯部(11)】和【柄部(12)】。

(3)定型フレーズについての分析

要約からは、**複合名詞**と**サ変動詞**(**動作名詞**＋「する」)が多く抽出された。なお、**複合名詞**については、発明の名称との重複も多く見られる。

サ変動詞については、一般にはあまり使わない専門用語としての**動詞**が多く見られる。ここでは、フレーズ(2語以上の複合語)を抽出しているため、たとえば「説明する」などのように、形態素解析器の辞書にすでに1語として登録されている一般的な**サ変動詞**は抽出しない。従って、抽出されるものには、形態素解析用辞書にない専門的な**サ変動詞**が多くなっていると考えられる。

(a) 名詞句

- 例1：本発明 (形容詞＋名詞) (※1)
- 例2：液晶表示装置 (名詞＋名詞) (※2)
- 例3：所定値 (形容詞＋名詞)

(b) 動詞句

- 例1：嵌合する (名詞＋動詞)
- 例2：当接する (形容詞＋動詞) (※1)
- 例3：対向配置する (名詞＋形容詞＋動詞)

※1 当調査に用いた形態素解析器では、**接頭辞**を**形容詞**として処理する。

※2 ここでの**名詞**には**複合名詞**も含む。例えば、(名詞＋名詞)は、さらに細分化すれば((名詞＋名詞)＋名詞)や((形容詞＋名詞)＋名詞)の場合もある。これは、機械翻訳用の形態素解析の辞書には、翻訳の精度向上のために、一般的に**名詞**として**複合名詞**も登録されている事に起因する。

技術分野ごとに見てみると、「本発明」や「当接する」のようなフレーズが全技術分野で用いられ、技術分野ごとに異なるフレーズとしては、たとえば、技術分野 B(処理操作;運輸)では「記録ヘッド」、技術分野 G(物理学)では「液晶表示装置」など、技術分野と関連が深いフレーズが用いられていた。

(4)主文パターン文についての分析

定型パターン文と同系統であるが、定型パターン文より複雑なパターンが抽出された。

主文パターン文	@は、@を備える。
実例	【このモータホルダ】は、【モータ本体の胴部14を収容する収容空間部32】を備える。
模範訳	【马达容纳件】具有【容纳马达主体部(14)的容纳空间部(32)】。

主文パターン文	@には、@が設けられている。
実例	【カーテンエアバッグ1】には、【ガイドロッド5に沿ってバッグの上端から下端まで延在する縦室30】が設けられている。
模範訳	在【窗帘气囊(1)】上设有【沿导向杆(5)从气袋上端延伸到下端的纵向室(30)】。

これらは、変数部分に**動詞**を含むため、通常の定型パターン文では抽出されなかったものである。厳密に言えば、より具体的な(変数部分の多い)パターン文として抽出されたため、頻度が低くなり、上位に残らなかったものである。

(5)節パターンについての分析

要約では、**名詞節**、**連用中止節**及び**副詞節**の順に頻度が高くなった。なお、**連体修飾節**は、ほとんど抽出されなかった³⁸。

(a)**名詞節**は**動詞**の「有する」を使用したものが飛びぬけて多く、以下、「構成する」、「設けられる」、「形成される」及び「応じる」が続いた。

節パターン	@を有する@
実例	【挿入部】を有する【内視鏡】
模範訳	具有【插入部】的【内窥镜装置】

(b)**連用中止節**は、「を有し」、「に優れ」、「が高く」、「が少なく」及び「とし」等の**動詞**／**形容詞**を使用したものが高頻度となった。

節パターン	@を有し、
実例	【所定の位置決め精度】を有し、
模範訳	具有【预定定位精度】

(c)**副詞節**では「であって」、「を抑制しつつ」、「を維持しつつ」等の**動詞**を使用したものが高頻度となった。

節パターン	@を抑制しつつ、
実例	【機械的強度の低下】を抑制しつつ、
模範訳	抑制【机械强度的降低】、

³⁸ ここでは連体修飾節は読点を含むものを扱っているため、例えば「前記シート状外包体が略矩形である、請求項10に記載のシート状二次電池セルの製造方法。」のように長い複合名詞を修飾する動詞の連体形とともに用いて、体言止めの文に使用されることが多い。このような文は請求項にはよく見られるが、それ以外ではあまり見られないため、抽出されなかったものと考えられる。

要約について作成した定型文・定型パターン文・定型フレーズ・主文パターン・節パターンの結果の中から特に上位のものを抜粋して頻度順に以下に示す。

なお、要約については、定型フレーズは、技術分野ごとに異なる表現が多いため技術分野のセクションごとに示すことにする。

4.5.1.2. 定型文

日本語	中国語	頻度
【選択図】	[选择图]	9917
【課題】	[课题]	9843
【解決手段】	[解决手段]	9808
図 1	图 1	4749
図 2	图 2	1415
なし	无	895
図 3	图 3	794

4.5.1.3. 定型パターン文

日本語	中国語	頻度
図<\$1>	图<\$1>	57
本発明は、<\$1>に適用できる。	本发明可以应用于<\$1>。	10
本発明は、<\$1>に適用することができる。	本发明可以应用于<\$1>。	8
<\$1>に優れた<\$2>を提供する。	提供<\$1>优异的<\$2>。	8
<\$1>は、<\$2>と<\$3>とを備える。	<\$1>具有<\$2>和<\$3>。	8
<\$1>を有する<\$2>を提供する。	提供一种具有<\$1>的<\$2>。	8
図<\$1> (<\$2>)	图<\$1> (<\$2>)	6
<\$1>が<\$2>を提供する。	<\$1>提供了<\$2>。	6
<\$1>は、<\$2>を有する。	<\$1>具有<\$2>。	6
<\$1>および<\$2>を提供する。	提供<\$1>以及<\$2>。	5

4.5.1.4. 定型フレーズ

各セクション共通：

日本語	中国語	品詞	頻度
本発明	本发明	名詞	1132
当接する	接触	動詞	256
回動	转动	名詞	171

A セクション：

日本語	中国語	品詞	頻度
被検体	受检体	名詞	46
遊技機	游戏机	名詞	41
遊技者	游戏者	名詞	39

B セクション :

日本語	中国語	品詞	頻度
記録ヘッド	记录头	名詞	38
回動する	转动	動詞	37
シート材	片材	名詞	30

C セクション :

日本語	中国語	品詞	頻度
粘着剤層	粘合剂层	名詞	46
含有してなる	包含	動詞	36
成形体	成形体	名詞	32

D セクション :

日本語	中国語	品詞	頻度
回転ドラム	转鼓	名詞	22
回転駆動する	旋转驱动	動詞	12
駆動手段	驱动方式	名詞	10

E セクション :

日本語	中国語	品詞	頻度
回動可能	可转动	名詞	9
嵌合する	接合	動詞	8
係止する	锁定	動詞	8

F セクション :

日本語	中国語	品詞	頻度
空気調和機	空调机	名詞	40
連通する	连通	動詞	34
燃料噴射弁	燃料喷射阀	名詞	33

G セクション :

日本語	中国語	品詞	頻度
液晶表示装置	液晶表示装置	名詞	160
画像形成装置	图像形成装置	名詞	149
消費電力	功率消耗	名詞	88

H セクション :

日本語	中国語	品詞	頻度
絶縁膜	绝缘膜	名詞	105
ゲート電極	栅电极	名詞	89
出力信号	输出信号	名詞	78

4.5.1.5. 主文パターン文

日本語	中国語	頻度
<\$1>を提供する。	提供<\$1>。	1565
<\$1>ことを特徴とする<\$2>。	<\$2>, 其特征在于: <\$1>。	161
<\$1>及び<\$2>を提供する。	提供<\$1>以及<\$2>。	151
<\$1>および<\$2>を提供する。	提供<\$1>以及<\$2>。	122
<\$1>である。	是<\$1>。	103
<\$1>は、<\$2>とを備える。	<\$1>具有<\$2>。	98
<\$1>ことを目的とする。	目的在于: <\$1>。	97
<\$1>を提供することを目的とする。	目的在于提供<\$1>。	72
<\$1>を得る。	得到<\$1>。	69
<\$1>とを備える。	具备<\$1>。	65

4.5.1.6. 節パターン

日本語	中国語	頻度
<\$1>を有する<\$2>	具有<\$1>的<\$2>	529
<\$1>を構成する<\$2>	构成<\$1>的<\$2>	138
<\$1>に設けられた<\$2>	设置于<\$1>上的<\$2>	106
<\$1>に形成された<\$2>	形成于<\$1>上的<\$2>	97
<\$1>に応じた<\$2>	响应<\$1>的<\$2>	96
<\$1>を有し、	具有<\$1>，	75
<\$1>を含有する<\$2>	含有<\$1>的<\$2>	74
<\$1>を形成する<\$2>	形成<\$1>的<\$2>	69
<\$1>が形成された<\$2>	形成有<\$1>的<\$2>	60
<\$1>となる<\$2>	作为<\$1>的<\$2>	58
⋮		
<\$1>に優れ、	<\$1>优秀，	42
<\$1>が高く、	<\$1>高，	41
⋮		
<\$1>が少なく、	<\$1>少，	32
⋮		
<\$1>であって、	是<\$1>，	19
⋮		
<\$1>を抑制しつつ、	抑制<\$1>，	14
⋮		
<\$1>を維持しつつ、	维持<\$1>，	10

4.6. 特許請求の範囲

4.6.1. 定型化の分析及び結果

4.6.1.1. 分析

(1) 定型文についての分析

特許請求の範囲については、高頻度文は皆無といってよい。これは、通常、特許請求の範囲の文は全てその発明の構成に欠くことのできない事項を述べる文になっており、また、他の請求項を引用することが多いこともあり、他の文献と一字一句同じ特徴や請求項の番号を持つことがまれであるためと考えられる。

(2) 定型パターン文についての分析

特許請求の範囲からは以下のようなパターンが抽出された。

定型パターン文	@が@である請求項@記載の@。
実例	【シリカ粒子】が【コロイダルシリカ粒子】である請求項【1】記載の【研磨液組成物】。
模範訳	如权利要求【1】所记载的【抛光液组合物】，其中【硅石粒子】是【胶体硅石粒子】。

定型パターン文	前記@は、@であることを特徴とする請求項@記載の@。
実例	前記【入力映像信号】は、【受信装置又はセットトップボックスからの信号】であることを特徴とする請求項【5】記載の【表示装置】。
模範訳	根据权利要求【5】的【显示设备】，其特征在于，所述【输入视频信号】是【从接收机或机顶盒发出的信号】。

これらの例にはいくつか**名詞句**が含まれているが、それらは、意図的に残したものである。特許請求の範囲から有用なパターンが取れるように、「請求項」や「前記」等のような、特許請求の範囲でよく使用される単語を残すようにした。その他の具体例については、「4.3.1.2.2. 定型パターン文の抽出方法の詳細」の「[2]定型パターン文抽出」の[注 2]を参照されたい。

(3) 定型フレーズについての分析

特許請求の範囲からは大きく分けて 2 種類のフレーズが多く抽出された。

(a) 名詞句型

例 1：絶縁膜 (名詞+名詞)

例 2：半導体装置の製造方法 (名詞+助詞+名詞)

例 3：情報処理装置 (名詞+名詞)

(b) 請求項に関するフレーズ型

例 1：請求項 1 に記載 (名詞+名詞+助詞+名詞)

例 2：請求項 1 記載 (名詞+名詞+名詞)

例 3：いずれか 1 項に記載 (名詞+名詞+助詞+名詞)

上記(a)については、発明の名称及び要約に共通して出現することが多い。また、この種のフレーズは、技術分野ごとに異なるフレーズが多い。

上記(b)については、全技術分野で共通して使用されている表現である。なお、この種のフレーズは、数字を含んでいる為に、通常は辞書に登録するには向かないと考えられる。しかし、請求項に特徴的な定型的な表現である上、助詞「に」の係り先を間違えて解析する可能性があるため、あえて数字部分を含め³⁹、さらに次に出現する助詞「の」も含めて、以下のように辞書に登録すると、訳質の向上が期待できる。

■辞書登録例 1

見出し語	請求項 1 に記載の
訳語	権利要求 1 所述的
品詞	連体詞

■辞書登録例 2

見出し語	請求項 1 記載の
訳語	権利要求 1 所述的
品詞	連体詞

■辞書登録例 3

見出し語	いずれか 1 項に記載の
訳語	任一项所述的
品詞	連体詞

※ **連体詞**は、**名詞**を修飾する点で**形容詞**の連体形に似ているが、活用がなく、終止形を持たない。

(4)主文パターン文についての分析

特許請求の範囲の定型パターン文では、**動詞**が「である」のものが多く抽出されたが、主文パターン文ではその**動詞**部分を変数内に含むため、より広範囲に適用可能なパターン文が抽出された。

主文パターン文	@ことを特徴とする請求項@に記載の@。
実例	【サーマルヘッドを用いて前記可逆表示部に画像を形成する】ことを特徴とする請求項【16】に記載の【画像処理方法】。
模範訳	按照权利要求【16】记载的【图像处理方法】，其特征在于，【使用热敏头在上述可逆显示部形成图像】。

主文パターン文	@ことを特徴とする請求項@又は@に記載の@。
実例	【ヒアルロン酸塩をさらに含有する】ことを特徴とする請求項【1】又は【2】に記載の化粧品。
模範訳	如权利要求【1】或【2】所述的化妆品，其特征在于，【还含有透明质酸盐】。

³⁹ 数字部分を含めて辞書に登録する場合、数字の部分だけ変えて、複数個を辞書に登録する必要がある。一見、非効率的に思えるが、数字を二つ以上含むわけでなければ、組み合わせ爆発で登録数が数百・数千に膨れ上がるわけではないため、精度向上に有効であれば数字を含めて辞書に登録することができる。

(5)節パターンについての分析

特許請求の範囲では、**名詞節**、**連用中止節**、**連体修飾節**及び**副詞節**が、高頻度で抽出された。

(a)**名詞節**は、発明の名称及び要約同様、**動詞**「有する」を使用したものが飛びぬけて多く、以下、「構成する」、「検出する」、及び「設けられる」が続いた。

節パターン	@を構成する@
実例	【液晶ディスプレイ】を構成する【液晶セルの裏側】
模範訳	构成【液晶显示器】的【液晶单元的内侧】

(b)**連用中止節**は、「を有し」、「であり」、「を備え」、「を含み」、及び「を形成し」等の**動詞**／**形容詞**を使用したものが高頻度となった。

節パターン	前記@は@であり、
実例	前記【柄部材】は【合成樹脂】であり、
模範訳	上述【手柄部件】为【合成树脂材料】、

(c)**連体修飾節**は、ほとんどが「である、」を使用したものだった。

節パターン	@が@である、
実例	【弾性変形部】が【螺旋状】である、
模範訳	【弹性变形部分】是【螺旋状的】

(d)**副詞節**では「であって、」を使用した節が非常に多く、それ以外の**動詞**／**形容詞**を使用した節はほとんど抽出されなかった。

節パターン	請求項@に記載の@であって、
実例	請求項【1】に記載の【印刷制御装置】であって、
模範訳	如权利要求【1】所述的【印刷控制装置】、

特許請求の範囲について作成した定型文・定型パターン文・定型フレーズ・主文パターン・節パターンの結果の中から特に上位のものを抜粋して頻度順に以下に示す。

なお、特許請求の範囲については、定型フレーズは、技術分野ごとに異なる表現が多いため技術分野のセクションごとに示すことにする。

4.6.1.2. 定型文

日本語	中国語	頻度
【請求項4】	[权利要求4]	4
【請求項5】	[权利要求5]	3
【請求項3】	[权利要求3]	3
不揮発性半導体記憶装置。	非易失性半导体存储器。	2
【請求項7】	[权利要求7]	2
【請求項6】	[权利要求6]	2
【化2】	[式2]	2
【化1】	[式1]	2

4.6.1.3. 定型パターン文

日本語	中国語	頻度
<\$1>が<\$2>である請求項<\$3>記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,<\$1>为<\$2>。	52
前記<\$1>は、<\$2>であることを特徴とする請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	44
前記<\$1>が<\$2>であることを特徴とする請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	39
前記<\$1>は、<\$2>であることを特徴とする請求項<\$3>記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	34
<\$1>が<\$2>である請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,<\$1>为<\$2>。	34
前記<\$1>は<\$2>であることを特徴とする請求項<\$3>記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	32
前記<\$1>が<\$2>であることを特徴とする請求項<\$3>記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	29
<\$1>が<\$2>である、請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,<\$1>为<\$2>。	29
前記<\$1>は<\$2>であることを特徴とする請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,所述<\$1>为<\$2>。	27
<\$1>が<\$2>であることを特徴とする請求項<\$3>に記載の<\$4>。	如权利要求<\$3>所述的<\$4>,其特征在于,<\$1>为<\$2>。	26

4.6.1.4. 定型フレーズ

各セクション共通：

日本語	中国語	品詞	頻度
請求項 1 に記載の	权利要求 1 所述的	連体詞	8062
請求項 1 記載の	权利要求 1 所述的	連体詞	6901
いずれかに記載の	任一项所述的	連体詞	6401
いずれか 1 項に記載の	任一项所述的	連体詞	3473
請求項 2 に記載の	权利要求 2 所述的	連体詞	2487
請求項 2 記載の	权利要求 2 所述的	連体詞	1593
請求項 3 に記載の	权利要求 3 所述的	連体詞	1533
請求項 4 に記載の	权利要求 4 所述的	連体詞	1379
請求項 5 に記載の	权利要求 5 所述的	連体詞	1264
いずれか一項に記載の	任一项所述的	連体詞	1062

A セクション：

日本語	中国語	品詞	頻度
遊技機	游戏机	名詞	176
血糖値測定装置	血糖值测定装置	名詞	162
記憶部	存储部	名詞	160

B セクション：

日本語	中国語	品詞	頻度
記録ヘッド	记录头	名詞	363
インクジェット記録装置	噴墨记录装置	名詞	223
記録装置	记录装置	名詞	219

C セクション：

日本語	中国語	品詞	頻度
プレス成形	冲压成型	名詞	130
粘着剤層	粘合剂层	名詞	107
炭化水素基	烃基	名詞	94

D セクション：

日本語	中国語	品詞	頻度
回転ドラム	转鼓	名詞	98
縫製装置	缝制装置	名詞	74
ドラム式洗濯機	滚筒式洗衣机	名詞	67

E セクション：

日本語	中国語	品詞	頻度
回動	转动	名詞	23
回動可能	可转动	名詞	19
付勢する	增势	動詞	19

F セクション：

日本語	中国語	品詞	頻度
空気調和機	空调机	名詞	181
燃料噴射弁	燃料喷射阀	名詞	173
移動体	移动体	名詞	160

G セクション：

日本語	中国語	品詞	頻度
液晶表示装置	液晶表示装置	名詞	1403
情報処理装置	信息处理装置	名詞	1169
画像形成装置	图像形成装置	名詞	1147

H セクション：

日本語	中国語	品詞	頻度
絶縁膜	绝缘膜	名詞	1776
半導体装置の製造方法	半导体装置的制造方法	名詞	1542
ゲート電極	栅电极	名詞	1001

4.6.1.5. 主文パターン文

日本語	中国語	頻度
<\$1>ことを特徴とする<\$2>。	<\$2>, 其特征在于: <\$1>。	17762
<\$1>ことを特徴とする請求項<\$2>に記載の<\$3>。	如权利要求<\$2>所述的<\$3>, 其特征在于: <\$1>。	8111
<\$1>ことを特徴とする請求項<\$2>記載の<\$3>。	如权利要求<\$2>所述的<\$3>, 其特征在于: <\$1>。	6872
<\$1>ことを特徴とする請求項<\$2>又は<\$3>に記載の<\$4>。	如权利要求<\$2>或<\$3>所述的<\$4>, 其特征在于: <\$1>。	807
<\$1>ことを特徴とする請求項<\$2>のいずれか<\$3>に記載の<\$4>。	如权利要求<\$2>中任<\$3>所述的<\$4>, 其特征在于: <\$1>。	764
<\$1>ことを特徴とする<\$2>のいずれか<\$3>に記載の<\$4>。	<\$2>中任<\$3>所述的<\$4>, 其特征在于: <\$1>。	764
<\$1>ことを特徴とする<\$2>のいずれかに記載の<\$3>。	<\$2>中任一项所述的<\$3>, 其特征在于: <\$1>。	719
<\$1>ことを特徴とする請求項<\$2>のいずれかに記載の<\$3>。	如权利要求<\$2>中任一项所述的<\$3>, 其特征在于: <\$1>。	660
<\$1>ことを特徴とする請求項<\$2>または<\$3>に記載の<\$4>。	如权利要求<\$2>或<\$3>所述的<\$4>, 其特征在于: <\$1>。	614
<\$1>ことを特徴とする請求項<\$2>又は<\$3>記載の<\$4>。	如权利要求<\$2>或<\$3>所述的<\$4>, 其特征在于: <\$1>。	562

4.6.1.6. 節パターン

日本語	中国語	頻度
<\$1>を有する<\$2>	具有<\$1>的<\$2>	911
請求項<\$1>に記載の<\$2>であって、	为权利要求<\$1>中所述的<\$2>，	470
請求項<\$1>記載の<\$2>であって、	为权利要求<\$1>中所述的<\$2>，	252
<\$1>を構成する<\$2>	构成<\$1>的<\$2>	178
<\$1>を検出する<\$2>	检测<\$1>的<\$2>	158
<\$1>を有し、	具有<\$1>，	157
<\$1>に設けられた<\$2>	设置于<\$1>上的<\$2>	139
前記<\$1>は<\$2>であり、	所述<\$1>为<\$2>，	136
前記<\$1>を構成する<\$2>	构成所述<\$1>的<\$2>	136
受信した<\$1>	接收的<\$1>	127
	∴	
<\$1>を備え、	具备<\$1>，	54
	∴	
<\$1>を含み、	含有<\$1>	38
	∴	
<\$1>を形成し、	形成<\$1>，	33

4.7. 明細書

4.7.1. 定型化の分析及び結果

4.7.1.1. 分析

(1) 定型文についての分析

明細書については、要約の場合と同様に見出しに用いられる用語が高頻度を占める。具体的には「【発明が解決しようとする課題】」、「【発明の属する技術分野】」、「【図1】」及び「【図2】」等である。

しかし、以下のような、明細書固有の高頻度文も存在する。

- ・ 以下、本発明を詳細に説明する。
- ・ 例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。

これらの高頻度文は、翻訳メモリとして登録しておくが良い。
なお、これらの見出しや高頻度文は、技術分野を問わず共通で使用されるものであった。

(2) 定型パターン文についての分析

明細書に関しては、大きく分けて4種類のパターンに分類される。

(a) 図の説明型

定型パターン文	@を示す@である。
実例	【電気掃除機】を示す【側面図】である。
模範訳	是表示【电动吸尘器】的【侧视图】。

定型パターン文	@を説明する@である。
実例	【A社製減圧室の構造】を説明する【断面構造図】である。
模範訳	是说明【A公司制减压室结构】的【截面结构图】。

定型パターン文	図@は、@を示す@である。
実例	図【10】は、【装置内部の構成例】を示す【模式図】である。
模範訳	图【10】是表示【装置内部结构例子】的【示意图】。

(b) 【発明の属する技術分野】に頻出型

定型パターン文	本発明は、@に関する。
実例	本発明は、【椅子の背凭れ】に関する。
模範訳	本发明涉及一种【椅子靠背】。

定型パターン文	本発明は、@に関するものである。
実例	本発明は、【湿式集塵装置】に関するものである。
模範訳	本发明涉及一种【湿式集尘装置】。

※ パターン自体のバリエーションは少ないが、この2つのパターンが頻出であった。

(c) 順序や書面上の場所を表す表現を含む型

定型パターン文	次に、@について説明する。
実例	次に、【ゲーム装置10の動作】について説明する。
模範訳	下面，对【游戏装置10的动作】加以说明。

定型パターン文	以下、@について説明する。
実例	以下、【各部の具体的構成】について説明する。
模範訳	下面，对【各部分的具体结构】加以说明。

定型パターン文	なお、@については後述する。
実例	なお、【ヘッド114同士の位置関係】については後述する。
模範訳	另外，关于【头114之间的位置关系】在后面叙述。

(d) その他の型

定型パターン文	得られた@は@であった。
実例	得られた【合金粉末の酸素量】は【0.08質量%】であった。
模範訳	得到的【合金粉末的氧含量】为【0.11wt%】。

定型パターン文	また、@には、@が設けられている。
実例	また、【利用ユニット3a】には、【各種のセンサ】が設けられている。
模範訳	另外，【利用单元3a】中设有【各种传感器】。

定型パターン文	@は、@と、@とを備えている。
実例	【このシステム100】は、【コンピュータ200】と、【カラープリンタ300】とを備えている。
模範訳	【此系统100】包括【计算机200】和【彩色打印机300】。

定型パターン文	@は、@であることが好ましい。
実例	【硬化膜の膜厚】は、【0.1~20μm】であることが好ましい。
模範訳	【固化膜的厚度】优选为【0.1-20μm】。

(3) 定型フレーズについての分析

明細書に関しては、大きく分けて以下の4種類の表現が抽出された。

- (a) 図表に関する表現型
- (b) 複合名詞型
- (c) サ変動詞型
- (d) 連体詞型

(a) 図表に関する表現型

- 例1：構成図
- 例2：図1に示す
- 例3：表1に示す

例 2 や例 3 の**動詞**は、実際に使用される文脈を見ると、終止形ではなく連体形で**名詞**を修飾していて、定型的なフレーズとして典型的な訳が存在する。従って、以下のように登録すると良い。

■辞書登録例 1

見出し語	図 1 に示す
訳語	图 1 中所示的
品詞	連体詞

■辞書登録例 2

見出し語	表 1 に示す
訳語	表 1 中所示的
品詞	連体詞

また上記例に「～した」や「～された」といった**助動詞**を含めた形は、日本語でも大きな意味の差はなく、中国語に訳した場合に過去形や受身形で訳されるわけではないため、下記のように上記例と同じ訳を登録しておく、精度の向上につながる。

■辞書登録例 3

見出し語	図 1 に示した
訳語	图 1 中所示的
品詞	連体詞

■辞書登録例 4

見出し語	図 1 に示された
訳語	图 1 中所示的
品詞	連体詞

なお、この型のフレーズは、特定の技術分野でのみ用いられる表現ではなく、全ての技術分野で高頻度で用いられる表現であった。

(b)複合名詞型

- 例 1：液晶表示装置 (名詞+名詞)
- 例 2：絶縁膜 (名詞+名詞)
- 例 3：固体撮像装置 (名詞+名詞+名詞)

これらは発明の名称や要約や特許請求の範囲と共通で出現しているものも多い。また、この型のフレーズは、技術分野ごとに異なるものが高頻度で抽出された。

(c)サ変動詞型

- 例 1：当接する (形容詞+動詞)
- 例 2：回動する (名詞+名詞+動詞)
- 例 3：出射する (名詞+動詞)

これも、(b)複合名詞型と同様に、発明の名称、要約及び特許請求の範囲と共通で出現しているものが多い。

なお、この型では、「当接する」のように複数の技術分野で使用される語もあれば、特定の技術分野で使用される語もある。

(d)連体詞型

例 1：本発明に係る

例 2：本実施形態に係る

これらは、「係る」を動詞として翻訳するよりも「的」を使用して翻訳することが多いようなので、下記のように登録しておくこと、自然な翻訳となる。

■辞書登録例 1

見出し語	本発明に係る
訳語	本発明的
品詞	連体詞

■辞書登録例 2

見出し語	本実施形態に係る
訳語	本実施例的
品詞	連体詞

なお、この型のフレーズは、全ての技術分野で共通して用いられるフレーズが多かった。

(4)主文パターン文についての分析

主文パターン文が定型パターン文と違う点の1つに、**副詞**を変数に入れるかどうかがある。定型パターン文では、極力変数部分に当てはめる語を減らし、固定部分を多くして訳質の向上を図る事を目的として、**副詞**を変数に入れないようにした。一方、主文パターン文では、固定部分をできるだけ減らし、変数部分に当てはめる語を多くして、最低限の骨格の構造を破たんさせないことを目的とするため、**副詞**を変数に入れている。

このため、明細書の主文パターン文では、以下のように、**文頭副詞**が変数になっているものが非常に多く抽出された。

主文パターン文	@、@について説明する。
実例 1	【次に】、【このような構成の基板搬送装置 2 4 1 における基板 2 3 3 の支持台 2 3 8 への供給及び排出の動作】について説明する。
模範訳 1	【其次】，对【这样的构成的基板输送装置 241 中的基板 233 的向支撑台 238 的供给以及排出的动作】进行说明。
実例 2	【図 8 を参照して】、【図 7 の S 1 0 0 における制御前提条件処理】について説明する。
模範訳 2	【参见图 8】，对【图 7 中 S100 的控制前提条件处理】进行说明。
実例 3	【ここで】、【描画処理時における一連の動作】について説明する。
模範訳 3	【这里】，对【描画处理时的一连串的动作】进行说明。

主文パターン文	@、@は@であり、@は@である。
実例 1	【ここで】、【5 0 1】は【処理順序の管理に用いるタスク番号】であり、【5 0 2】は【アクセス情報テーブルへのポインタ】である。
模範訳 1	【在此】，【501】是【用来管理处理顺序的任务号】，【502】是【指向存取信息表的指针】。

実例 2	【なお】、【図 4 (a)】は【図 2 に係る a - a 断面図】であり、【図 4 (b)】は【同 b - b 断面図】である。
模範訳 2	【再者】、【图 4(a)】是【图 2 的 a-a 剖面图】、【图 4(b)】是【其 b-b 剖面图】。
実例 3	【これらのうち】、【ボタン F 8】は【初期の設定を行うための機能を立ち上げるためのボタン】であり、【ボタン F 9】は【建設機械の後方の視野を確保するために設けられる映像を映すためのボタン】である。
模範訳 3	【在这些按钮中】、【按钮 F8】是【用于启动进行初始设定功能的按钮】、【按钮 F9】是【用于放映为确保工程机械后方视野而设置的摄像的按钮】。

また、要約同様、変数部に**動詞**を含む、より複雑なパターンも抽出された。

主文パターン文	本発明は、@及び@に関する。
実例	本発明は、【有機物や細菌等の分解や殺菌機能を有する EL ファイバー】及び【かかる EL ファイバーを用いた光触媒反応容器】に関する。
模範訳	本发明涉及【对有机物、细菌等具有分解或杀菌作用的 EL 纤维】，以及【包含该 EL 纤维的光催化反应器】。

主文パターン文	@は、@を説明するための@である。
実例	【図 1 5】は、【図 1 4 に示すフローチャートにおけるブロックの画像位置調整の候補から削除処理（ステップ S 1 4 0 3）の詳細】を説明するための【フローチャート】である。
模範訳	【图 15】是用于说明【图 14 所示的流程图中的从图像位置调整的候选中删除块的处理（步骤 S1403）的详细】的【流程图】。

(5)節パターンについての分析

明細書では、**名詞節**、**連用中止節**及び**副詞節**が、高頻度で抽出された。**連体修飾節**については、抽出自体はされたが、他のものに比べて頻度が非常に低かった為、分析対象からは外した。

(a)**名詞節**は、他のセクション同様、**動詞**「有する」を使用したものが飛びぬけて多く、以下、「示した」、「構成する」、「設けられた」及び「後述する」が続いた。

節パターン	@に示した@
実例	【図 1 3 ~ 図 1 5】に示した【ヒステリシス曲線の形状】
模範訳	【图 13 ~ 图 15】中所示的【磁滞曲线的形状】

(b)**連用中止節**は、「であり、」を使用したものが飛びぬけて多く、それ以外には「を用い、」、「を有し、」、「となり、」及び「を示し、」等の**動詞**、**形容詞**を使用したものが高頻度となった。

節パターン	@を用い、
実例	【アルゴン等の不活性ガスに酸素を 1 0 % 程度混合したガス】を用い、
模範訳	使用【在氩等惰性气体中混合 10% 左右氧的气体】、

(c)副詞節では、「であれば」、「に代えて」、「を超えると」、「である場合、」及び「の場合には、」等の動詞、形容詞を使用したものが高頻度となった。

節パターン	@が@であれば、
実例	【被写体】が【低輝度】であれば、
模範訳	如果【被摄体】是【低亮度】，

明細書について作成した定型文・定型パターン文・定型フレーズ・主文パターン・節パターンの結果の中から特に上位のものを抜粋して頻度順に以下に示す。

なお、明細書については、定型フレーズは、技術分野ごとに異なる表現が多いため技術分野のセクションごとに示すことにする。

4.7.1.2. 定型文

日本語	中国語	頻度
符号の説明	符号的说明	5525
【発明が解決しようとする課題】	[发明内容]	3741
【発明の属する技術分野】	[技术领域]	3730
【発明の実施の形態】	[具体实施方式]	3727
【従来技術】	[背景技术]	3722
【課題を解決するための手段】	[解决课题的方案]	3673
【図1】	[图1]	3572
【図面の簡単な説明】	[附图说明]	3569
【発明の効果】	[发明效果]	3492
【図2】	[图2]	3442
⋮		
以下、本発明を詳細に説明する。	以下，详细地说明本发明。	100
例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。	例如，也可以从实施例所示的所有构成要素中删除几个构成要素。	99

4.7.1.3. 定型パターン文

日本語	中国語	頻度
<\$1>を示す<\$2>である。	是表示<\$1>的<\$2>。	3169
図<\$1>は、<\$2>を示す<\$3>である。	图<\$1>是表示<\$2>的<\$3>。	1125
次に、<\$1>について説明する。	下面，对<\$1>加以说明。	595
<\$1>を図<\$2>に示す。	<\$1>显示于图<\$2>中。	394
<\$1>を説明する<\$2>である。	是说明<\$1>的<\$2>。	392
図<\$1>は<\$2>を示す<\$3>である。	图<\$1>是表示<\$2>的<\$3>。	373
本発明の<\$1>を示す<\$2>である。	表示本发明<\$1>的<\$2>。	360
<\$1>を説明するための<\$2>である。	用于说明<\$1>的<\$2>。	307
図<\$1>に示す<\$2>である。	是图<\$1>所示的<\$2>。	267
図<\$1>に<\$2>を示す。	<\$1>示出了<\$2>。	263

4.7.1.4. 定型フレーズ

各セクション共通：

日本語	中国語	品詞	頻度
本発明	本发明	名詞	165668
実施形態	实施例	名詞	64982
実施の形態	实施例	名詞	25904
本実施形態	本实施例	名詞	20479
図1に示す	图1中所示的	連体詞	10817
本実施例	本实施例	名詞	10160
実施形態に係る	实施例的	連体詞	9127
符号の説明	符号的说明	名詞	8963
図2に示す	图2中所示的	連体詞	8186
当接する	接触	動詞	8142

Aセクション：

日本語	中国語	品詞	頻度
遊技者	游戏者	名詞	1448
被検体	受检体	名詞	753
遊技機	游戏机	名詞	723

Bセクション：

日本語	中国語	品詞	頻度
記録ヘッド	记录头	名詞	1566
上下方向	上下方向	名詞	993
幅方向	宽度方向	名詞	992

Cセクション：

日本語	中国語	品詞	頻度
メッキ液	电镀液	名詞	1139
配合量	配合量	名詞	875
粘着剤層	粘合剂层	名詞	813

Dセクション：

日本語	中国語	品詞	頻度
ドラム式洗濯機	滚筒式洗衣机	名詞	378
回転速度	旋转速度	名詞	360
回転ドラム	转鼓	名詞	349

Eセクション：

日本語	中国語	品詞	頻度
閉状態	闭状态	名詞	108
回動可能	可转动	名詞	104
係止する	锁定	動詞	75

F セクション：

日本語	中国語	品詞	頻度
連通する	连通	動詞	732
空気調和機	空调机	名詞	563
燃料噴射弁	燃料喷射阀	名詞	536

G セクション：

日本語	中国語	品詞	頻度
液晶表示装置	液晶表示装置	名詞	5951
画像形成装置	图像形成装置	名詞	3764
情報処理装置	信息处理装置	名詞	2260

H セクション：

日本語	中国語	品詞	頻度
絶縁膜	绝缘膜	名詞	4333
ゲート電極	栅电极	名詞	2888
半導体装置の製造方法	半导体装置的制造方法	名詞	2413

4.7.1.5. 主文パターン文

日本語	中国語	頻度
<\$1>である。	是<\$1>。	57348
<\$1>は、<\$2>である。	<\$1>是<\$2>。	29053
<\$1>を示す<\$2>である。	是表示<\$1>的<\$2>。	20021
<\$1>は<\$2>である。	<\$1>是<\$2>。	10235
<\$1>は、<\$2>を示す<\$3>である。	<\$1>是表示<\$2>的<\$3>	7419
<\$1>、<\$2>について説明する。	对<\$1>、<\$2>进行说明。	5218
<\$1>を示す<\$2>。	表示<\$1>的<\$2>。	4815
<\$1>、<\$2>は、<\$3>である。	<\$1>、<\$2>是<\$3>。	3439
本発明は、<\$1>に関する。	本发明涉及一种<\$1>。	2589
<\$1>は<\$2>を示す<\$3>である。	<\$1>是表示<\$2>的<\$3>。	2349

4.7.1.6. 節パターン

日本語	中国語	頻度
<\$1>を有する<\$2>	具有<\$1>的<\$2>	8387
<\$1>に示した<\$2>	<\$1>中所示的<\$2>	5261
<\$1>を構成する<\$2>	构成<\$1>的<\$2>	4691
<\$1>に設けられた<\$2>	设置于<\$1>上的<\$2>	3087
<\$1>は<\$2>であり、 後述する<\$1>	<\$1>为<\$2>， 下述的<\$1>	2639 2503
<\$1>に形成された<\$2>	形成于<\$1>上的<\$2>	2383
<\$1>となる<\$2>	作为<\$1>的<\$2>	2294
図示しない<\$1>	图中未示的<\$1>	2221
<\$1>に応じた<\$2>	响应<\$1>的<\$2>	2165
⋮		
<\$1>を有し、	具有<\$1>，	639
⋮		
<\$1>となり、	作为<\$1>	514
⋮		
<\$1>に代えて、	代替<\$1>，	396
⋮		
<\$1>が<\$2>を超えると、	如果<\$1>超过<\$2>，	385
⋮		
<\$1>が<\$2>である場合、	<\$1>为<\$2>时，	373
⋮		
<\$1>は<\$2>を示し、	<\$1>表示<\$2>，	350
⋮		
<\$1>が<\$2>の場合には、	<\$1>是<\$2>时，	343

第5章 技術分野別特性の分析

5.1. 調査の目的

技術分野別の特性を分析するために、化学(構造)式・塩基配列に着目して調査を行った。化学式・塩基配列が含まれた文を機械翻訳した場合、この範囲をひとくくりに捉えられず、分断して構文を解釈することにより誤訳になることがある。

化学式・塩基配列として抽出可能な範囲は、機械翻訳においては、「翻訳対象外」にして名詞句として扱う(即ち、数字や英字等をそのまま表示する)方がよいと考えられる。化学式・塩基配列がどのように記述されているかの分析を行い、その範囲の特定方法を検討する。

5.2. 数式・化学式の特定方法

本調査では、まず、元素記号を含む文字の抽出を行い、化学式に特徴的な記号や単語で該当箇所を検索する事によって、その先頭文字と終端文字を類推しようと試みたが、数式との区別が難しい為、この方策は断念した。そこで、化学式に限定せず、化学式と数式をひとくくりに捉えて翻訳する方針を採用し、数式・化学式を対象として調査を実施した。

まず、本文から漢字・平仮名・カタカナ以外の全ての文字の抽出を行い、その中から数式・化学式で利用されるであろう文字を目視で抽出した。

数式・化学式で利用される文字としたのは、次の 19 種の文字である。

表 5.2.-1

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
≠	≡	∠	√	∞	∩	∪	∫	∞	≤	≥	≤	≥	∈	Σ	×	÷	-	+

これらの文字から前後に日本語や区切文字が出現するまでの文字列を抽出し、数式・化学式として扱うための範囲指定を試みる。その際、抽出した数式・化学式の精度確認、そして抽出した数式・化学式に過不足がないか、目視で精度確認を行うために、前後 20 文字の KWIC⁴⁰ (Keyword In Context) を作成する。下記表 5.2.-2 に KWIC の例を示す。

表 5.2.-2

前 20 文字	数式・化学式	後 20 文字
向く F の圧力はほぼ釣り合うように設計している。この場合、式	$3 : D 2 m = ((D 2 o^{2} + D 2 i^{2}) / 2) / 2$	の関係を保つようそのパターンの寸法が設計されている。【00
の図において有効再生時間 T_e は次式 (1) で表すことができる。	$T_e = t_{x1} + t_{x2} + t_{x3} + t_{x4} \dots (1)$	有効再生時間 T_e とは、前述したとおり目標ベッド温度 T_x を超え
周面の曲率半径 ρ と、ラップ部の半径方向の間隔寸法 T とを用いて	$\alpha_{min} = 2 \times \{ \sqrt{(T / 2 \times h)} \} / \rho$	として設定したので、例えばエンドミル等の切削工具によりスクロ

⁴⁰ 文書からキーワード等の特定の表現を抽出する際に、キーワードだけではなく、キーワード前後の文字列も含めて抽出したデータ又はその形式を指す。

特徴的な記号や文字を手がかりに KWIC の分析を行って数式・化学式を抽出し、構成要素および、構成規則を作成していく。

(抽出例)

数式をキーとして文章を検索・抽出し、その前後の文字を合わせて抽出している。

また、隣合う突起のなす角度 α の下限値 α_{\min} は、突起の高さ h と、突起の形成部位における周面の曲率半径 ρ と、ラップ部の半径方向の間隔寸法 T とを用いて $\alpha_{\min} = 2 \times \{ \sqrt{(T / 2 \times h)} \} / \rho$ として設定したので、例えばエンドミル等の切削工具によりスクロールの母材を切削加工してラップ部を形成するときには、ラップ部の内周面と外周面とに沿った切削動作を行うだけで、母材のうち内周面と外周面との間に位置する底面部分を速やかに削取ることができる。

※上記例の太字部分が、数式として抽出した文字列である。

※上記例の四角で囲まれている部分が、数式として抽出した文字列の前後 20 文字である。

5.3. 数式・化学式・塩基配列の抽出

式特有の文字が含まれるものを数式、元素記号を含むものを化学式、塩基配列と推測されるものを塩基配列として、分けて各パターンのルールを作成した。

(1)数式・化学式・塩基配列の抽出ルール

数式・化学式と判断し、翻訳対象外の名詞句を構成する文字と判断してよいものを後述する表 5.3.-1(数式),表 5.3.-3(化学式)及び表 5.3.-5(塩基配列)に示す。本パターンのルールは、調査対象の文に対して、前節で述べた KWIC の分析を行い作成したものである。

同じ文字でも数式内と判定すべきケース、数式外と判定すべきケースがあり、KWIC リストを目視で確認しながら何度か抽出パターンを調整した。

例：区切文字

カンマは、通常は区切文字として扱う文字であるが、下記のような数式の場合には、カンマで区切ると数式が分割されてしまう。数式で利用される文字に隣接する場合には、数式内と判定した方がよい。

$$\text{例: } \alpha = \max\{R_{\min}, \min(WL1 - WW1/2, WL2 - WW2/2)\}$$

カンマで区切ってしまうと、

①" $\alpha = \max\{R_{\min}$ "、②" $\min(WL1 - WW1/2$ "、③" $WL2 - WW2/2)\}$ "
の3つに分割されてしまう。

(2) 数式・化学式・塩基配列の正規表現⁴¹

上記の方法によって数式・化学式・塩基配列の範囲を推定することが可能となる。

(a) 数式

文中の数式を構成する文字列の正規表現は下記の通りである。

表 5.3.-1 数式の正規表現

正規表現	$[\text{^\{Han\} \{Katakana\} \{Hiragana\} :, , : . ; \, \} \text{¥}]$ $[\text{^\{Han\} \{Katakana\} \{Hiragana\} . ; \, \}^*]$ $[\neq \equiv \angle \sqrt{\infty} \cap \cup \int \infty \leq \geq \in \Sigma \times \div - +]^+$ $[\text{^\{Han\} \{Katakana\} \{Hiragana\} :, , : . ; \, \}^*]$ $[\text{^\{Han\} \{Katakana\} \{Hiragana\} :, , : . ; \, \} \text{¥} (]$
------	--

上記の「数式の正規表現の説明」を下記表に示す。

表 5.3.-2 数式の正規表現の詳細

構成要素	正規表現	内容
開始文字	$[\text{^\{Han\} \{Katakana\} \{Hiragana\} :, , : . ; \, \} \text{¥}]$	以下のいずれでもない文字が 1 個
		$\text{^\{Han\}}$ 漢字
		$\text{^\{Katakana\}}$ カタカナ
		$\text{^\{Hiragana\}}$ 平仮名
		記号 $:, , : . ; \, \}$
中間文字	$[\text{^\{Han\} \{Katakana\} \{Hiragana\} . ; \, \}^*]$	以下のいずれでもない文字が 0 個以上
		$\text{^\{Han\}}$ 漢字
		$\text{^\{Katakana\}}$ カタカナ
		$\text{^\{Hiragana\}}$ 平仮名
		記号 $. ; \, \}$
	$[\neq \equiv \angle \sqrt{\infty} \cap \cup \int \infty \leq \geq \in \Sigma \times \div - +]^+$	以下のいずれかの文字が 1 個以上
		\neq 特有文字
		\equiv 特有文字
		\angle 特有文字
		$\sqrt{\quad}$ 特有文字
		∞ 特有文字
		\cap 特有文字
		\cup 特有文字
		\int 特有文字
		∞ 特有文字
\leq 特有文字		
\geq 特有文字		
\leq 特有文字		

⁴¹ 文字列のパターンを表現する表記法。例：数値は[0-9] で表す。

		≧	特有文字
		≡	特有文字
		Σ	特有文字
		×	特有文字
		÷	特有文字
		—	特有文字
		+	特有文字
[[^] ¥p{Han} ¥p{Katakana} ¥p{Hiragana}。;、]*	以下のいずれでもない文字が 0 個以上		
	¥p{Han}	漢字	
	¥p{Katakana}	カタカナ	
	¥p{Hiragana}	平仮名	
	記号	;, , :。;、	
終端文字	[[^] ¥p{Han} ¥p{Katakana} ¥p{Hiragana};, , :。;、【¥(]	以下のいずれでもない文字が 1 個	
		¥p{Han}	漢字
		¥p{Katakana}	カタカナ
		¥p{Hiragana}	平仮名
		記号	;, , :。;、【(

※漢字、カタカナ及び平仮名は、Unicode において、日本や中国、台湾、韓国等東アジア圏の各国で使用されている文字⁴²。

(b)化学式

数式で利用される文字を含まないが、数式同様にひとくくりにした方がよいものに化学式がある。下記のように化学式を検出し、例えば、変数“X”に置き換えることにより、誤訳の可能性を低減することができる。以下にその例を示す。下記の例では、SF₆, N₂, CF₄をそれぞれ変数 X, Y 及び Z に置換している。

化学式の置換の例

日本語	中国語
実験 1、2 の結果から、SF ₆ と N ₂ とを含み、且つ CF ₄ を含まない混合ガス	从实验一个或两个的结果和 SF ₆ N ₂ 以及包括と, 并且是不里面含有 CF ₄ 的混合煤气

↓置換

日本語	中国語
実験 1、2 の結果から、X と Y とを含み、且つ Z を含まない混合ガス	以及, 从实验一个或两个的结果, 包括 X 和 Y, 并且是不里面含有 Z 的混合煤气

⁴² コードポイントの詳細については、下記 URL を参照。
<http://www.unicode.org/Public/UNIDATA/Scripts.txt>

	(He Li Be Ne Na Mg Al Si Cl Ar Ca Sc Ti Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr Rb Sr Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te Xe Cs Ba La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu Hf Ta Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn Fr Ra Ac Th Pa Np Pu Am Cm Bk Cf Es Fm Md No Lw Rf Db Sg Bh Hs Mt Ds Rg Cp H B C N O F P S K V Y I U W)*	元素記号が 0 個以上
	(<sub> <sup>)	上付タグ、下付タグの開始
	[0-9a-zA-Z¥-¥+--+]+	数字、アルファベット、プラス記号、マイナス記号が 1 個以上
	(<¥/sub> <¥/sup>)	上付タグ、下付タグの終了
終端文字	(<sub> <sup> <¥/sub> <¥/sup> ¥/ ¥(¥) ¥[¥ ¥{ ¥ }[0-9] He Li Be Ne Na Mg Al Si Cl Ar Ca Sc Ti Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr Rb Sr Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te Xe Cs Ba La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu Hf Ta Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn Fr Ra Ac Th Pa Np Pu Am Cm Bk Cf Es Fm Md No Lw Rf Db Sg Bh Hs Mt Ds Rg Cp H B C N O F P S K V Y I U W ¥= ¥+ ¥- _ + ¥s · →)*	元素記号、上付タグ、下付タグ、開き括弧、プラス記号、マイナス記号等の記号及び数字が 0 個以上

(c)塩基配列

上記の方法によって数式・化学式・塩基配列の範囲を推定することが可能となる。文中の塩基配列を構成する文字列の正規表現は下記の通りである。

表 5.3.-5 塩基配列の正規表現

正規表現	[0-9 0-9 ¥-¥']*[ATCGA T C G] {6, } [ATCGA T C G 0-9 0-9 ¥-¥' ¥s]*
------	---

上記の「塩基配列の正規表現の説明」を下記表に示す。

表 5.3.-6 塩基配列の正規表現の説明

構成要素	正規表現	内容
開始文字	[0-90-9¥-¥'']*	数字、マイナス記号、ダブルクォーテーション記号が 0 個以上
開始文字 又は 中間文字	[ATCGATCG]{6,}	塩基記号 6 個以上
中間文字 及び 終端文字	[ATCGATCG0-90-9¥-¥' ' ¥s]*	塩基記号、数字、マイナス記号、ダブルクォーテーション記号又は空白

(3) 上述の正規表現を適用して抽出した例

上述した正規表現を適用して数式・化学式・塩基配列を抽出した例を示す。文中の数式・化学式・塩基配列を名詞句としてまとめ、機械翻訳の翻訳対象外にすることにより、機械翻訳精度の向上が期待できる。上述の正規表現を適用せずに翻訳した例を翻訳結果項の適用前に、上述の正規表現を適用し名詞句としてまとめ、機械翻訳の翻訳対象外として翻訳した例を翻訳結果項の適用後に示す。

表 5.3.-7 抽出例

原文	抽出	翻訳結果	
スポンジ白金中に残存する (NH ₄) ₂PtCl₆ s ub>上に乾燥した水素ガスを 通じると爆発する危険がある。	(NH ₄) ₂PtCl₆	適用前	有当在₂PtCl₆ 残留在全海绵白金(NH ₄) 上通干燥氢煤气的時候爆炸的危險。
		適用後	有 当 在 (NH ₄) ₂PtCl₆ 残留在全海绵白金上通干燥氢煤气的時候爆炸的危險。
母子相互作用値 = (-2) × 30 + (+2) × 20 + (+1) × 10 = -10 となる。	= (-2) × 30 + (+2) × 20 + (+1) × 10 = -10	適用前	和 10 母子相互作用价值 = (-2)*30+(+2)*20+(+1)*10=成为。
		適用後	变成母子相互作用价值 = (-2)*30+(+2)*20+(+1)*10=-10。

5.4. 特殊タグ・フォーマットの有無

機械翻訳にとって問題となる特殊タグ・フォーマットの有無に関して調査を行った。

5.4.1. 利用されている特殊タグ・フォーマット

(1)特殊タグ

日本国の調査対象文献データを調査し、テキスト内に含まれる注意すべき特殊タグ・フォーマットを抽出して目視による確認を行ったところ、数式・化学式に関連すると思われる下記の2種類の特殊タグが確認できた。

表 5.4.1.-1 利用されている特殊タグ

特殊タグの意味	特殊タグの表記
開始	<maths>
	<chemistry>
終了	</maths>
	</chemistry>

上記の特殊タグは、下記の例に示すように画像データの指定に利用されている。

例 1	<pre><maths num="1"> </maths></pre>
例 2	<pre><chemistry num="1"> </chemistry></pre>

上述した特殊タグで指定された画像には例えば、下図のような数式・化学式が記載されている。

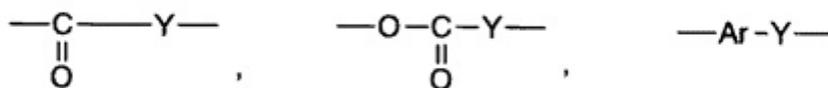


図 5.4.1.-1

(2)特殊フォーマット

KWIC⁴³を利用して分析を行ったところ、数式・化学式の右側に下記に太字で例示するような出現頻度の高いパターンが存在している事が分かった。

出現頻度の高いパターン例

例 1	$I y = a \langle \sup \rangle 2 \langle \sup \rangle c o s \langle \sup \rangle 2 \langle \sup \rangle (\Delta \gamma / 2)$	…(10)
例 2	$(1 - d i f x) \times (1 - d i f y) \times s (s x + 1, s y + 1)$	…式(11)
例 3	$\therefore (VM0 / I p) \times [(m-1) / n] \ll R Y$	…式(1)

上記太字部分のパターンを正規表現で示すと以下のようになる。

表 5.4.1.-2 高頻度の特殊フォーマット

[…] {3,} ¥s*式? ([¥ ([0-90-9A-Za-z]+ [¥])

上記の「高頻度の特殊フォーマットの正規表現の説明」を下記表に示す。

5.4.1.-3 高頻度の特殊フォーマットの正規表現の説明

構成要素	正規表現	内容
開始文字	[…] {3,}	中点または三点リーダーが 3 個以上
中間文字	¥s*	空白文字が 0 個以上
	式?	「式」という文字が 1 個以下
	[¥ ([全角または半角の開始括弧
	[0-90-9A-Za-z]+	数字、アルファベットが 1 個以上
終端文字	[¥]	全角または半角の閉じ括弧

⁴³ KWIC については「5.2. 数式・化学式の特定方法」を参照。

5.5. 数式・化学式・塩基配列の技術分野別の割合

前節「5.3. 数式・化学式・塩基配列の抽出」の手法により抽出した数式・化学式・塩基配列がどのような割合で出願しているか調査を行った。得られた数式、化学式、塩基配列の技術分野別の総抽出数、1 文献当たりの抽出数(各セクションの総抽出数/各セクションの総文献数)及び分布を示す表・グラフは下記図 5.5.-1~5.5.-3 の通りである。なお、下記のグラフでは 1 文献当たりの抽出数に基づいて技術分野別の分布を示している。

技術分野 (IPC)	抽出数	1 文献当たりの抽出数
A:生活必需品	933	1.41
B:処理操作;運輸	2,403	1.79
C:化学;冶金	5,922	5.11
D:繊維;紙	61	0.32
E:固定構造物	17	0.19
F:機械工学; 照明;加熱; 武器;爆破	874	1.18
G:物理学	11,233	4.22
H:電気	11,205	3.55
合計	32,648	3.26

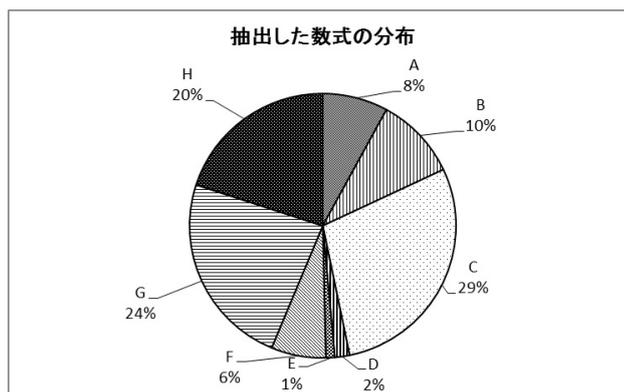


図 5.5.-1 数式の抽出割合

技術分野 (IPC)	抽出数	1 文献当たりの抽出数
A:生活必需品	527	0.80
B:処理操作;運輸	1,104	0.82
C:化学;冶金	11,340	9.78
D:繊維;紙	38	0.20
E:固定構造物	48	0.53
F:機械工学; 照明;加熱; 武器;爆破	428	0.58
G:物理学	4,567	1.72
H:電気	7,275	2.30
合計	25,327	2.53

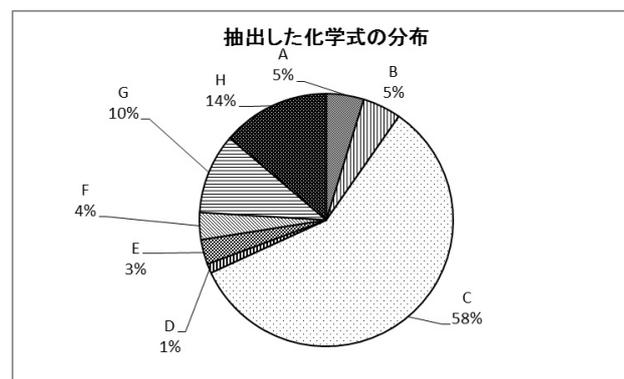


図 5.5.-2 化学式の抽出割合

技術分野 (IPC)	抽出数	1 文献当たりの抽出数
A:生活必需品	0	0
B:処理操作;運輸	0	0
C:化学;冶金	199	0.17
D:繊維;紙	0	0
E:固定構造物	0	0
F:機械工学; 照明;加熱; 武器;爆破	0	0
G:物理学	0	0
H:電気	0	0
合計	199	0.02

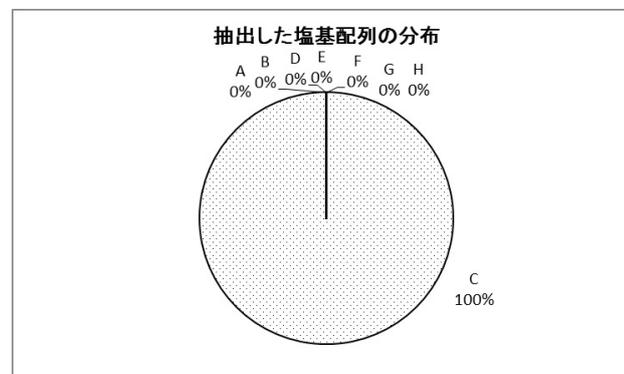


図 5.5.-3 塩基配列の抽出割合

5.6. 特殊タグ・フォーマットの出現頻度

前節「5.4. 特殊タグ・フォーマットの有無」で述べた特殊タグ(表 5.4.1.-1 利用されている特殊タグ)及び特殊フォーマット(表 5.4.1.-2 高頻度の特殊フォーマット)について技術分野別の出現数(抽出数)及び分布(総出現数に占める各技術分野の出現数の割合)を調査したところ、以下のような結果になった。なお、下記のグラフでは1文献当たりの出現数に基づいて技術分野別の分布を示している。

技術分野 (IPC)	出現数	1文献当たりの出現数
A:生活必需品	89	0.134
B:処理操作;運輸	55	0.041
C:化学;冶金	1,243	1.071
D:繊維;紙	4	0.021
E:固定構造物	0	0
F:機械工学; 照明;加熱; 武器;爆破	1	0.001
G:物理学	64	0.024
H:電気	222	0.07
合計	1,678	0.168

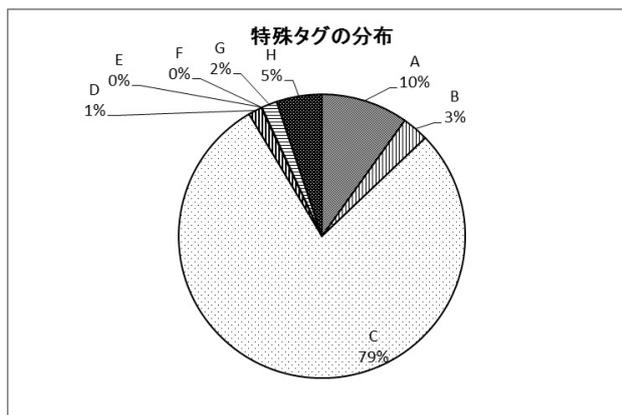


図 5.6.-1 特殊タグの分布

技術分野 (IPC)	出現数	1文献当たりの出現数
A:生活必需品	23	0.035
B:処理操作;運輸	70	0.052
C:化学;冶金	212	0.183
D:繊維;紙	0	0
E:固定構造物	8	0.089
F:機械工学; 照明;加熱; 武器;爆破	78	0.105
G:物理学	469	0.176
H:電気	454	0.144
合計	1,314	0.131

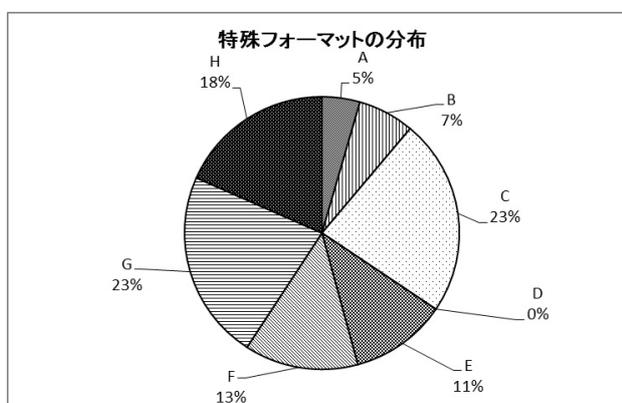


図 5.6.-2 特殊フォーマットの分布

5.7. 技術分野別特性の考察

前節 5.1.~5.6.までの分析結果から次の傾向が読み取る事ができた。

- ・数式は、技術分野 C,G 及び H セクションでの出現割合が高い。
- ・化学式では、技術分野 C セクションの割合が圧倒的に高く、次いで技術分野 H,G セクションとなる。
- ・塩基配列は、技術分野 C セクション以外では確認できなかった。
- ・数式・化学式は分野横断的に出現しているため、上述した数式・化学式の抽出手法を分野に関係なく利用可能であると言える。

第6章 調査結果の検証

6.1. 調査目的

本章では、前章までに得られた特許文献の日中機械翻訳に対する改善策を実施した場合の改善効果の検証を目的とする。検証した改善策は、「3.4.2. ミクロ分析」の「3.4.3. 各不備要因に対する改善策」で得られた翻訳不備に対する改善策と、「第4章 定型化可能な表現の分析」の分析結果から得られた定型化による改善策である。

6.2. 改善効果の検証方法

改善効果の検証は、「3.4.2. ミクロ分析」で調査対象としたミクロ分析用の文献データを再度、調査対象とし、改善策を実施する前後での機械翻訳結果の比較分析をする事により行う。改善策を実施する前の機械翻訳結果は、「3.4.2. ミクロ分析」で既已取得している為、ここでは、機械翻訳システムに改善策を実施し、改善策が実施された後の機械翻訳結果を取得する。

6.2.1. 改善策の実装

本調査で改善策として機械翻訳システムに実装したのは、以下の4つである。

(1)	改善用の不備対策辞書データ
(2)	改善用の不備対策翻訳メモリ
(3)	改善用の定型フレーズ辞書データ
(4)	改善用の定型文翻訳メモリ

以下、個別に説明する。

(1)改善用の不備対策辞書データ

「3.4.3. 各不備要因に対する改善策」では、ミクロ分析用の文献データのミクロ分析から改善策を検討した。この改善策を検討する際に得られたミクロ分析用の文献データに対する改善用の不備対策辞書データを機械翻訳システムの辞書に登録した。

下記に改善用の不備対策辞書データの品詞の内訳を示す。

品詞	語数
名詞	1,575
動詞	566
副詞	58
形容詞	52

表 6.2.1.-1 改善用の不備対策辞書データの品詞

(2)改善用の不備対策翻訳メモリ

上記(1)の改善用の不備対策辞書データと同様に「3.4.3. 各不備要因に対する改善策」でマイクロ分析から改善策を検討した際に得られた知見からマイクロ分析用の文献データに対する改善用の不備対策翻訳メモリを登録した。

なお、この不備対策翻訳メモリとして、後述の(4) 改善用の定型文翻訳メモリを利用する場合もある。

(3)改善用の定型フレーズ辞書データ

「第 4 章 定型化可能な表現の分析」では、定型化可能な表現を調査し、実際に定型文、定型パターン文、定型フレーズ、主文パターン文及び節パターンを取得する事ができた。改善用の定型フレーズ辞書データとは、辞書に登録する事によって改善策を実施する事が可能なデータを指し、上述した 5 つのパターンのうち、定型フレーズが該当する。改善用の辞書データと同様に、改善用の定型フレーズ辞書データを機械翻訳システムの辞書に登録した。

なお、改善用の定型フレーズ辞書データに登録された定型フレーズの詳細に関しては「第 4 章 定型化可能な表現の分析」を参照されたい。

(4)改善用の定型文翻訳メモリ

「第 4 章 定型化可能な表現の分析」で得られた定型可能な表現のうち、定型文と定型パターン文については翻訳メモリとして登録した。これら 3 つを改善用の定型文翻訳メモリと呼ぶ。

なお、改善用の定型文翻訳メモリに登録された定型文と定型パターン文の詳細に関しては「第 4 章 定型化可能な表現の分析」を参照されたい。

6.2.2. 改善策を実施した後の分析結果の差異

前節「6.2.1. 改善策の実装」で説明した改善策を機械翻訳システムに実装し、マイクロ分析用の文献データの機械翻訳結果を再取得し、改善策によって訳が変化したそれぞれの箇所について、改善前と改善後の訳を改善 1 点、同等 0 点、悪化-1 点として評価した。その結果、各得点の分布は以下の表 6.2.2. のようになった。

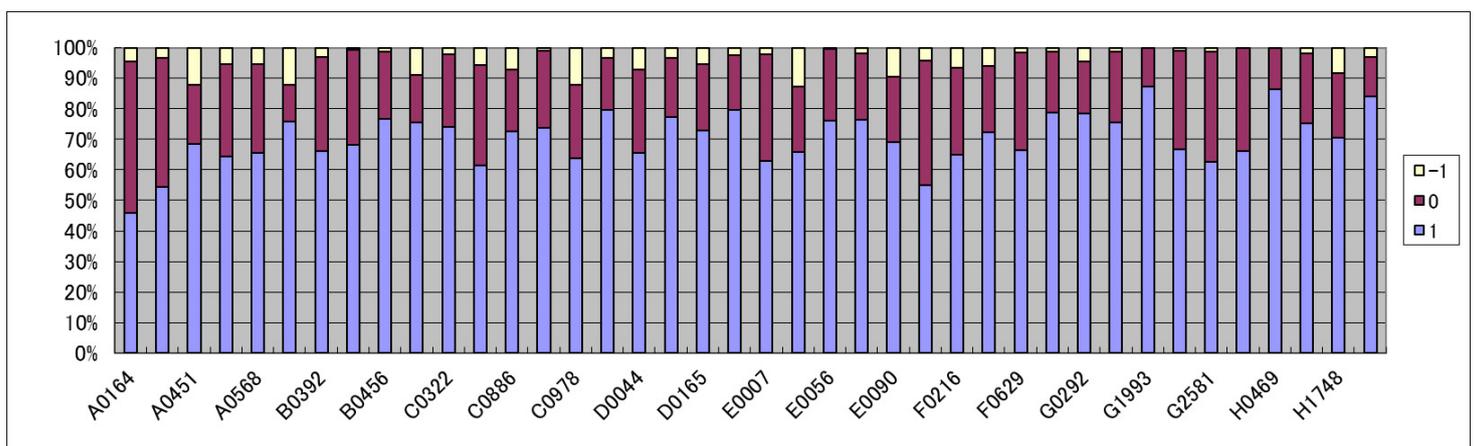


表 6.2.2. 改善効果の評価点分布図

上記の表 6.2.2.-1 を見ると、全体的に「改善」(1点)が 50%~80%と最も多く、「同等」(0点)が 20%前後であった。この結果から改善策が非常に有効である事が分かる。

しかしながら、わずかだけ「低下」(-1点)も見られた。この「低下」の原因は、後述する「第 8 章 課題と対策」で考察する。

6.2.3. 改善効果の分析

以下に、「3.2. 翻訳不備の要因分析」で挙げた改善策のうち辞書によるものについての改善効果を具体的に検証する。

6.2.3.1. 機械翻訳の訳語の不備に対する改善効果

(1)一般的な訳語の他に特許文にふさわしい訳語がある場合

「3.2.1. 機械翻訳の訳語の不備に基づく問題点」の「(1)一般的な訳語の他に特許文にふさわしい訳語がある場合」で例示した文献データ A0164 に関係した改善策は以下の 2 つである。

- (a)改善用の不備対策辞書データを、動詞として「1 が 2 に 関する = 1 涉及 2」のように登録した。
- (b)「3.2.6.その他の問題点」の「(2)原文にはないが、訳文には補った方がよい場合」に対する改善策としても以下の改善用の不備対策翻訳メモリを登録した。

改善用の不備対策翻訳メモリの例

原文パターン	本発明は、<\$1>に関する。
訳文パターン	本發明涉及一種<\$1>。

改善策(b)の結果、改善用の不備対策翻訳メモリにマッチし、下記に示すように改善効果が得られた。

改善例 1) 文献データ A0164

原文	【本発明は、】 魚を釣り上げた後に、この魚を収納しておくバッグ 【に関する】 。
機械翻訳文(改善策前)	钓鱼, 并且放了之后 【本發明关于】 收藏这条鱼的手提包。
機械翻訳文(改善策後)	【本發明涉及一種】 在鱼钓到了之后在这条鱼收容的手提包。
対応する中国特許文	【本發明涉及一種】 在钓到鱼之后可以存放钓到的鱼的鱼护。

改善後は特許文献のごく定型的な表現である「本發明涉及一種…」と訳出することが可能になった。

また、上記の改善策(a)も、以下に挙げるように別の文献データ D0165 の改善につながった。

改善例 2) 文献データ D0165

原文	この発明は、水噴射式織機の機上脱水装置に【 関する 】。
機械翻訳文 (改善策前)	这项发明在水喷射式织机的时机上【 关于 】脱水装置。
機械翻訳文 (改善策後)	这个发明【 涉及 】脱水喷水式纺织机的台式装置。
対応する 中国特許文	本发明【 涉及 】一种喷水式织机的机上脱水装置。

こちら、「～に関する」の訳語が、「关于」から、より特許文献で一般的な「涉及」に変化した。

上記改善例 1)と改善例 2)との違いは、原文の主語が「本発明は、」から「この発明は、」に変わったことである。このように、全てのバリエーションに翻訳メモリがカバーしきれていない場合でも、上記のような辞書登録があることによって、ある程度特許文らしい訳を出すことができた。

(2)専門的な訳語がある場合

「3.2.1. 機械翻訳の訳語の不備に基づく問題点」の「(2)専門的な訳語がある場合」で例示した文献データ B0392 に関係した改善策を以下に示す。

(c)改善用の不備対策辞書データを、名詞として「ハンガー＝悬挂件」のように登録した。

改善策(c)の結果、改善用の不備対策辞書データにより、下記に示すように改善効果が得られた。

改善例 3) 文献データ B0392

原文	2個のクランプ31を備える【 ハンガー 】30は、案内装置32により上下方向に位置決め自在である。
機械翻訳文 (改善策前)	至于拥有2个的防滑钉片31的【 衣架 】30,定位被向导装置32在上下方向自在。
機械翻訳文 (改善策後)	至于拥有2个的夹钳31的【 悬挂件 】30,定位在垂直地被引导装置32自由。
対応する 中国特許文	另外,在【 平板 】16上具有插通针杆1用的开口部17。

改善後は辞書に登録したとおりに訳されるようになった。

(3)訳し分けの問題がある場合

「3.2.1. 機械翻訳の訳語の不備に基づく問題点」の「(3)訳し分けの問題がある場合」で例示した文献データ A0164 に関係した改善策を以下に示す。

(d)改善用の不備対策辞書データを、格支配情報⁴⁴を添えて動詞として「…に = 到…内」のように登録した。

⁴⁴ 格支配情報とは、動詞がどのような格助詞（が、を、に、へ、から等）から係り受けすることができるかを示す情報である。

改善策(d)の結果、改善用の不備対策辞書データにより、下記に示すように改善効果が得られた。

改善例 4) 文献データ A0164

原文	そして、釣人は魚を釣り上げるとバッグの袋部内【に】魚を【収納する】。
機械翻訳文 (改善策前)	以及当钓鱼人提高鱼钓鱼的时候【在】手提包的袋子部【里收藏】鱼。
機械翻訳文 (改善策後)	以及当钓鱼者在鱼钓【到】的时候将鱼手提包的袋子部【内存放】。
対応する 中国特許文	钓鱼者钓到鱼后，将鱼放【到】鱼护的袋部分【内存放】。

「3.2.1. 機械翻訳の訳語の不備に基づく問題点 (3)訳し分けの問題がある場合」で述べたように、改善前は「に」が誤って「在」と訳されていたが、正しく「到」と訳されるようになった。

6.2.3.2. 訳の位置に関する問題点に対する改善効果

「3.2.3. 訳の位置に関する問題点」で例示した文献データ H2965 に関係した改善策を以下に示す。

(e)改善用の不備対策辞書データを、副詞として「本発明によれば = 根据本发明」のように登録した。

改善策(e)の結果、改善用の不備対策辞書データにより、下記に示すように改善効果が得られた。

改善例 5) 文献データ H2965

原文	【本発明によれば】、配線基板の孔に透光性蓋部の少なくとも一部が嵌入されているので、従来に比較して薄い固体撮像装置を得ることが出来る。
機械翻訳文 (改善策前)	因为至少透光性盖子部的1部被【据本发明说】电线敷设基础的洞嵌入さ所以在从来比较,能得到薄固体撮像装置。
機械翻訳文 (改善策後)	【根据本发明】在以往比较,因为至少透光性盖子部的1部被印刷电路板的洞嵌入さ所以能获得薄固态图像传感装置。
対応する 中国特許文	【根据本发明】的固态图像传感装置的特征在于透光盖部分的至少一部分安装于印刷电路板的孔中,可以得到与常规固态图像传感装置相比更薄的固态图像传感装置。

改善前は、文中に訳出され、訳語も特許文献で一般的な表現ではなかったが、改善後は位置、訳語いずれも特許文献でよく見られるものに改善された。

6.2.3.3. 訳抜けに関する問題点に対する改善効果

「3.2.5. 訳抜けに関する問題点」で例示した文献データ E0059 に関係した改善策を以下に示す。

(f)改善用の不備対策辞書データを、連体助詞として「…等の = …等」のように登録した。

改善策(f)の結果、改善用の不備対策辞書データにより、下記に示すように改善効果が得られた。

改善例 6) 文献データ E0059

原文	屋外配電盤収納ボックス【等の】固定枠体側から扉を開閉するために、下記特許文献に見られる、図5と図6で示すL型のハンドルが扉側に取り付けられている。
機械翻訳文 (改善策前)	由于图5和为开关门在下列专利文献能够被看见的图6显示型L车把被在门一侧屋外配电网盘收藏箱【】的固定范围身体一侧安装。
機械翻訳文 (改善策後)	型L手柄示出由于图5和为从屋外配电网盘收藏箱【等】的固定框体一侧开关门在下述专利文献能够被看见的图6被门一侧安装。
対応する 中国特許文	为了从屋外配电网盘收藏箱【等】的固定框体侧对门进行开闭，有一种在门侧安装有下述专利文献中记载的、如图5和图6所示的L形手柄。

改善前は訳出されていなかった「等」が訳出されるようになった。

6.2.3.4. その他の問題点に対する改善効果

「3.2.6. その他の問題点」の「(3)単語の切れ目で誤りが発生する場合」で例示した文献データ A0164 に関係した改善策を以下に示す。

(g)改善用の不備対策辞書データを、動詞として「複数形成する = 成有多個」のように登録した。

改善策(g)の結果、改善用の不備対策辞書データにより、下記に示すように改善効果が得られた。

改善例 7) 文献データ A0164

原文	発明3のバッグは、発明1のバッグであって、袋部の上面及び下面の開口側縁にはロープ穴が【複数形成され】、ロープがロープ穴を挿通することで開口が開閉自在に綴じられており、閉塞具をロープが兼ねている。
機械翻訳文 (改善策前)	发明3的手提包是发明1的手提包,被绳索洞孔袋子部的外表以及开口一侧预先方面的缘【形成复数形式】,并且因为绳索在绳索洞孔插通所以被把开闭自在地开口订起来,并且绳索正兼任闭塞工具。
機械翻訳文 (改善策後)	发明3的手提包是发明1的手提包,绳索孔被【成有多个】在袋子部的上表面以及下表面的开口边缘,并且打开和闭合自由,并且因为绳索穿过绳索孔所以开口被订起来,并且绳索正兼任封闭构件。
対応する 中国特許文	发明3中的鱼护是发明1中的鱼护,其特征在于: 上述袋部分的上表面及下表面的开口边缘处形【成有多个】绳孔,上述绳子穿过上述绳孔,将上述开口可自由开闭地束起,其中,所述绳子兼用作上述封闭构件。

改善前は「成される = 形成」と「複数形 = 复数形式」のようにバラバラに訳されていたのが、改善後はまとめて「成有多個」と訳されるようになった。

第7章 平成 21 年度調査結果との比較・分析

7.1. 調査の目的

平成 21 年度調査では、日中機械翻訳に関する本調査とは逆に中国公開特許公報から日本語への中日機械翻訳に関する調査が行われた。平成 21 年度調査では、あらかじめ機械翻訳をした結果をデータベースに蓄積して検索を行う方式（コンテンツ方式）等による翻訳精度向上のための分析を行った。ここでは、日中機械翻訳に関する分析結果及び考察を、平成 21 年度調査の調査結果と比較対照する事により、いくつかの課題を検討する。

7.1.1. 言語の文法的特性・言語の特性について

平成 21 年度調査では以下の文法的特性について述べられている。ここでは、これらの文法的特性が、NewJC の翻訳結果に反映されているかどうかについて述べる。

1. 形態素の曖昧性
文が漢字で区切りなく書かれるため、単語の境界が明確でないという特性。
2. 多品詞語
ある単語が文脈に応じて例えば動詞と名詞のいずれかに使用されるといった特性。
3. 分離前置詞
例えば「～の上に」を中国語では「在～上」と言うように、2 語に分かれた語がひとつの前置詞に相当するという特性。
4. 離合詞
例えば「見他的面」（彼の顔を見る→彼に会う）というような表現で、動詞「見面」（会う）の間に目的語「他的」（彼の）が挿入されるという特性。
5. 重ね型表現
例えば「散散步」（ちょっと散歩する）というような表現で、「散歩」（散歩する）の「散」を重ねることで、「ちょっと」の意味を付与するという特性。

まず、1 と 2 は中国語の解析の問題であり、日中翻訳エンジンは中国語の解析は行わないので、無関係である。また、4 と 5 は特許表現にはあまり使用されないということなので、分析は省略する。

残るは分離前置詞であるが、NewJC で以下の原文を翻訳したところ、下記の翻訳結果を得た。

原文	こうして得られた第 1 の分散液を <u>基体 1 0 0 上</u> に塗布する。
模範訳	将这样得到的第 1 分散液涂布在 <u>基体 100 上</u> 。
機械翻訳文	<u>在基础身体 100 上</u> 涂这样能够得到了的第 1 的分散液。

ここで、「基体 1 0 0 上に」は「在基础身体 100 上」と訳されている。これは「～上に」の部分に関しては「在～上」という分離前置詞を用いて正しく翻訳されていることを示している。

以下にもうひとつ例を挙げる。

原文	これは、 <u>被施療者が足先を入れたときに</u> 、足先の安定性を高めるためである。
模範訳	这是因为 <u>在被治疗者放入脚时候</u> ，能够提高脚的稳定性。
機械翻訳文	这个 <u>在被施療者把脚的前方装进去了的时候</u> ，是为提高脚的前方的稳定。

ここで、「被施療者が足先を入れたときに」は、「在被施療者把脚的前方装进去了的时候」と訳されている。これは、「～ときに」の部分に関しては「在～时候」という分離前置詞を用いて正しく翻訳されていることを示している。

このように、NewJC には分離前置詞を利用できる機能が備わっているため、平成 21 年度の中日調査で挙げた分離前置詞の中で訳出されないものがあったとしても、辞書に登録すれば利用できるようになる。

7.1.2. 定型化可能な表現について

本調査の「第 4 章 定型化可能な表現の分析」では、特許文献を日中機械翻訳する際の原文となる日本語の定型化可能な表現について分析した。同様に、平成 21 年度調査では、特許文献を中日機械翻訳の際に原文となる中国語の定型化可能な表現について分析がされている。これらの分析結果を比較・考察する。

(1) 発明の名称

発明の名称の定型化可能な表現は、日本語及び中国語の場合のいずれも、通常は名詞句で構成されているという特徴が判明した。

(2) 要約

要約については、中国語の定型化可能な表現では、「本发明提供～」(本発明は～を提供する)のように「本发明」(本発明)を含む定型文が多かった。一方、日本語の定型化可能な表現では、「～を提供する。」のように「本発明」がない場合が多かった。これは、日本語の特許文献の要約は、【課題】の見出しの直後に記載される形式になっている事に起因すると思われる。

(3) 特許請求の範囲

特許請求の範囲については、中国語の定型化可能な表現では、「本发明的特征在于：」(本発明の特徴は以下の通り)のような定型文が多いのに対して、日本語の場合にはこのような表現の定型文は皆無であった。これは、中国語の請求項の記載では、文をコロンとセミコロンで区切る傾向があり、文頭からコロンまでの文が定型文として高頻度になった結果であると考えられる。

従って、中国語の定型化可能な表現では、コロンを使用したものが多く存在するという点で日本語の定型化可能な表現と大きく相違していた。

<中国語の定型化可能な表現>及びその<日本語対訳>と、<日本語の定型化可能な表現>及びその<中国語対訳>とのペア同士を比較したところ、<中国語の定型化可能な表現>でコロンが使用されていない場合には同様の結果が抽出されていた。一方、<中国語の定型化可能な表現>でコロンが使用されている場合にはこれらのペアの合致が見られなかった。例えば、以下の例は、<中国語の定型化可能な表現>及びその<日本語対訳>として抽出されたものである。

例) <中国語の定型化可能な表現>及びその<日本語対訳>

原文	一种<\$1>, 其特征在于:
訳文	以下を特徴とする<\$1>:

上記の例に該当する<日本語の定型化可能な表現>及びその<中国語対訳>は本調査では得られなかった。これは、上記の例の訳文のような表現は実際の日本語の特許文献では用いられないが、中国語から日本語に翻訳する際には、このような表現にしないとうまく翻訳できない為である。

(4)明細書

明細書については、日本語の定型化可能な表現と中国語の定型化可能な表現との間で以下の共通点が見られた。

- (a) 「4.7.1.1. 分析」の(1)で説明した見出し系の高頻度文が多い。
- (b) 「4.7.1.1. 分析」の(2)(a)で説明した「図の説明型」の定型パターン文が多い。
- (c) 「4.7.1.1. 分析」の(2)(b)で説明した「【発明の属する技術分野】に頻出型」の定型パターン文が多い。
- (d) 「4.7.1.1. 分析」の(2)(c)で説明した「順序や書面上の場所を表す表現を含む型」の定型パターン文が多い。

7.1.3. 外来語表記について

本節では日本語と中国語の外来語の表記方法を比較し、平成 21 年度調査における中国語の外来語表記および未知語の分析結果を日中機械翻訳に応用する方法に関して検討する。

日本語では外来語は多くの場合カタカナで表記する。一方、中国語では外来語は「意識」又は「音訳」という形で漢字表記にする（平成 21 年度調査の報告書の「3.3.1.2. 外来語の中国語表現と表記のゆれ」を参照）。日本語で外来語の表記に使う文字であるカタカナは日本固有語にはほとんど使われない文字であり、外来語が未知語となった場合でもカタカナの並びを単語とすればよいのでその範囲の特定は容易である。一方、中国語では外来語を中国固有語と同様に漢字で表記するため、未知語となった場合その範囲の特定が困難である（平成 21 年度調査の報告書の「3.3.2. 未知語の分析」を参照）。

平成 21 年度調査では漢字表記される外来語が未知語となった場合にその範囲を特定するために外来語の音訳で使用頻度の高い漢字を収集し、音訳文字リスト（平成 21 年度調査の報告書の表 3.3.2.5-1 を参照）を作成した。このリストの各漢字にカタカナ表記をあてることで、日本語でカタカナ未知語が発生したときにその音訳を自動生成することができるものと思われる。

ここでは例を用いて想定される処理の概略を示す。例として「ボルツマン」を用いる。ボルツマンは人名であるが「ボルツマン定数」「ボルツマン方程式」などの形で技術文献にも出現しうる。ボルツマンの漢字表記としては「玻尔兹曼」「玻耳兹曼」「波尔兹曼」などが多くみられる。このように音訳では何種類かの表記方法が混在していることがある（平成 21 年度調査の報告書の「3.3.1.2. 外来語の中国語表現と表記のゆれ」の「(5)表記のゆれ」を参照）。

「ボルツマン」について表の上位から発音の近い漢字を拾っていくと以下の様になる⁴⁵。

⁴⁵ 平成 21 年度調査の報告書の表 3.3.2.5-1 には 30 位までの記載しかないが、同様の手法で調査して、49 位の「ツ」=「茨」を得た。

カナ読み	漢字 (ピンイン)	出現順位
ボ	bó 伯	24
ル	ěr 尔	1
ツ	cǐ 茨	49
マン	màn 曼	23

表 7.1.3.-1 ボルツマンの音訳

これにより「伯尔茨曼」が得られる。これは「玻尔兹曼」「玻耳兹曼」「波尔兹曼」とは異なるが、ボルツマン("Boltzmann")の音訳として「伯尔茨曼」を示しているサイト⁴⁶もあるようにありえない表記ではなく、該当技術を熟知している者であれば類推可能であると思われる。

実用的な処理のためにはどのようなカタカナ表記でも処理できる網羅的な詳細な漢字リストの整備が必要であり、今後の課題である。

7.1.4. 化学式(構造式)・塩基配列の表記法について

本調査の「第5章 技術分野別特性の分析」では、特許文献を日中機械翻訳する際の原文となる日本語に含まれる化学式・塩基配列について分析した。同様に、平成21年度調査では、特許文献を中日機械翻訳する際に原文となる中国語に含まれる化学式・塩基配列について分析がされている。

両者を比較した結果、2つの特徴が判明した。

(1)中国語の特許文献に比べ、日本語の特許文献では文中に出てくる数式・化学式が少ない。

これは、中国語の特許文献では MathML⁴⁷を使用して文中に数式を書くことが多いのに対し、日本語の特許文献では化学式をイメージ画像等の挿入図として別途参照することが多いと考えられる。

(2)日本語及び中国語の特許文献中の数式・化学式・塩基配列の抽出の為の正規表現が非常に類似していた。

なお、類似はしているものの、本調査では日本語に対応する必要があった為、平仮名、カタカナ、全角数字及び全角アルファベット等を考慮に入れた正規表現とした。

7.1.5. 句読点の表記法について

平成21年度調査では、特許文献を中日機械翻訳する際の原文となる中国語の句読点の表記法について分析がされている。その分析結果によると、中国語の特許文献では、コロン及びセミコロンが多用されている事が判明している。一方、日本語の一般的な文章では、

⁴⁶ Wikipedia 中国語版(<http://zh.wikipedia.org/zh-cn>)「德语姓名列表 (B)」

⁴⁷ Mathematical Markup Language の略で、数式や化学式を記述するための言語。詳細は平成21年度調査「3.2.2.1. MathMLによる記述」を参照。

通常、コロン及びセミコロンはほとんど使用されない。日本語の特許文献も同様で、文の区切りにコロン及びセミコロンはほとんど使用されていない。

中国語の特許文献で出現頻度の高い、コロン及びセミコロンの後で改行している請求項の記載例を、対応する日本語の特許文献と共に以下に挙げる。下記例では、改行位置を明示する為に
と表記する。

例 1) コロン及びセミコロンの後に改行が有る例

中国語 特許文献	<p>一种空调服装，其特征在于，具有：
 衣料部，其用于覆盖身体的至少上半身及臂部，且在与身体或内衣之间的空间内，沿身体或内衣的表面引导空气；
 一个或多个送风单元，其用于在前述衣料部与身体或内衣之间的空间，强制性地产生空气流；
 电源单元，其向前述送风单元提供电力，其中，
 在前述衣料部的背面侧且在与臂部的一部分或全部对应的部分，作为衬里敷设有高吸湿性布料。
</p>
日本語 特許文献	<p>身体の少なくとも上半身及び腕部を覆うと共に、身体又は下着との間の空間において空気を身体又は下着の表面に沿って案内するための服地部と、
 前記服地部と身体又は下着との間の空間に空気の流れを強制的に生じさせるための一又は複数の送風手段と、
 前記送風手段に電力を供給する電源手段と、
 を具備し、
 前記服地部の裏面側であって腕部の一部又は全部に対応する部分に、吸湿性の高い布地を裏地として取り付けたことを特徴とする空調衣服。</p>

上記例の対応する日本語の特許文献を見ると明らかであるように、日本語の特許文献では、コロン及びセミコロンは使用しないが、読点で同じように改行している場合が多い。

しかしながら、下記の例のように改行がされない例もある。

例 2) コロン及びセミコロンの後に改行がない例

中国語 特許文献	<p>一种电动吸尘器，其特征在于，具备：内装电动送风机并且在前方具有吸气口的电机室；配置在电机室的前方并且在前方具有吸气口的纸袋；在前方备有吸气软管接口并可使纸袋的吸气口与该吸气软管接口连通地内装纸袋的集尘室；使被吸入纸袋内含有尘埃的空气在纸袋内回旋而可离心分离尘埃的回旋流产生机构。</p>
日本語 特許文献	<p>電動送風機を内蔵し、前方に吸気口を有するモータ室と、モータ室の前方に配置され、かつ、前方に吸気口を有する紙パックと、吸気ホース接続口を前方に備え、この吸気ホース接続口に紙パックの吸気口を連通して紙パックを内蔵可能とした集塵室と、紙パック内に吸入された塵埃を含む空気を、紙パック内において旋回させて塵埃を遠心分離可能にする旋回流発生機構とを備えたことを特徴とする電気掃除機。</p>

基本的には、日本語の特許文献及び中国語の特許文献のいずれも、長文の請求項の記載では、改行して記載される傾向が強いが、上記例のように、長文でも改行されない場合も多い。

さらに、以下のように、中国語の特許文献でコロン及びセミコロンが使用されない場合もある。

例 3)コロン及びセミコロンを使用しない例

中国語 特許文献	一种育发剂组合物，其特征在于，在该育发剂组合物中，配合有按照干燥物换算占 0.00005~0.1%的生姜的提取物，并且还配合有牡丹的提取物。
日本語 特許文献	ショウガの抽出物を乾燥物換算で0.00005~0.1%配合し、さらにボタンの抽出物を配合することを特徴とする育毛剤組成物。

上記例 1)及び 2)のコロン及びセミコロンを使用する場合と、上記例 3)の使用しない場合とを比較すると、やはり、請求項の記載が長文である場合に、コロン及びセミコロンが使用される傾向が強い事が分かる。従って、日本語から中国語に翻訳する際には、以下のようなパターンを作成し、日本語の読点を中国語でよく使用されるコロン及びセミコロンに変換すると実際の中国語の特許文献により近い翻訳結果を得ることが可能である。

例)読点をコロン及びセミコロンに変換するパターン例

原文	<\$1>と、<\$2>と、<\$3>と、<\$4>とを備えたことを特徴とする<\$5>。
訳文	一种<\$5>，其特征在于，具备：<\$1>; <\$2>; <\$3>; <\$4>。

第8章 課題と対策

前述の第3章～第7章までで、日本国の文献データを日中機械翻訳するに当たって留意すべき点、及び中国の文献データを中日機械翻訳する場合との差異について分析してきた。それらを踏まえた上で、本章では、主に日本国の文献データを日中機械翻訳するに当たっての問題点・課題点及びそれに対する今後の対策について解説する。本調査により、考察される課題は、以下の通りである。

- (1) 機械翻訳不備の対策を実施する上での困難性
- (2) 定型可能表現の定型化

8.1. 機械翻訳不備の対策を実施する上での困難性

本調査の「第6章 調査結果の検証」では機械翻訳の不備を分析し、その対策として各種の辞書登録を行った。辞書登録によりその対象とした文や類似の文では訳が改善されるが、問題となるのが対象以外の文への悪影響（弊害）である。

ここでは弊害を起こす事例とその対策を類型ごとに分けて検討する。

(1) 特定の状況でのみ有効な訳語

例えば以下に示す文献データ D0121 の文では「補助ベッド」の訳語「補助台」としての辞書登録により訳が改善している。

例) 「補助ベッド」の登録による改善例 (文献データ D0121)

原文	本発明は、ミシンに関し、特にフリーアーム部に補助ベッドを装着したときに前記フリーアーム部と前記補助ベッドがフラットベッド部を形成するようにしてなるミシンに関する。
機械翻訳文 (改善策前)	本发明对缝纫机关于在在自由摇臂部特别安装补助床了的时候无上述摇臂部和上述补助床形成平地床部, 成为的缝纫机。
機械翻訳文 (改善策後)	本发明涉及缝纫机, 特别是涉及在在自由臂部安装辅助台了的时候上述上述自由臂部和 辅助台 形成平台部, 得的缝纫机。
対応する 中国特許文	本发明涉及一种缝纫机, 特别是涉及在自由臂部中安装 辅助台 时, 上述自由臂部与上述 辅助台 形成平台部的缝纫机。

上記文献ではミシンの部品について述べているのでこの訳が適切であるが、一般の文では「補助ベッド」は「ホテルなどでの追加の寝台」の意味であり、例えば「加床」などの訳が適切である。そのため、このような辞書登録は他の文の訳を悪化させる恐れがある。

名詞に複数の訳がある場合の訳し分けの手法としては、それぞれの訳になんらかの意味情報を付与し、共起する単語（主に係り先の動詞）で訳し分けるという手法があるが、この場合の「作業台」と「寝台」の意味の差異はわずかであり、詳細な意味情報体系を用意しないと弁別できない。このような詳細な意味情報の付与は非常に手間がかかる作業であり、また意味情報を付与しても「補助ベッドを設置した」のような文では係り先の動詞を見てもどちらの訳語なのかは判定できない。

このような場合に適用可能な一つの手法として辞書を分野ごとに用意するという方法が考えられる。「補助ベッド」であれば、通常使用する辞書（基本辞書）には「寝台」の意味で登録し、機械工学分野の専門語辞書には「作業台」の意味で登録するという手法である。機械工学でも「寝台」について述べる場合もあり完全ではないが、このような分類は意味情報の付与に比べて容易に行うことが可能で、現実的な手法であるといえる。

(2) 想定外の場合に使われる場合

例えば以下に示す文献データ B0433 の文では動詞「往復する」の訳語「往复运动」としての辞書登録により訳が改善している。

例)「往復する」の登録による改善例(文献データ B0433)

原文	前記研磨用ヘッドの拡張部 3 全体にダイヤモンドペーパー 4 を巻き付け、石英ガラス管内に挿入し、研磨用ヘッドの回転数を 100rpm、速度を 3m/min にして石英ガラス管の全長にわたって 80 回 往復 させた。
機械翻訳文 (改善策前)	把全体扩充部 3 前記研磨用脑袋的钻石纸 4 上绕在, 并且在石英玻璃管里插入, 并且让在 3m/min 用研磨用脑袋的旋转数做 100rpm, 速度, 持续, 石英玻璃管的全长 往返 于 80 回了。
機械翻訳文 (改善策後)	把全体扩大部 3 上述抛光头的金刚石纸 4 上绕在, 并且在石英玻璃管里插入, 并且在 3m/min 在抛光头的转速做 100rpm, 速度, 对石英玻璃管的全长持续, 80 回让 往复运动 了。
対応する 中国特許文	包括扩大部分 3 的整个抛光头用金刚石纸 4 围绕包裹和用钩和圈紧固件固定在其上面, 如在图 3 中所示。将抛光头插入到石英玻璃管中和沿着石英玻璃管的整个长度 往复运动 80 次, 同时抛光头以 100rpm(转/分钟)的转速转动和以 3m/min(米/分钟)的速度移动, 如在图 1 中所示。

この動詞「往復する」の訳語「往复运动」としての辞書登録は、「往復する」「往復させる」といった形であれば問題ないが、「往復運動させる」の形で出現する文献データ D0171 では以下に示すような弊害が起きている。

例)「往復する」の登録による弊害(文献データ D0171)

原文	本発明の方法を実施する二重環縫いミシン(1本針本縫いミシン)は、図1～図3に示すように、ミシン本体1に、周知のように、ミシン主軸2と、このミシン主軸2の回転に連動して、下端に1本の針3を取付けた針棒4を上下方向に 往復運動させる 針棒上下運動機構5と、ミシン主軸2の回転に連動してルーパー6を所定の楕円軌道の運動をさせるルーパー揺動運動機構7とを備えている。
機械翻訳文 (改善策前)	缝制实施本发明的方法双重的环, 并且好象在图 1~图 3 显示缝纫机(缝 1 部针书, 缝纫机), 并且, 在缝纫机本体 1, 在缝纫机主轴 2 和这个缝纫机主轴 2 的旋转众所周知连动, 正具有标准打数摆动活动构造让在针棒子上下活动构造让 运动往复 5 和缝纫机主轴 2 的旋转在在下端安装 1 部的针 3 了的针棒子 4 在上下方向连动, 规定的椭圆轨道在黄油炒面标准打数 6 运动黄油炒面 7。
機械翻訳文 (改善策後)	用本发明的方法实施的双线锁缝缝纫机(单针平缝缝纫机)具备有好象图 1~需要缝纫机主体图 3 中所示的的 1 公知是的缝纫机主轴 2 和机构针杆在在针一本也在在这台缝纫机主轴 2 的旋转连动的下端的 3 安装了了的针杆 4 垂直地 往复运动 , 运动 让的上下运动的 5 和弯针摆动运动机构椭圆在缝纫机主轴 2 的旋转连动的弯针 6 预定的轨道的运动让的 7。
対応する 中国特許文	实施本发明方法的双线锁缝缝纫机(单针平缝缝纫机)如图 1~图 3 所示那样, 在缝纫机主体 1 如公知的那样具有缝纫机主轴 2、针杆上下运动机构 5、及弯针摆动运动机构 7; 该针杆上下运动机构 5 与该缝纫机主轴 2 的回转连动地使在下端安装了 1 根针 3 的针杆 4 朝上下方向 往复运动 ; 该弯针摆动运动机构 7 与缝纫机主轴 2 回转连动地使弯针 6 进行预定的椭圆轨道的运动。

ここでは「往復運動させる」を「往復し、運動させる」と解釈し「往復运动, 运动」としてしまっている。「…する」の形のサ変動詞は例えば「…間を往復、あるいは…する」のように「する」なしでも動詞となることがあるためこのような訳出方法が行われる。このような場合には「往復運動する」の登録も行うことがひとつの対策となるが、「往復する」の登録時にそのような事例を想定することは困難である。

一方でこういった弊害は、ある程度の量のテキストを用いて登録前と登録後の翻訳の差異を調査することで容易に発見することができる。特許文では大量の原文を用意することが可能であり、また大規模分散処理など大量の計算機資源を用いて翻訳差異を短時間で算出することも近年では容易になってきているため、こういったシステムを利用して調査することで弊害を事前に発見し対策することが可能であると思われる。

8.2. 定型可能表現の定型化

本調査の「第4章 定型化可能な表現の分析」では、日本語の特許文献を日中機械翻訳する際の定型化について分析し、実際に定型化を行った。しかしながら、定型化に関しては問題点・課題点が存在する。ここでは、その問題点・課題点を説明する。

8.2.1. 定型化の問題点

定型化により得られたパターンは、うまい具合にマッチすれば訳質を大幅に向上させる事ができる。しかしながら、変数部分にマッチする範囲が広くなり過ぎて、構文構造が破壊され、逆に訳質を低下させてしまう場合もある。さらに、意図しない部分とマッチする事により訳質を低下させてしまう場合もある。このような問題点を解決する為の対策としては、以下の4つの方法(1)~(4)を実現する事でパターンに望ましくないマッチをさせない方法が考えられる。

(1)数字のみにマッチする変数を導入する方法

下記に例を示すパターン例1を用いて説明する。

例)パターン例1

原文 パターン	図<\$1>に、<\$2>を示す。
訳文 パターン	在图<\$1>中示出了<\$2>。

上記のパターン例1は、下記に例を示すような原文にマッチし、模範訳のような好ましい機械翻訳文が訳出される事を想定している。

例)パターン例1が有効な例

原文	図【3】に、【プレアニール条件】を示す。
模範訳	在图【3】中示出【退火条件】。

しかしながら、上記のパターン例1は、下記に例を示すような原文の場合に、下記のマッチ結果のような意図しないマッチ結果となり、模範訳とはかけ離れた好ましくない機械翻訳文が訳出されてしまう。

例)パターン例 1 が不適切な例

原文	図 5 および図 6 では、便宜的に、ポリカルボン酸エステルとしてコハク酸ジメチルを用い、有機溶媒として n-ブタノールを用いた例を示す。
模範訳	在图 5 及图 6 中，为了便于说明，示出了作为聚羧酸酯使用琥珀酸二甲酯、作为有机溶剂使用 n-丁醇的例子。
マッチ結果	図【5 および図 6 では、便宜的】に、【ポリカルボン酸エステルとしてコハク酸ジメチルを用い、有機溶媒として n-ブタノールを用いた例】を示す。
機械翻訳文	在图【5 以及图 6 权宜】中示出了【作为聚羧酸酯使用琥珀酸二甲酯、作为有机溶剂使用 n-丁醇的例子】。

このような問題を解決する対策としては、数字のみにマッチする変数を導入し、<\$1>をそのような変数にする事で、上述した不適切な例の原文をマッチさせなくする方法が挙げられる。マッチさせなくすることで、パターンを使用しない機械翻訳が行われ、機械翻訳の訳質低下を避けることができる。

(2) 読点を含まない変数を導入する方法

下記に例を示すパターン例 2 を用いて説明する。

例)パターン例 2

原文 パターン	まず、<\$1>について説明する。
訳文 パターン	首先，对<\$1>进行说明。

上記のパターン例 2 は、下記に例を示すような原文にマッチし、模範訳のような好ましい機械翻訳文が訳出される事を想定している。

例)パターン例 2 が有効な例

原文	まず、【半導体記憶装置 100 ヘデータを書込む動作】について説明する。
模範訳	首先，对【向半导体存储器 100 写入数据的工作】进行说明。

しかしながら、上記のパターン例 2 は、下記に例を示すような原文の場合に、下記のマッチ結果のような意図しないマッチ結果となり、模範訳とはかけ離れた好ましくない機械翻訳文が訳出されてしまう。

例)パターン例 2 が不適切な例

原文	まず、無機配向膜の形成方法の説明に先立ち、本発明の液晶パネルについて説明する。
模範訳	首先，在说明无机取向膜的形成方法之前，对本发明的液晶面板进行说明。
マッチ結果	まず、【無機配向膜の形成方法の説明に先立ち、本発明の液晶パネル】について説明する。
機械翻訳文	首先，对【在说明无机取向膜的形成方法之前，本发明的液晶面板】进行说明。

上記例の場合、変数<\$1>に「無機配向膜の形成方法の説明に先立ち、」が入ってしまうことにより、「対」の前にその句が生成されることができなくなってしまう。

このような問題を解決する対策としては、読点を含まない変数を導入し、<\$1>をそのような変数にすることにより、短いフレーズにマッチさせる事を意図するパターンで誤マッチを防止させる方法が挙げられる。

なお、この方法は、後述する「(4)マッチする品詞を指定できる変数を導入する方法」を導入し、変数<\$1>を名詞等に限定すれば不要となるが、読点を含まない変数を導入する方法の方が比較の実装が容易な為、先にこの方法を採用する事は有効である。

(3) 形態素解析を行った結果の単語単位でのマッチを行う方法

下記に例を示すパターン例 3 を用いて説明する。

例)パターン例 3

原文 パターン	図<\$1>は図<\$2>の<\$3>を示す。
訳文 パターン	图<\$1>表示图<\$2>的<\$3>。

上記のパターン例 3 は、下記に例を示すような原文の場合に、下記のマッチ結果のような意図しないマッチ結果となり、模範訳とはかけ離れた好ましくない機械訳文が訳出されてしまう。

例)パターン 3 が不適切な例

原文	図 1 1 は、図 6 又は図 8 のピーク抽出部 6 1 4 の代替例を示す図である。
模範訳	图 11 表示图 6 或图 8 的峰值提取部 614 的替代例。
マッチ結果	图【1 1 は、図 6 又】は图【8】の【ピーク抽出部 6 1 4 の代替例】を示す図である。
機械訳文	图【11 图 6 另外】表示图【8】的【峰值提取部 614 的替代例】。

上記例の場合、変数<\$1>に「又は」の「又」が入ってしまい、「又は」を分断してしまっている。

このように助詞等が単語の一部に誤マッチしてしまう問題を解決する対策としては、単純な文字列単位のマッチではなく、形態素解析を行った結果の単語単位でのマッチを行う方法が挙げられる。ただし、形態素解析を実行することにより、速度が低下するというデメリットが考えられる。

(4) マッチする品詞を指定できる変数を導入する方法

上記の「(2) 読点を含まない変数を導入する方法」で述べたように、変数の中身の品詞を指定し、誤マッチを防止する方法が考えられる。

例えば、上記の「(2) 読点を含まない変数を導入する方法」のパターン例 2 の変数<\$1>を名詞句と限定する。すると、不適切なマッチ結果である「無機配向膜の形成方法の説明に先立ち、」がその変数内に入っている場合は、変数に対応する部分が名詞句ではなくなるので、マッチしなくなり、誤マッチを防止することができる。

8.2.2. 定型化の課題

前節「8.2.1. 定型化の問題点」では、パターンに望ましくないマッチをさせない方法により訳質の低下を防止する対策を述べた。ここでは、逆に、適切なマッチを行う事で訳質を向上させる為の課題を述べる。

(1)文頭の副詞を抜いてマッチさせ、合成する改善策

下記に例を示すパターン例 4 を用いて説明する。

例)パターン例 4

原文 パターン	<\$1>としては、<\$2>、<\$3>、<\$4>、<\$5>等が挙げられる。
訳文 パターン	作为<\$1>，可以举出<\$2>、<\$3>、<\$4>、<\$5>等。

上記のパターン例 4 は、下記に例を示すような原文の場合に、下記のマッチ結果のような意図しないマッチ結果となり、模範訳とはかけ離れた好ましくない機械翻訳文が訳出されてしまう。

例)パターン例 4 が不適切な例

原文	ここで、電子部品としては、リードピン、半導体チップ搭載用基板、放熱板、半導体チップ自身等が挙げられる。
模範訳	此处，作为电子部件，可以举出引线脚、半导体芯片搭载用基板、放热板、半导体芯片自身等。
マッチ結果	【ここで、電子部品】としては、【リードピン】、【半導体チップ搭載用基板】、【放熱板】、【半導体チップ自身】等が挙げられる。
機械翻訳文	【作为此处，电子部件】，可以举出【引线脚】、【半导体芯片搭载用基板】、【放热板】、【半导体芯片自身等】。

上記例の場合、文頭の変数<\$1>に副詞が入ってしまい、副詞「ここで、」(此处，)の訳出位置がおかしくなっている。

このような状況を改善する対策としては、原文の文頭副詞「ここで、」を抜いてマッチさせて機械翻訳結果を作成してから、最後に「ここで、」の訳の「此处，」を訳文に付与する方法が挙げられる。

この方法は誤マッチを防止するだけでなく、文頭副詞を残した以下のようなパターン例 5 を別途用意する必要がなくなり、より多くの文に適用できるようになるという利点も持ち合わせている。

例)パターン例 5

原文 パターン	ここで、<\$1>としては、<\$2>、<\$3>、<\$4>、<\$5>等が挙げられる。
訳文 パターン	此处，作为<\$1>，可以举出<\$2>、<\$3>、<\$4>、<\$5>等。

(2) 主文パターンと節パターンを使用する改善策

「4.3.2.1.2. 主文パターン文の定義」及び「4.3.2.2.2. 節パターンの定義」において、これらを実現する翻訳エンジンが実用化されていない事を述べた。また、主文パターン文の中のサブパターンとして節パターンを利用することにより、精度向上を期待する事ができるとも述べた。主文パターン文は、主文内に含む節をすべて変数化するため、変数がマッチする範囲が広がる。すると、マッチする文が増える反面、誤マッチが発生しやすくなる。そのため、上記で述べてきた訳質の低下を解決する方法のうち、特に「(4) マッチする品詞を指定できる変数を導入する方法」の導入が必要となる。主文パターンの変数に品詞が指定できると、名詞節・連用中止節・連体修飾節・副詞節等の指定もできるようになる。これにより、誤マッチを防止しつつ、それらの品詞の節パターンを組み合わせることで訳質の向上を図ることができる。

つまり、主文パターンでマッチ率が向上し、品詞指定変数で誤マッチが回避され、指定品詞の節パターンで訳質が向上することが期待される。

参考文献:

- [1] 中华人民共和国国家知识产权局（中華人民共和国国家知識産権局）
<http://www.sipo.gov.cn/>
- [2] 中国知的産権局
<http://www.cnipr.com/>
- [3] 特許電子図書館 <http://www.ipdl.inpit.go.jp/homepg.ipdl>
- [4] 平成 21 年度「中国公開特許公報の機械翻訳による日本語での提供に関する調査」
特許庁
- [5] 平成 20 年度「中国語機械翻訳技術に関する調査」特許庁
- [6] 平成 22 年度 平成 21 年度特許版・産業日本語委員会報告書「産業日本語」～産業技術文書を明晰に記述するための日本語仕様～ 一般財団法人日本特許情報機構 特許情報研究所
- [7] 平成 21 年度「中国における特許翻訳の現状」AAMT/Japio 特許翻訳研究会 シンポジウム発表資料

凡例
<p>行の使い方には「原文訳文行」「訳語不備指摘行」「訳語変化評価行」「翻訳メモリマッチ評価行」の4種類あり、それぞれで列の使い方が異なる。「原文訳文行」以外は背景をグレーにしてある。</p>

原文訳文行

日中の公開特許公報より取得した文を対応付けて並べ、機械翻訳を付与し、評価を付与した行。背景は白。

日本公開特許公報	日本公開特許公報より取得した翻訳対象文
機械翻訳	改善策実施前の機械翻訳結果
機械翻訳(修正後)	改善策実施後の機械翻訳結果
訳ふり(日中)	改善策実施前の機械翻訳結果の見出し語と訳語の対応を示す。分析の参考のために用いる。「日本語【中国語】」形式で、日本語の語順で並んでいる。
訳ふり(中日)	改善策実施前の機械翻訳結果の見出し語と訳語の対応を示す。分析の参考のために用いる。「中国語【日本語】」形式で、中国語(機械翻訳結果)の語順で並んでいる。 ★ マークは日本語の単語に対応する中国語単語が複数ある場合を示している。例えば「…際に」が「在…的时候」と訳されている場合は、在【際に、★】…的时候【際に、★】となる。
中国公開特許公報	日本公開特許公報と優先権主張によって対応付けられる中国公開特許公報から文を取得し、日本公開特許公報の文との対応をとって並べたもの。日本語の一文が中国語で複数文になっている場合は、最初の行に日本語文と中国語の最初の文を入れ、続く行には中国語文のみを一行に一文ずつ入れる。
訳語	(原文訳文行では空欄)
辞書評価	(原文訳文行では空欄)
TM評価	(原文訳文行では空欄)
位置	訳出位置の不備がある場合に「1」を記入。
コメント	「位置」の誤りの内容を具体的に記入。
した	助動詞「～した」の訳し方が誤っている場合に「1」を記入。
される	助動詞「～される」の訳し方が誤っている場合に「1」を記入。
させる	助動詞「～させる」の訳し方が誤っている場合に「1」を記入。
助動	その他の助動詞の訳し方が誤っている場合に「1」を記入。
コメント	「した」「される」「させる」「助動」のいずれかに誤りのある場合に内容を具体的に記入。
訳抜	訳抜けがあった場合に「1」を記入。
コメント	「訳抜」の内容を具体的に記入。
句切不備	カンマの位置が不適切な場合に「1」を記入。
コメント	「句切不備」の内容を具体的に記入。
其他	上記以外の不備があれば、その内容を具体的に記入。

注)不備の分類については報告書第3章「表3.4.2-1 機械翻訳不備の分類」および「3.2. 翻訳不備の要因分析」を参照のこと。

訳語不備指摘行

機械翻訳で訳語に不備のあった単語を切り出し、適切な訳語と品詞を付与した行。背景はグレー。また、改善策実施後にその通りの訳になったかの評価を行った。

日本公開特許公報	日本語文より切り出した訳語に不備のある単語。
機械翻訳	改善策実施前の訳語。改善策で訳が変わった時のみ示す。「 」で区切られているものは複数の単語を組み合わせて訳されていたことを示し、「…」は文中の離れた場所に訳出されていたことを示す。
機械翻訳(修正後)	改善策実施後の訳語。改善策で訳が変わった時のみ示す。
訳ふり(日中)	(訳語不備指摘行では空欄)
訳ふり(中日)	(訳語不備指摘行では空欄)
中国公開特許公報	適切な訳語を評価者が記入。
訳語	切り出した単語の品詞を記入(辞書化に際しては記入内容を、名詞、動詞、形容詞、副詞、その他に整理した)。
辞書評価	改善策実施後の変化を、1=改善、0=同等、-1=悪化で評価。変化がなかった場合は空欄になっているが、集計の際には同等扱いにしている。
TM評価	(訳語不備指摘行では空欄)
	(これ以降の列は訳語不備指摘行では空欄)

注)不備の分類については報告書第3章「表3.4.2-1 機械翻訳不備の分類」および「3.2. 翻訳不備の要因分析」を参照のこと。

訳語変化評価行

改善策実施後に訳の変化のあった単語についてその評価を付与した行。背景はグレー。ただし、訳語不備指摘行で指摘のあった単語の訳語変化の評価は訳語不備指摘行で行った。

日本公開特許公報	改善策により訳語が変化した日本語単語
機械翻訳	改善策実施前の訳語。「 」で区切られているものは複数の単語を組み合わせて訳されていたことを示し、「…」は文中の離れた場所に訳出されていたことを示す。
機械翻訳(修正後)	改善策実施後の訳語。
訳ふり(日中)	(訳語変化評価行では空欄)
訳ふり(中日)	(訳語変化評価行では空欄)
中国公開特許公報	(訳語変化評価行では空欄)
訳語	その訳語の出自を示す。[0A]がマイクロ分析の不備対策辞書、[0B]が定型フレーズ辞書。
辞書評価	改善策実施後の変化を、1=改善、0=同等、-1=悪化で評価。
TM評価	(訳語変化評価行では空欄)
	(これ以降の列は訳語変化評価行では空欄)

翻訳メモリマッチ評価行

改善策実施後に翻訳メモリのマッチによって訳の変化のあった文についてその評価を付与した行。背景はグレー。

日本公開特許公報	マッチした翻訳メモリの日本語内容(パターンの場合は<\$1>などで変数を示す)
機械翻訳	(翻訳メモリマッチ評価行では空欄)
機械翻訳(修正後)	マッチした翻訳メモリの中国語内容(パターンの場合は<\$1>などで変数を示す)
訳ふり(日中)	(翻訳メモリマッチ評価行では空欄)
訳ふり(中日)	(翻訳メモリマッチ評価行では空欄)
中国公開特許公報	(翻訳メモリマッチ評価行では空欄)
訳語	(翻訳メモリマッチ評価行では空欄)
辞書評価	(翻訳メモリマッチ評価行では空欄)
TM評価	改善策実施版の翻訳メモリによる変化を1=改善、-1=悪化で評価。翻訳メモリにマッチした場合訳は大きく変化するので、「同等」とは評価せず、改善か悪化かのどちらかにした。
	(これ以降の列は翻訳メモリマッチ評価行では空欄)