

## 5 調査内容

調査の流れ、および調査観点について記述する。

---

5.1 調査の流れ

5.2 調査観点

---

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	5-1

### 5.1 調査の流れ

以下の流れで調査を行い、目標とする成果を明示する。

- ①目標とする成果を元に、調査観点を洗い出し、具体的な調査項目に細分化する。
- ②ユーザ、および業者にて調査項目の調査を実施する。
  - ユーザは、モデル検証システムを利用し、アンケートを記載する。
  - 業者は、ツールを用いてモデル検証システムへ定量的な検索処理を実施する。
- ③アンケートによる定性評価と各種ログによる定量評価を集計・算出し、多面的に分析する。
- ④調査結果を「調査報告書」としてまとめ、目標とする成果を明示する。

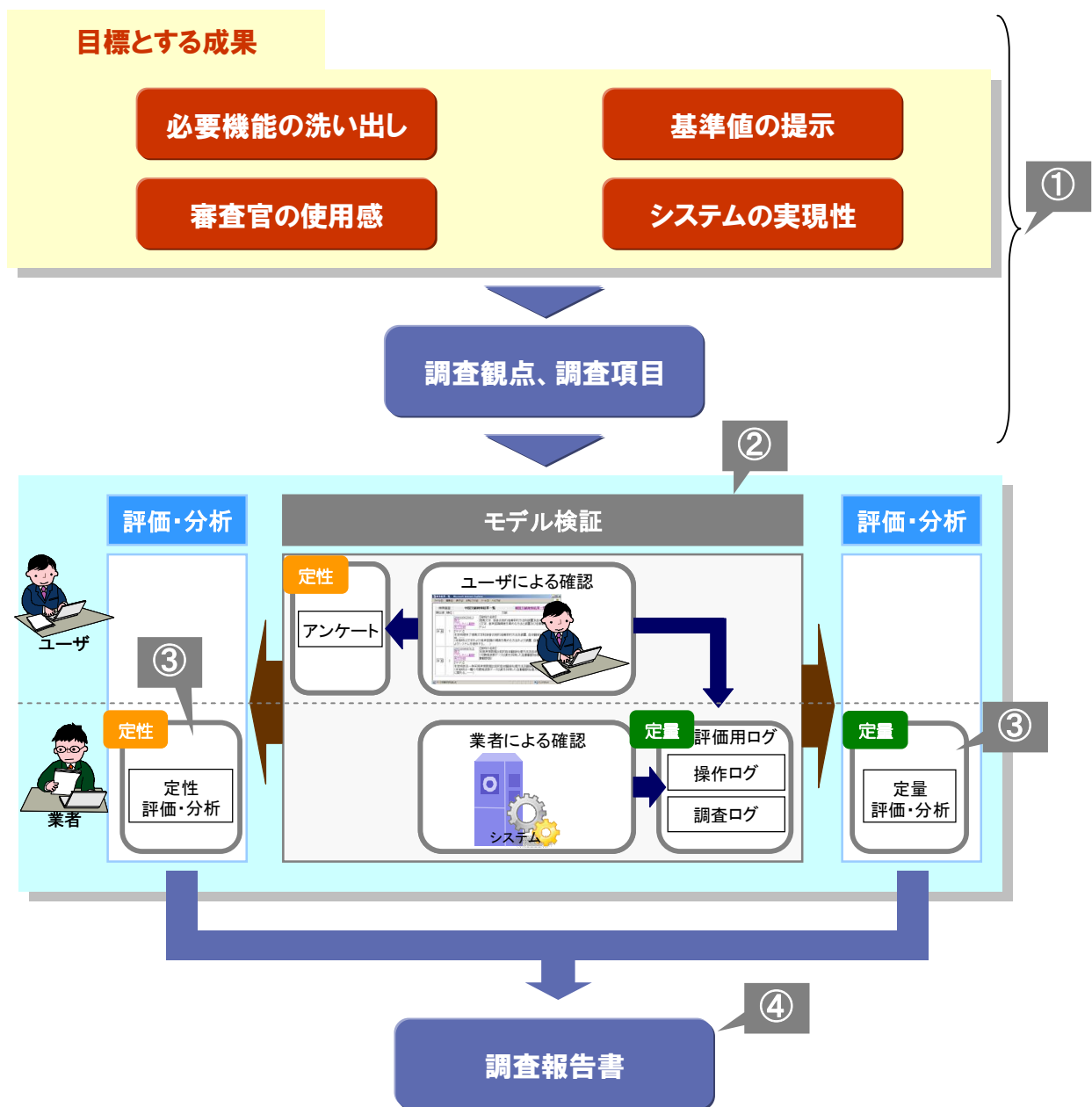


図 5.1-1 調査の流れ

5.2 調査観点

「目標とする成果」を導く調査観点は、『表 5.2-1 調査観点』の通りである。

表 5.2-1 調査観点

項番	観点項目	目的	観点内容	対象言語 中 韓	実施者			
					定量	定性	机上	
<b>1 必要機能の洗い出し</b>								
1-1	翻訳方式 (コンテンツ翻訳方式/キーワード翻訳方式)	各翻訳方式の有効性を確認する	各翻訳方式での検索精度の比較、および定性評価から分析	● ●	- ○	○	-	-
1-2	翻訳技術							
1-2-1	辞書自動メンテナンス	辞書自動メンテナンス方法の有効性を確認する	1. 中韓の言語/特許文書の特性調査 2. 辞書自動メンテナンスの実施・活用 (韓国は、手動ユーザ辞書登録を実施)	●	- ○	-	-	-
1-2-2	シソーラス辞書を用いた訳し分け規則	訳し分け規則(シソーラス辞書)を用いる技術が有効であるかを確認する	訳し分け規則(シソーラス辞書)を活用し、活用前後の検索精度、翻訳精度を比較	● ●	- ○	-	-	-
1-3	検索技術							
1-3-1	概念検索							
1-3-1-1	検索方式 (自然文検索方式/全文検索方式)	各検索方式の有効性を確認する	自然文検索と全文検索の有効性を定性評価から分析	●	- -	- ○	-	-
1-3-1-2	類似度順表示	検索結果の類似度順表示の有効性を確認する	類似度順表示の有効性を定性評価から分析	● ●	- -	○	-	-
1-3-2	訳語候補展開	全文検索における訳語候補展開機能の有効性を確認する	全文検索における訳語候補展開の有効性を定性評価から分析、および検索精度の分析	●	- ○	- ○	-	-
1-3-3	検索対象範囲 (全文、要約、実施例)	全文、要約、実施例のいずれの検索対象範囲が有効であるかを確認する	1. 中韓の特許文書の特性調査 2. 各範囲での検索精度を比較	● ●	- ○	-	-	-
1-4	審査関連情報の活用							
1-4-1	分野別の辞書活用	メンテナンス辞書を使用することの有効性を確認する	メンテナンス辞書の有無による有効性を定性評価から分析、および検索精度を比較	●	- ○	○ ○	-	-
1-4-2	書誌情報による検索	多言語文献において検索条件に書誌情報を使用することの有効性を確認するため。	検索条件の書誌情報の有無による有効性を定性評価から分析、および検索精度を比較	● ●	○	- ○	-	-
1-4-3	対応特許表示	多言語文献に対する対応特許、英文抄録、日本語訳を表示させることの有効性を確認する	検索結果の対応特許表示の有効性を定性評価から分析	● ●	- -	○	-	-
<b>2 基準値の提示</b>								
2-1	システムの検索精度	“審査に使える”システムの基準値を算出するための参考値	審査官が“審査に使える”と判断した全てのシステムの検索精度を算出する。	● ●	○ ○	○	-	-
<b>3 審査官の使用感</b>								
3-1	審査業務への有効性	多言語横断検索が審査に有効であるかを確認する	モデル検証システム(多言語横断検索)を利用し、システム全体に対する定性評価を分析	● ●	- -	○	-	-
<b>4 システムの実現性</b>								
4-1	対象言語のスケールビリティ							
4-1-1	辞書(専門辞書・訳し分け規則)	新検索システム導入後のスケールビリティを検証	対象言語の拡張時における専門辞書・訳し分け規則の拡張性評価(技術的課題)	-	- -	- -	-	○
4-1-2	検索エンジン		対象言語の拡張時における検索エンジンの拡張性評価(技術的課題)	-	- -	- -	-	○
4-2	多言語横断検索導入時の新検索システム構成案	多言語横断検索の新検索システム導入システム構成の立案	多言語横断検索機能を用いた新検索システムのシステム構成案を検討	-	- -	- -	-	○

## 6 技術的内容の説明

多言語横断検索技術の基本的なアルゴリズムについて記述する。

---

6.1 翻訳技術の説明

6.2 検索技術の説明

---

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	6-1

6.1 翻訳技術の説明

6.1.1 翻訳方式

機械翻訳における一般的な翻訳方式は、『表 6.1-1 一般的な翻訳方式』の通りである。

表 6.1-1 一般的な翻訳方式

項番	翻訳方式	概要
1	規則ベース	人手で記述した規則に基づいて翻訳処理を行う。 商用システムで主流の方式で、「トランスファ方式」との組み合わせが多い。
2	統計ベース	大量の対訳用例から単語や句のレベルで統計学習。 近年研究は非常に活発だが実用システム(Google 翻訳等)は少数。
3	用例ベース	大量の対訳用例を節や文のレベルで利用。規則ベースとの併用が有効。

6.1.2 中国語翻訳

本調査の日中／中日翻訳に用いた「The 翻訳」は、「規則ベースのトランスファ方式」を採用している。『図 6.1-1 日中／中日翻訳の処理概要』に、「The 翻訳」の処理概要を示す。

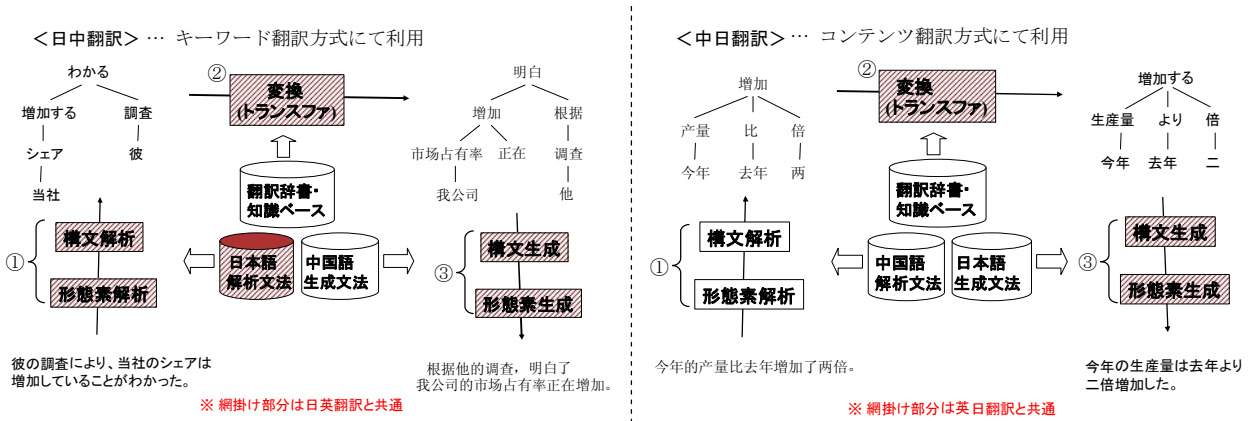


図 6.1-1 日中／中日翻訳の処理概要

トランスファ方式による機械翻訳処理は、以下の3ステップから成る。

① 原文解析

解析文法を参照して処理を行う(中日翻訳は中国語解析文法、日中翻訳は日本語解析文法)。

両者とも、入力文を単語分割し、各単語に品詞を割り当てる「形態素解析」(詳細については、『6.1.2(2)1形態素解析:単語分割と品詞付与』を参照)を行った後、単語間の関係、すなわち文の構造を決定する「構文解析」(詳細については、『6.1.2(2)2構文解析:単語間の関係を求める』を参照)を行う。

② 変換(トランスファ)

翻訳辞書を参照して、入力言語の語彙・構造を出力言語の語彙・構造に変換する。

③ 訳文生成

生成文法を参照して処理を行う(中日翻訳は日本語生成文法、日中翻訳は中国語生成文法)。

両者とも、語順を決定する構文生成を行った後、語尾変化などを処理する「形態素生成」を行い、最終的な訳文が出力される。

翻訳技術における留意ポイントを以下に示す。

(1) 訳し分け規則を用いて、複数訳語を持つ語に対応

変換(トランスファ)時に「訳し分け規則」という、係り受け関係にある単語を参照し、適切な訳語を選択する処理を行う。

(例)「わかる」

彼の調査により、当社のシェアは増加していることがわかった。

中国語がわかる

だれがいつ/どこで/どう~するかわかる

中国語では「わかる」に対して、「明白」「懂」「知道」の複数の意味が存在するため、訳し分け規則で決定している。

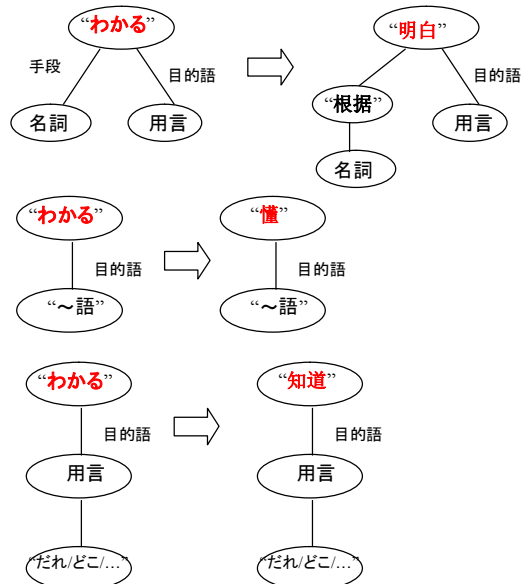


図 6.1-2 日中翻訳での訳し分け規則の例

「The 翻訳」では日中／中日辞書に、対訳や文法属性(品詞等)だけでなく、この「訳し分け規則」の情報を持たせているため、日英／英日の変換処理を利用して日中／中日の変換処理が可能である。

(2) 中国語の文章解析は、文例集を元にした統計モデルから最適解を決定

コンテンツ翻訳方式で利用する中日翻訳は、中国文献の中国語を、「形態素解析」→「構文解析」の2ステップで解析している。

1) 形態素解析: 単語分割と品詞付与

中国語は語の切れ目が明示されない膠着(こうちやく)語であるため、単語分割を行う際に曖昧性が生じる。

形態素解析あいまい性の例

「不断开发引领世界的新技术」 (世界をリードする新技术を絶え間なく開発する)

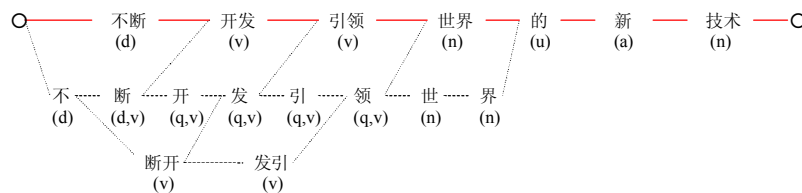


図 6.1-3 形態素解析の曖昧性の例

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	6-3

このため、正しく単語分割された品詞が付与された大量のテキスト(品詞タグ付きコーパスと言う)を用いて正解例を学習し、そこから確立統計を用いて最適解を算出する。

「The 翻訳」では、中華人民共和国教育部 言語文字応用研究所が開発した「品詞タグ付きコーパス」の一部(訳 1,700 万語)を利用している。

12月/t 31日/t ,/w 中共中央/nt 总书记/n 、/w 国家/n 主席/n 江/nr 泽民/nr 发表/v  
 1998年/t 新年/t 讲话/n 《/w 迈向/v 充满/v 希望/n 的/u 新/a 世纪/n 》/w 。/w  
 同胞/n 们/k 、/w 朋友/n 们/k 、/w 女士/n 们/k 、/w 先生/n 们/k :/w  
 在/p 1998年/t 来临/v 之际/f ,/w 我/r 十分/m 高兴/a 地/u 通过/p 中央/n 人民/n  
 广播/vn 电台/n 、/w 中国/ns 国际/n 广播/vn 电台/n 和/c 中央/n 电视台/n ,/w  
 向/p 全国/n 各族/r 人民/n ,/w 向/p 香港/ns 特别/a 行政区/n 同胞/n 、/w 澳门/ns  
 和/c 台湾/ns 同胞/n 、/w 海外/s 侨胞/n ,/w 向/p 世界/n 各国/r 的/u 朋友/n  
 们/k ,/w 致以/v 诚挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn !/w

図 6.1-4 品詞タグ付きコーパスの例

2) 構文解析:単語間の関係を求める

構文解析では、形態素解析結果に対し、チャート法というアルゴリズムを用いて文脈自由文法\*1 (CFG:Context Free Grammar)をボトムアップに適用する。『図 6.1-5 構文解析の曖昧性の例』は、CFG を適用して得られた構文森である。意味的に正しい解釈は、実線で示された動詞句②であるが、構文的には動詞句①や名詞句③の解釈も可能であり、曖昧性の解消が必要である。

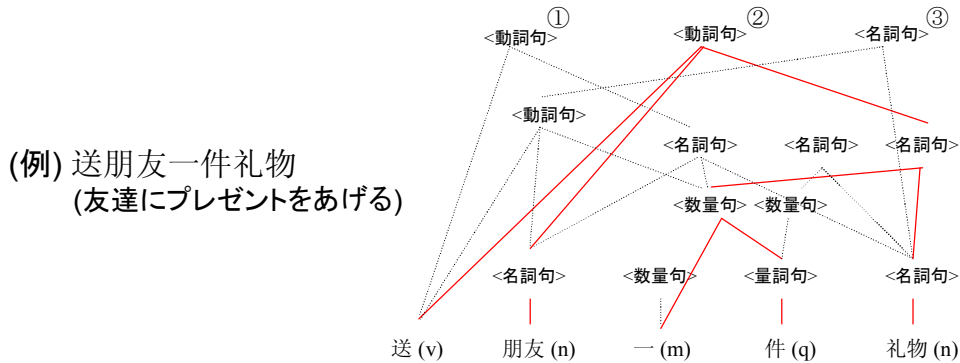


図 6.1-5 構文解析の曖昧性の例

\*1 句の構成規則を、<動詞句>→<動詞>+<名詞句>のような形式で定義する。”→”の右辺の語句の並びが左辺の句を構成する。



### 6.1.3 韓国語翻訳

本調査の日韓／韓日翻訳に用いた「J-Server」は、「規則ベース」を採用している。

多くの語は複数の訳語を持っているため、適切な訳し分けが必要であり、各動詞に対し複数の各フレーム辞書を構築し、動詞と名詞の意味的整合性を調べる事により訳し分けを可能にしている。



図 6.1-7 日韓／韓日翻訳の処理概要

韓国語は、ハングルが表音文字であり、発音が同じで意味が異なる同音異義語が多い。同音異義語は、各語の関係、属性、文脈情報により、訳し分けを行う。

- ・ 경기 (景気、競技)
- ・ 지방 (地方、脂肪)
- ・ 의사 (意志、医師)
- ・ 사전 (事典、事前)

図 6.1-8 韓国語における同音異義語の例

また、翻訳辞書を強化して、複数単語の組み合わせにより意味を成す複合語を登録することにより、同音異義語の処理を強化している。

#### 複合語の登録例

- ・ 競技種目
- ・ 親善協議
- ・ 景気回復
- ・ 景気変動
- ・ 皮下脂肪
- ・ 地方行政

図 6.1-9 複合語登録の例

## 6.2 検索技術の説明

### 6.2.1 単語索引

単語索引とは、データ登録時に検索対象となる文書を走査して単語を取り出し、高速な検索が可能となるような索引データとして作成されるものである。

本調査では、日本語を対象とした自然言語検索機能をベースとして、試験的に中国語、韓国語にも拡張して対応した。対応箇所は、単語索引を作成するために文書から単語を取り出す処理である。

自然言語検索のため、中国語・韓国語それぞれの単語索引を作成する必要があり、以下の 2 つの方法を用いた。

#### (1) 形態素解析

辞書や文法を用いた処理によって、単語を取り出す方法である。本調査では、中国語の単語索引を作成するために用いた。

辞書や文法を用いた処理であるため、抽出単語数を一定の数に抑えることが可能であるが、辞書の品質影響を受けやすく、検索漏れが生じる可能性がある。

#### (2) 2-gram

文字列を 2 文字単位(2-gram)で切り出し、切り出した文字列を単語と見なして処理する方法である。本調査では、韓国語の単語索引を作成するために用いた。

2文字単位で単語を抽出するため、意味を成さない単語抽出を行い単語索引の肥大化を招く危険性があるが、検索漏れが少ない利点がある。

### 6.2.2 概念検索

本調査で利用した概念検索エンジンでは、中国語の場合は形態素解析、韓国語の場合は 2-gram という技術を用いて、取り出された単語に基づく単語索引を利用し実現される。処理の流れは以下の通りである。

- ① 検索質問(自然文)を形態素解析(韓国語は 2-gram)により単語に分割する。
- ② 分割された単語の中から品詞により検索に利用する検索語を選択する。
- ③ 検索語と検索対象の文書を照合し、各文書の検索スコアを計算する。
- ④ 検索スコア順に文書を並べる。

検索スコアの計算は、検索質問から抽出された単語の文書中の出現頻度やその単語を含む文書数、文書の長さなどの統計情報に基づく方式で行われる。

このスコアは、大きく以下の 3 つのルールで決定する。

- ✓ 単語の出現文書数が少ないほどスコアが高い。
- ✓ 単語の文書中の出現頻度が多いほどスコアが高い。
- ✓ 文書長が短いほどスコアが高い。

なお、このルールは、本調査の概念検索エンジンに限らず、広く一般的に使われるルールである。

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	6-7

### (1) 単語の出現文書数が少ないほどスコアが高い

基本的には、検索質問中に含まれる単語と共通する単語を多く含む文書の検索スコアが高くなり、その文書が上位にランキングされる。

但し、文書をランキングする上で、すべての単語を等しく扱うわけではなく、検索質問のどの単語が重要であるかを判定するため、検索対象のデータベース中で少数の文書にしか含まれない単語を重視する重み付けを行う。これは、多くの文書に含まれる単語は、検索における弁別能力が低いという特徴に基づいている。

### (2) 単語の文書中の出現頻度が多いほどスコアが高い

文書側の単語については、その文書中の単語のうち、出現頻度が高い単語を重視する重み付けを行う。これは、ある文書で重要な単語は、その文書で繰り返し用いられるという特性に基づいている。

### (3) 文書長が短いほどスコアが高い

この単語の出現頻度は、文書が長くなると多くの単語の出現頻度が高くなる傾向がある。検索スコアの計算では、文書長の影響が出ないように、文書長が短いほどスコアが高くなる様な補正を行う。

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	6-8

## 7 調査方法

調査の具体的な方法、検索課題、チューニング内容などについて記述する。

- 
- 7.1 調査項目
  - 7.2 実施方法
  - 7.3 検索課題
  - 7.4 モデル検証データ
  - 7.5 評価方法
- 

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-1

## 7.1 調査項目

「調査観点」を達成するための調査は、『表 7.1-1 実施概要』で示す大きな 7 つの実施項目に基づき、ユーザ検証および業者検証にて検証を実施する。

表 7.1-1 実施概要

項番	実施項目	実施概要
1	処理方式の比較	1回目の検証と4回目の検証にて調査する。 キーワード翻訳方式とコンテンツ翻訳方式のそれぞれにおいて、同一の検索条件で検索し、検索精度の観点から比較を行い、多言語横断検索において効果のある方式を検証する。
2	自動辞書メンテナンスの有効性	2回目から4回目の検証を通して調査する。 同一の検索条件で検索し、検索精度の観点から比較を行い、自動辞書メンテナンス機能の有効性を検証する。
3	訳語候補展開機能の有効性	4回目の検証にて調査する。 訳語候補展開機能と文脈を意識して翻訳する自然文検索にて検索し、検索精度の観点から比較を行い、訳語候補展開機能の有効性を比較する。
4	翻訳精度の確認	3回目の検証にて調査する。 同一文献において以下の項目より翻訳精度の観点から比較し検証する。 ①分野別専門辞書を使用せずに翻訳 ②製品版の分野別辞書を使用して翻訳 ③自動辞書メンテナンスにて作成した辞書を使用して翻訳 ④シソーラス辞書を使用して翻訳した文献を審査官が確認(4回目の検証)
5	IPC指定の有効性	2回目の検証にて調査する。 IPCの指定の有無のみが異なる、同一の検索条件で検索し、検索精度の観点から比較を行い、IPC指定の有効性を検証する。
6	検索対象範囲の比較	1回目から4回目の検証にて調査する。 検索対象範囲のみ異なる、同一の検索条件で検索し、検索精度の観点から比較を行い、検索対象範囲の違いによる検索結果の特徴を検証する。

## 7.2 実施方法

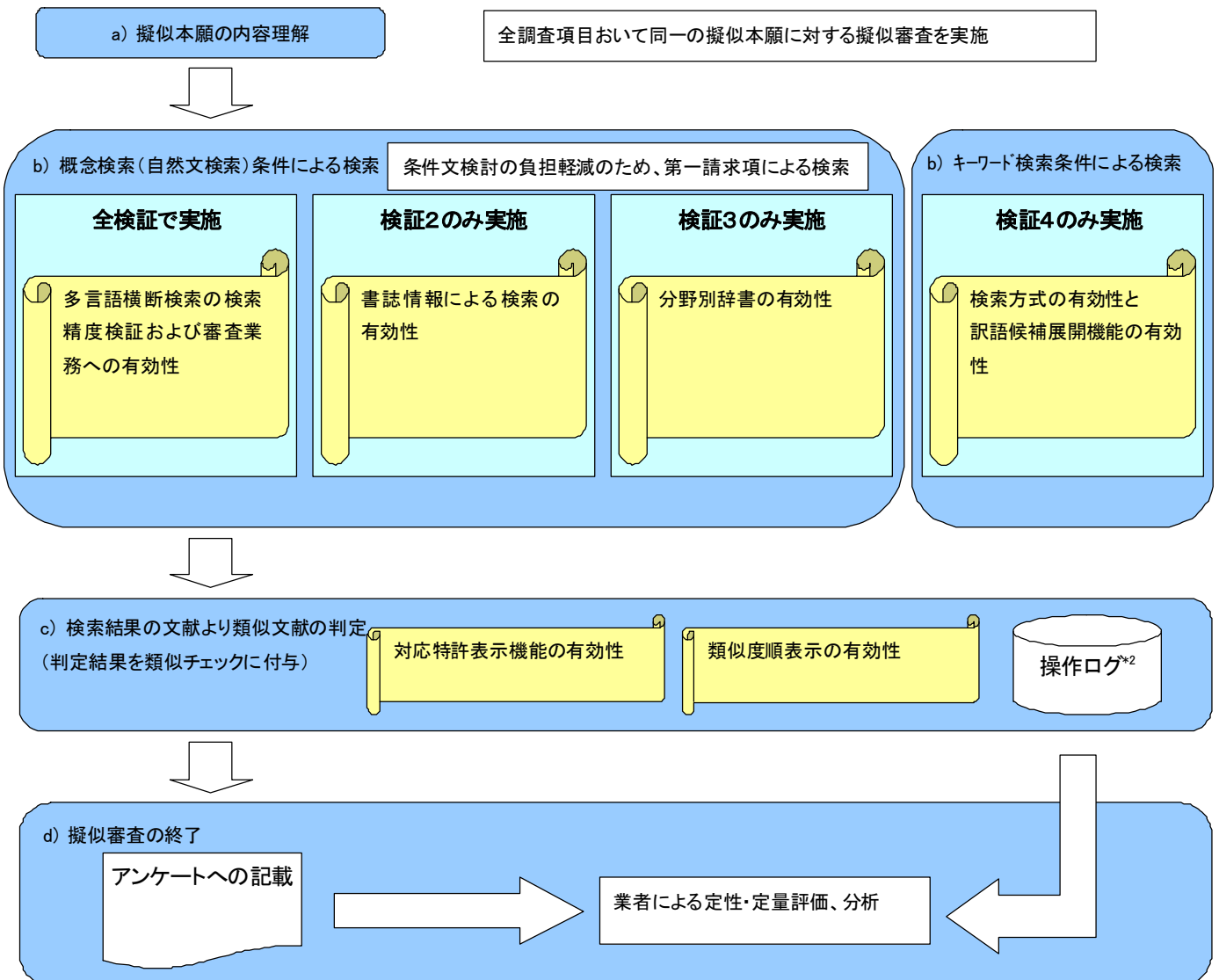
### 7.2.1 ユーザ検証

#### (1) 擬似審査検証

モデル検証システムを用いて、擬似本願に対し実体審査を行う。

擬似審査検証の流れを以下に示す。

- a) 擬似本願の内容を理解
- b) モデル検証システムの利用
- c) 類似文献の判定（負担軽減のため、上位指定件数<sup>1</sup>までの判定を実施）
- d) アンケートへの記載



\*2 : ユーザに操作していただいた内容のログ (類似文献の順位情報等)

図 7.2-1 ユーザ検証実施手順

<sup>1</sup> ユーザには、上位 30 件程度を目安として、類似判断を行って頂いた。

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-3

## (2) フリーオペレーション検証

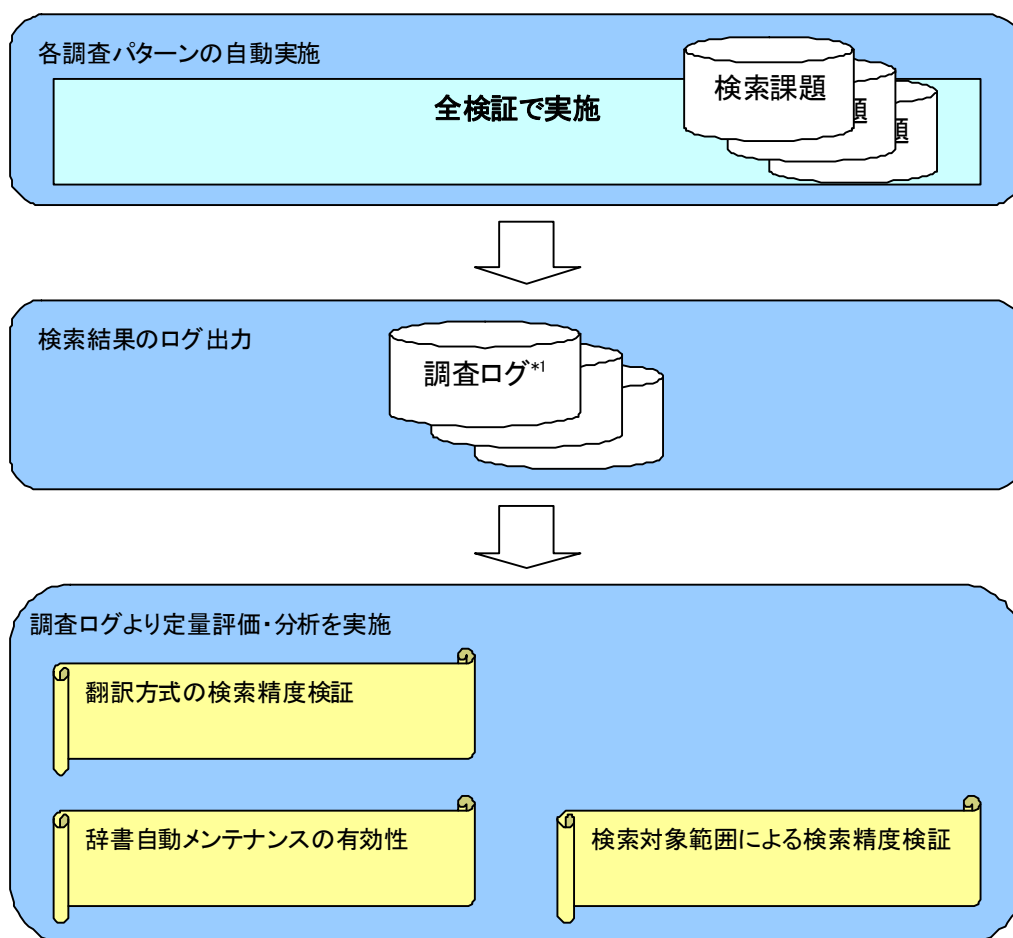
検証期間中、自由にモデル検証システムを使用し、使用感などの定性評価(アンケート記載)を行う。対象者は、全審査官を対象とし、実施については任意とする。

但し、「(1) 擬似審査検証」と異なり操作ログによる評価、分析等を行わない。アンケートの記載結果については、机上分析を実施する。

## 7.2.2 業者検証

## (1) 検索課題実施検証

各調査パターンの自動実施により、定量的に検証する調査ログの収集し、評価・分析する。また、操作ログ、モデル検証データ作成時における蓄積ログより机上分析を実施する。



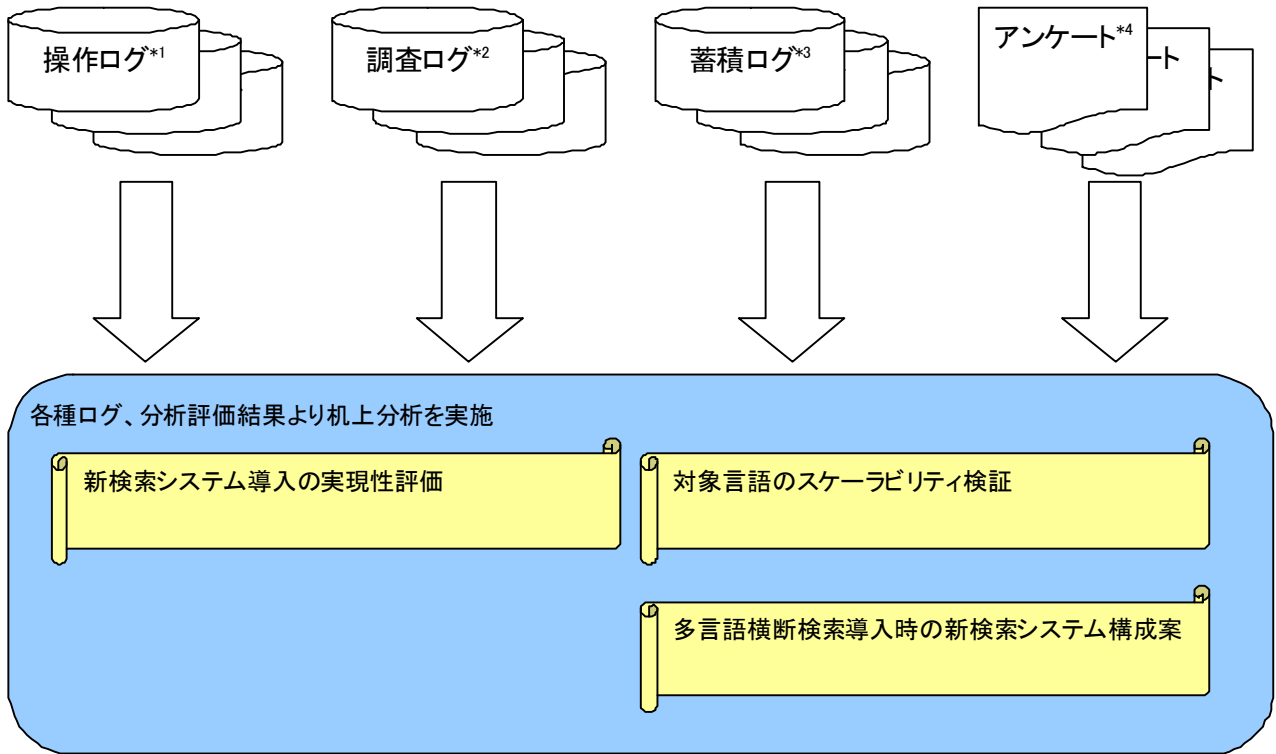
\*1：業者が調査した内容のログ（正解の出現順位等）

図 7.2-2 業者検証実施手順 (1/2)

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-4

(2) 総合分析

ユーザ検証による検証結果(操作ログ)、業者検証による検証結果(調査ログ)、モデル検証データ作成時の蓄積ログ、ユーザによるアンケート結果を踏まえ、システムの実現性について分析する。



\*1 : 「(1) 1 擬似審査検証」にてユーザが操作した内容のログ

\*2 : 「(2) 1 検索課題実施検証」にて業者が調査した内容のログ

\*3 : モデル検証データの蓄積時に採取するログ (翻訳性能等)

\*4 : 「(1) 1 擬似審査検証」「(1) 2 フリーオペレーション検証」にてユーザに記載していただくアンケート

図 7.2-3 業者検証実施手順 (2/2)

## 7.3 検索課題

### 7.3.1 ユーザ検証の検索課題

ユーザ検証では、擬似的な実体審査を行うために、任意の擬似本願(JP 公報)を検索課題として、中国・韓国文献内の類似文献を検索して検証する。その対象分野と検索課題は、以下の通りである。

#### (1) 検索課題の対象分野

ユーザ検証に用いる検索課題の対象分野は以下の通りである。分野は、①中国、韓国国内での出願が多い分野、②日本から中国、韓国への出願件数が多い分野を優先的に選択した。

①の「中国、韓国国内で出願件数が多い分野」は、それぞれの国の主要な技術分野であると考えられる。また、②の「日本から中国、韓国への出願件数が多い分野」は、日本企業が中国、韓国を意識する分野であると考えられる。

このため、以上の 2 点を兼ねる分野が、中国文献、韓国文献に対する先行技術調査のニーズが高い分野であると判断して分野選定を行った。対象分野は『表 7.3-1 検索課題の対象分野』の通り。

表 7.3-1 検索課題の対象分野

項番	言語	分野	(A)	(B)	タイトル
1	中国	A61K	1	586	医薬用、歯科用又は化粧品用製剤
2		C07C	—	417	非環式化合物または炭素環式化合物
3		C07D	10	443	複素環式化合物
4		H04L	2	1,110	デジタル情報の伝送
5		H01L	4	3,775	半導体装置,他に属さない電氣的固体装置
6	韓国	A61K	5	492	医薬用、歯科用又は化粧品用製剤
7		H01L	4	1,365	半導体装置,他に属さない電氣的固体装置
8		H04N	1	446	画像通信
9		H04B	1	329	伝送
10		G06F	2	405	電氣的デジタルデータ処理

(A) 中韓国内での出願件数順位(2007年)

(B) 中韓国における、出願人が日本国籍の公開文献数

#### (2) 検索課題抽出の条件

上記の各分野にて実際に使用する検索課題の抽出は、以下の条件で行った。なお、ユーザには通常業務の合間をぬって検証に参加して頂くため、ユーザ負担を軽減してスムーズに検証を実施して頂くことを優先した。その抽出条件は『表 7.3-2 検索課題抽出の条件』の通りである。

表 7.3-2 検索課題抽出の条件

項番	条件	理由
1	上記分野に該当する審査官 10 名をユーザとして選抜。	対象分野の専門家に実施して頂くことで、より深い考察ができるため。
2	各ユーザの審査実績のある案件で、かつ対象分野(IPC)に該当する本願を抽出。	審査実績のある本願を利用することで、本願理解の時間を短縮するため。
3	最新起案結果が拒絶査定の本願(具体的には、拒絶理由通知、意見拒絶査定、戻し拒絶査定)	特許査定の本願だと、正解となる類似文献が存在しない可能性があるため。
4	上記が複数ある場合は、出願年が新しい文献。	今回の検証データは、直近 1 年分の中国・韓国公報としている(詳細は『7.4 モデル検証データ』)。このため、古い本願の場合、類似文献が存在しない可能性があるため。

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-6

## (3) ユーザ検証対象者と検索課題の内訳

ユーザ検証対象者と検索課題の対象分野、本願の概要の内訳は以下の通りである。

なお、出願番号に※印が付く案件は、検証後により適切な課題にすべく、ユーザ検証 2 から変更した課題である。

表 7.3-3 ユーザ検証検索課題の内訳

項番	審査官	言語		分野	出願番号	本願の概要 【発明の名称】を抜粋
		中	韓			
1	審査官 A	○	—	A61K 9/00	特願 2003-545326 (特表 2005-514360)	ダイコンソウ(GeumJaponicumthunbvar.)の有機抽出物を含む組成物およびその使用
2		—	○		特願 2004-545003 (特表 2006-506370)	アッケンソウ抽出物を含むマイクロイド系白血病治療及び免疫増強用組成物
3	審査官 B	○	—	C07C	特願 2000-525148	アミノビグアニド化合物ならびにコンタクトレンズの消毒および医薬組成物の保存におけるその使用。
4		○	—	C07D	特願平 11-21702	1, 4-ジヒドロピリジン誘導体
5	審査官 C	—	○	G06F 12/00	特願 2005-016925	誤り訂正符号を有する固体記憶装置を構成するシステムおよび方法
6		—	○		特願 2003-521494	移動端末機におけるモデムとメモリとの間のインターフェース装置及び方法
7	審査官 D	○	—	H01L 21/00	特願 2005-053259	半導体素子製造用ウェーハ処理装置
8		○	—		特願 2005-000288	半導体製造システム及びクリーンルーム
9	審査官 E	—	○	H01L 21/00	特願 2001-564392	広範囲なワイヤを形成するためのナノスケール・パターン形成
10		—	○		特願 2001-332869	超高圧水銀ランプ及びそれを用いた半導体露光装置
11		—	○		特願 2006-272137 ※	炭素含有膜エッチング方法及びこれを利用した半導体素子の製造方法
12		—	○		特願 2008-140105 ※	半導体発光素子およびその製造方法
13	審査官 F	—	○	H04B 7/00	特願 2002-150884	通信端末及び無線通信方法
14		—	○		特願 2003-385887	通話料金節減装置及びシステム
15	審査官 G	○	—	H04L 12/00	特願平 10-247672	時分割多重伝送システム及びそれに用いるチャンネル識別方式
16		○	—		特願平 10-374226	光伝送路上の導通正常性確認方法
17		○	—		特願 2003-289965 ※	無線ネットワークのための制御方法及び制御装置
18		○	—		特願 2003-426170 ※	監視装置、基地局および無線LANシステム
19	審査官 H	○	—	H04L 12/00	特願 2004-236465	自動化機能を有する家庭用電気製品通信制御コード・トランスフォーム・モジュール
20		○	—		特願 2004-312122	HAVi規格に準拠したターゲット機器
21	審査官 I	—	○	H04N 7/00	特願 2006-281732	分割設置型情報処理システム、情報処理端末およびパーソナルコンピュータ
22		—	○		特願 2007-026845	端末監視装置とそのプログラム
23	審査官 J	—	○	H04N 7/00	特願 2002-093668	双方向通信システムにおけるチャンネル変更方法、ケーブルモデム、ケーブルモデム終端装置
24		—	○		特願 2002-357145	AV送信装置、AV受信装置、AV通信システム、AV送信プログラム及びAV受信プログラム
25		—	○		特願 2003-000399 ※	ネットワークシステムの監視方法、及びそのネットワークシステム。

### 7.3.2 業者検証の検索課題

業者検証もユーザ検証と同様であるが、その違いは、標本数を増やして統計的分析を元に、定量評価を行う点である。このため、大量の検索課題を用意し、検証を自動化して行った。

その対象分野と実際の検索課題は、以下の通りである。

#### (1) 検索課題の対象分野

ユーザ検証と同一の対象分野とした。これにより、ユーザ検証による定性評価(アンケート)を業者検証の定量評価(検索精度)によって裏付ける事が可能となる。

#### (2) 検索課題の抽出条件

業者で検索精度評価を行う場合は、一般的にも、予め正解データを用意しておき、それが検索結果として表れるか否かの検証が必要である。

このため、今回の検証では正解データをファミリー文献(中国・韓国公報)として、それに対応するJP公報を検索課題とした。

本来、検索の精度評価の正解データとしては、各検索クエリに対して、正解であると人が判断した文献を用いるべきである。例えば、特許であれば、請求項を検索クエリとし、引例となっている特許を正解文献とするなど、既に存在するデータの有効に利用すべきである。

しかし、中国特許や韓国特許では、引例情報が得られないということであった。また、このような正解データを人手で新たに作成するには、多大なコストが発生する。

そこで、今回は、国内優先あるいは国際出願によって日本から中国あるいは韓国に出願された特許に関して、元の日本特許を検索クエリ(実際には特許全文ではなく、第一請求項を検索クエリとする)とし、その元の日本特許に対応する中国あるいは韓国の特許を擬似正解文献とすることとした。

#### (3) 対象分野と検索課題の内訳

今回の検証で使用する中国特許、韓国特許のデータの中から、今回の調査対象とした各IPCの件数比率に対応して、対応する日本特許とともにそれぞれ2,000件を抽出した。

この擬似正解文献に基づく正解データについてIPC分類ごとの件数を『表 7.3-4 業者検証課題の内訳』に示す。なお、日本特許に対応特許がある特許は、例えば、中国特許については6,883件存在している。そこで、2,000件の特許を選定するにあたっては、対応する日本特許の出願番号の最近のものから優先して抽出するものとした。

表 7.3-4 業者検証課題の内訳

項番	国	分野	件数	項番	国	分野	件数
1	中国	A61K	196	6	韓国	A61K	133
2		C07C	139	7		G06F	275
3		C07D	142	8		H01L	1,113
4		H01L	1,198	9		H04B	187
5		H04L	325	10		H04N	293
		合計	2,000			合計	2,001

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-8

## 7.4 モデル検証データ

## 7.4.1 データ諸元

モデル検証で使用した、中国文献および韓国文献のデータ諸元を以下に示す。

表 7.4-1 データ諸元

項番	言語	データ種別	対象年	データ量	入手元
1	中国	中国特許公開公報 中国特許登録公報	2007年発行分 (IPC:ACGH*1) 2006年発行分 (IPC:BDEF*1)	キーワード翻訳方式 約 26 万件*2	株式会社 発明通信社 より借用
				コンテンツ翻訳方式 約 6 万件*3	
2	韓国	韓国特許公開公報 韓国特許登録公報	2007年発行分*4	キーワード翻訳方式 約 22 万件*2	特許庁、独立行政法人 工業所有権・情報館*4 より借用
				コンテンツ翻訳方式 約 22 万件	

\*1: 中国で出願傾向の多いIPC主要セクション(A、C、G、H)は2007年発行のデータを、それ以外のセクション(B、D、E、F、)は2006年発行のデータとした。

\*2: キーワード翻訳方式のデータは、中国文献、韓国文献ともに対象年の1年分のデータとした。

\*3: 検索課題のIPCと同じIPCの文献を蓄積した。

チューニング期間は、1週間程度である。中国文献を全文献に対し翻訳蓄積を実施する場合、以下の通り、1ヶ月程度の時間を要する。

このため、コンテンツ翻訳方式のデータは主分類と副分類に検索課題と同じIPCの文献のみ(6万件程度)とした。

但し、6万件を翻訳蓄積する場合、以下の通り7日程度要するため、初期環境構築時とチューニング3でのみ翻訳蓄積を実施した。(チューニング1回当たりの蓄積時間は2、3日を想定)

< 中国文献の蓄積時間: 全文 >

$26 \text{ 万件 (データ量)} \times 100 \text{ 秒 (翻訳時間/1 文献)} \div 10 \text{ (マシン 10 台での並列処理)} = \text{約 1 か月}$

< 中国文献の蓄積時間: 5 万件 (主分類、副分類に検索課題のIPCが含まれている文献) >

$6 \text{ 万件 (データ量)} \times 100 \text{ 秒 (翻訳時間/1 文献)} \div 10 \text{ (マシン 10 台での並列処理)} = \text{約 7 日}$

\*4: 以下、INPITと記載(National Center for Industrial Property Information and Training)。

## 7.4.2 準備方法

## (1) キーワード翻訳方式用データベースの作成

## 1) 中国語文献用データベース

今回入手した中国語文献は、書誌と全文 XML が別ファイルで提供されていた。このため、以下の4つのステップにて実施した。

## 1. 全文 XML データへの書誌事項の追加。

まず全文データ中に書誌データを追加する作業を行った。

## 2. 文字コードを UTF-8 に変換。

次に、中国語 XML ファイルの文字コードを UTF-8 にコード変換実施した。

## 3. XML データへの妥当性検証を実施

コード変換後、XML データの型や桁数などが正常であることを確認するために、妥当性検証<sup>2</sup>を実施。

## 4. XML データベースへ登録

最後に妥当性検証で正常と判断された XML ファイルのみ、XML データベースに登録を行った。

表 7.4-2 中国語文献入手データ

項番	データ種別	形式	データ単位	文字コード
1	書誌データ	テキスト	IPC セクション単位(1ファイル中に複数文献)	GB2312
2	全文データ	XML	文献単位(1ファイル中に1文献)	GB2312 または GBK

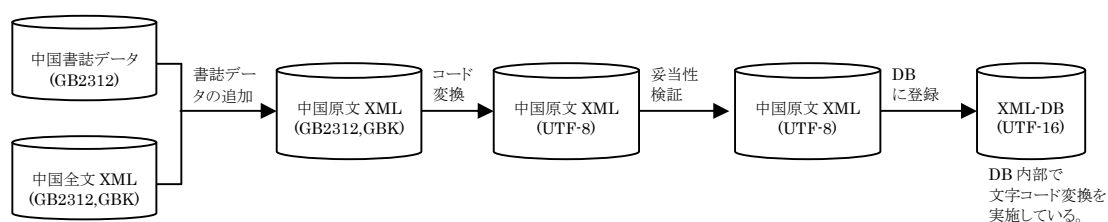


図 7.4-1 中国原文文献の登録処理

<sup>2</sup> 妥当性検証には、Apache XML Project の sax.Counter を使用した。

## 2) 韓国語文献用データベース

韓国文献は、1 文献の全データが 1 ファイルで提供されていたためそのまま使用した。

## 1. 文字コードを UTF-8 に変換

まず、韓国語 XML ファイルの文字コードを UTF-8 にコード変換実施。

## 2. XML データへの妥当性検証を実施

次に、XML データの型や桁数などが正常であることを確認するために、妥当性検証を実施。

## 3. XML データベースへ登録

最後に妥当性検証で正常と判断された XML ファイルのみ、XML データベースに登録を行った。

表 7.4-3 韓国原文文献入手データ

項番	データ種別	形式	データ単位	文字コード
1	書誌+全文	XML	文献単位(1 ファイル中に 1 文献)	EUC-KR

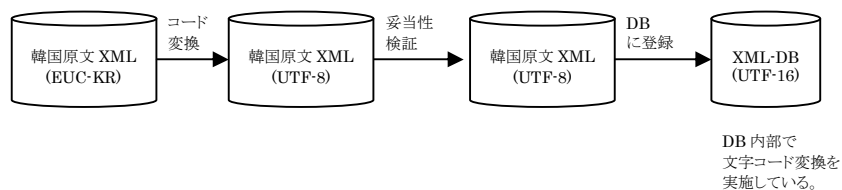


図 7.4-2 韓国原文文献の登録処理

## (2) コンテンツ翻訳方式用データベースの作成

## 1) 中国日本語翻訳文献用データベース

キーワード翻訳方式の場合と同様に、以下の 4 ステップで実施した。なお、キーワード翻訳方式では、原文を UTF-8 に文字コード変換した後に登録していたが、コンテンツ翻訳方式の場合、翻訳結果が Shift-JIS のため、そのままの文字コードでデータベースへの登録が可能となる。

## 1. 全文 XML データへの書誌事項の追加。

まず全文データ中に書誌データを追加する作業を実施。

## 2. 全文 XML ファイルの翻訳実施

次に、作成した中国語 XML ファイルに対して、日本語翻訳を実施。

文献の IPC 分類(主分類および副分類)に対応する分野別専門辞書を使用し翻訳を行った。また、文献内に複数の対象となる IPC(主・副)が存在する場合は、複数辞書指定により翻訳を行った。IPC と分野別辞書の対応は、『表 7.4-4 中国文献翻訳辞書 IPC 対応表』の通りである。

表 7.4-4 中国文献翻訳辞書 IPC 対応表

セクション	サブクラス	分野別専門辞書	
		初期構築	検証 4
A	A61K	化学	化学
C	C07C	化学	C07(自動メンテ辞書)
	C07D	化学	C07(自動メンテ辞書)
	その他	化学	化学
G	G06F	コンピュータ・電子	コンピュータ・電子
H	H01L	電子	H01L(自動メンテ辞書)
	H04B	コンピュータ・電子	コンピュータ・電子
	H04L	コンピュータ・電子	コンピュータ・電子
	H04N	コンピュータ・電子	コンピュータ・電子
	その他	コンピュータ・電子	コンピュータ・電子

## 3. XML データへの妥当性検証を実施

日本語翻訳 XML ファイルに対して妥当性検証を実行し、データの型や桁数などがあっていることを確認。

## 4. XML データベースへ登録

最後に妥当性検証で正常と判断された XML ファイルのみ、XML データベースに登録を行った。

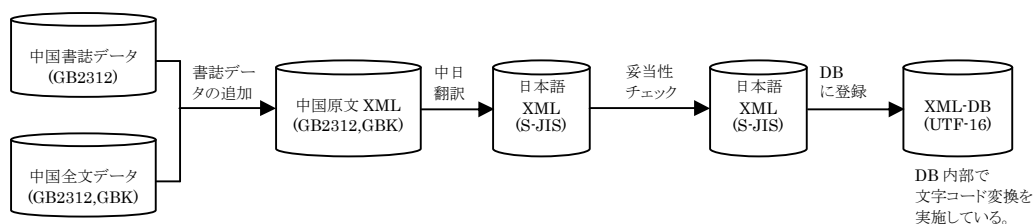


図 7.4-3 中国日本語翻訳文献の登録処理

## 2) 韓国日本語翻訳文献用データベース

中国文献のキーワード翻訳方式の場合と同様に、以下の3ステップで実施した。キーワード翻訳方式との違いは、中国日本語翻訳文献用データベースと同様である。

## 1. 全文 XML ファイルの翻訳実施

最初に、韓国語 XML ファイルに対して、日本語翻訳を実施。

初期構築時は基本辞書のみ、検証 4 では基本辞書＋ユーザ登録辞書を用いて全文の翻訳処理を行った。

## 2. XML データへの妥当性検証を実施

日本語翻訳 XML ファイルに対して妥当性検証を実行し、データの型や桁数などがあっていることを確認。

## 3. XML データベースへ登録

最後に妥当性検証で正常と判断された日本語翻訳 XML ファイルのみ、XML データベースに登録を行った。

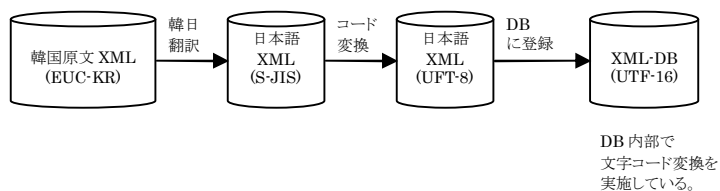


図 7.4-4 韓国日本語翻訳文献の登録処理

作成文書	多言語横断検索技術に関する 次世代検索システム開発に向けた調査	Rev.	頁
調査報告書		1.0	7-13

## 7.5 評価方法

### 7.5.1 アンケート

各ユーザ検証期間の検証終了後に、検証実施ユーザによるアンケート記載を行う。  
アンケート結果は、定性評価として分析を行う。

### 7.5.2 評価指標

本調査では、検索精度の評価指標として、MAP、MRR を用いて評価・分析を行う。

MRR は、正解が1つの場合に適した評価指標であり、1つのファミリー文献を正解とする業者検証結果(調査ログ)の評価に用いる。

MAP は、正解が複数の場合に適した評価指標であり、複数の類似文献を正解とするユーザ検証結果(操作ログ)の評価に用いる。また、ユーザ検証では検索課題数が少ないため、MAP については参考値として使用する。

2つの検索結果(①正解が1位と100位、②正解が1位と6位)についてのMAP(AP)、MRR(RR)の違いは、『表 7.5-1 MRRとMAPの違い』の通りである。

表 7.5-1 MRRとMAPの違い

MRR(検索課題が1件のため RR) … ①と②は同じ検索精度	
①正解が1位と100位に検索された場合 RR = 1 / 1 = 1.00	②正解が1位と6位に検索された場合 RR = 1 / 1 = 1.00
MAP(検索課題が1件のため AP) … ②の方が検索精度は良い	
①正解が1位と100位に検索された場合 AP = 1 / 2 (1 / 1 + 2 / 100) = 0.51	②正解が1位と6位に検索された場合 AP = 1 / 2 (1 / 1 + 2 / 6) = 0.67

## 7.5.3 MRR(Mean Reciprocal Rank)

MRR についての詳細は、『表 7.5-2 MRR 説明』の通りである。

表 7.5-2 MRR 説明

項番	項目	詳細
1	説明	1つの正解について正解が出現した順位を評価する指標である。 最小値を0、最大値を1とし、数値が大きい程、良い検索精度といえる。 検索課題毎に正解が最初に出現した順位の逆数を求め(RR(Reciprocal Rank))、 全検索課題のRRを平均すること(MRR)で、システムの検索精度を評価する。
2	計算式	<p>r : 正解が出現した順位</p> $RRi = \frac{1}{r}$ $MRR = \frac{1}{N} \sum_i^N RRi$ <p>例 1)RR 正解出現順位が5位の場合、 <math>RR = 1 / 5 = 0.2</math></p> <p>例 2)MRR 検索課題数が2件、それぞれのRRが、0.2、0.33の場合、 <math>MRR = 1 / 2 \times (0.2 + 0.33) = 0.265</math></p>

## 7.5.4 MAP(Mean Average Precision)

MAP についての詳細は、『表 7.5-3 MAP 説明』の通りである。

表 7.5-3 MAP 説明

項番	項目	詳細
1	説明	<p>検索結果のゴミの少なさとモレの少なさを評価する指標である。            最小値を 0、最大値を 1 とし、数値が大きい程、良い検索精度といえる。            検索課題の正解毎、検索された正解数を順位で割ることで、ゴミの少なさを評価、検索された全ての正解の精度を足し、最後に、全正解数で割ることで、モレの少なさを評価する(AP(Average Precision))。全検索課題の AP を平均すること(MAP)で、システムの検索精度を評価する。</p> <p>※)ユーザ検証において、検索結果の上位 30 文献内で、類似判断がされていない文献については、非類似文献(不正解)として計算する。</p>
2	計算式	<p>R<sub>i</sub> : 課題 i における全正解数            I(r) : r 位の文書が正解か否か(1or0)            R : 順位            count(r) : 検索された正解数            N : 検索課題数</p> $AP_i = \frac{1}{R_i} \sum_r I(r) \frac{\text{count}(r)}{r}$ $MAP = \frac{1}{N} \sum_i AP_i$ <p>例 1)AP            全正解数が 3 件、正解出現順位が 2 位、4 位、ランキング外の場合、  <math>AP = 1/3 \times (1/2 + 2/4 + 0) = 0.244</math></p> <p>例 2)MAP            検索課題数が 2 件、それぞれの AP が、0.244、0.33 の場合、  <math>MAP = 1/2 \times (0.244 + 0.33) = 0.287</math></p>