

10 考察と課題のまとめ

本調査の全体的な考察や課題について記述する。

10.1 考察のまとめ

10.2 結果と考察における課題と解決策

10.3 本検証方法における課題と解決策

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	10-1

10.1 考察のまとめ

10.1.1 結果

本検証を通じて判明した、主要な結果を『表 10.1-1 考察のまとめ』に示す。

表 10.1-1 考察のまとめ

項番	カテゴリ	概要
1	翻訳機能	韓国は実用可能(日本語訳も読んで理解できるレベル)。
2		中国は審査利用のレベルに達していない。今後、精度向上が必要。
3		大幅な辞書整備が必要(特に中国)。
4	検索機能	コンテンツ翻訳方式の評価が高い。
5		多言語への概念検索の導入はまだ早い。 まずは、現行クラスタ検索の踏襲が必要。
6	総論	韓国は実用可能。 中国は注力分野を絞って、完成度の高いシステム提供を目指すべき。
7		初期の多言語横断検索は、ローカル文献に対して利用するのが良い。

10.1.2 考察

(1) 翻訳機能

- 韓国は実用可能(日本語訳も読んで理解できるレベル)。
- 中国は審査利用のレベルに達していない。今後、精度向上が必要。

韓国語の翻訳機能は使えるレベルであるが、中国語の翻訳機能は審査に使えるレベルに達していないことが判明した。これは、中国語翻訳の難しさに強く関連している(詳細は『9.2.2(1)翻訳精度』を参照)。

韓国語は日本語と文法構造が似ているため、ある種、単純な置換作業で翻訳可能な言語である。対して、中国語は、日本語と文法的な構造が異なっていることにより、翻訳の難易度が高いという実情がある。(これは英語翻訳でも同様)。

しかし、日本語と英語では業界全体として 20~30 年近くの歳月を掛けて育てた結果、実用に耐えるレベルにまで改善が進んでいる。一方、中国語翻訳は近年 10 年程度で活発した技術であるため、まだ歴史が浅いという実情もあり、十分に翻訳精度の向上を実現するに至っていない面がある。辞書の登録語数の違いを見ても、その差は明らかである。

表 10.1-2 中国翻訳と英語翻訳の辞書語数(The 翻訳)

辞書種別	中日翻訳	英日翻訳	備考
基本語	25.0 万語	103 万語	中国は、英語の約 1/4
専門用語	23.8 万語	272 万語	中国は、英語の約 1/10
自動メンテナンス	1300 語	—	

※上記数値は、The 翻訳の中日(モデル検証で使用した試用版)と、英日の登録語数。

英語翻訳が、過去から現在に至るまでに、着実に精度向上を図ってきたように、中国語翻訳も、英語翻訳と同様の開発や改善のステップを踏む事で、精度向上が見込まれると考えられる。

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	10-2

- 大幅な辞書整備が必要(特に中国)。

辞書は正しい対訳を提示するだけでなく、中国語で難しい語句の切れ目を判断する際や構文の解析時にも利用するため、精度向上には必須である。現在の英語翻訳エンジンでは、100万語以上が当たり前となっているため、同等レベルの語数が必要な可能性は高い。但し、辞書構築は、人件費コストが掛る作業であるため、長期的な視点が必要と考える。

モデル検証にて確認した「辞書自動メンテナンス機能」を用いて「100万語」レベルの辞書を構築するには、50人体制で1年近くの歳月がかかる(詳細は『9.2.2(3)5a)作業工数』を参照)。

$$100 \text{ [万語]} \div 13.2 \text{ [語/時間・人]} = 75757.6 \text{ [時間・人]}$$

$$75757.6 \text{ [時間・人]} \div (7 \text{ [時間/日]} \times 20 \text{ [日/月]}) = 541 \text{ [人月]}$$

また、各製品ベンダが辞書整備することを待つという選択肢もある。しかし、特許庁で必要となる辞書分野に絞ってベンダが整備することは、ベンダ自身の費用対効果の面からも難しいため、各省庁あるいは、業界団体全体が連携していく必要があると考えられる。

以上の理由から、長期的な視点で、辞書整備のステップ毎に、方式案を組み合わせる辞書を作成することが望ましいと考える。以下に組み合わせの案を示す。(詳細は、『9.2.2(2)辞書』を参照。)

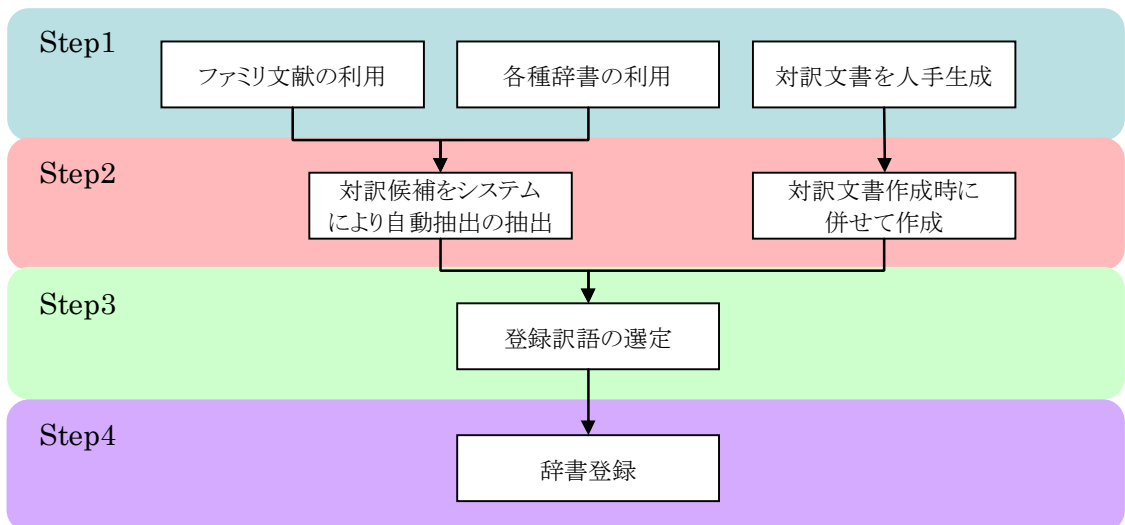


図 10.1-1 辞書整備の組み合わせ案

(2) 検索機能

- コンテンツ翻訳方式の評価が高い。

今回の検証では、コンテンツ翻訳方式の方がやや利があった。しかし、翻訳精度向上の結果を即時反映できるという面で、キーワード翻訳方式が優位な点もある。

これについては、メリット・デメリットを整理したうえで、今後の要件定義フェーズにて選定が必要と考えている。

<コンテンツ翻訳の主なメリット>

- ✓ 日本語用の検索エンジンが使えるため、精度チューニングしやすい。
 予め翻訳しているため、JP 公報と同じ扱いが可能。しかし、キーワード翻訳では検索クエリとデータベース、両方を原語のまま扱う為、中韓対応の検索エンジン導入が必要。
- ✓ 現行同様のスクリーニング性能が確保できる。
 予め翻訳しているため、JP 公報と同様の性能を担保。キーワード翻訳方式の場合、翻訳処理が常に入るとため、遅くなる(1 文献当たり、中国=33 秒、韓国=2.8 秒)。

<コンテンツ翻訳の主なデメリット>

- ✓ 誤訳を直すには再蓄積が必要 (但し、サーバ台数を増やすことで回避可能)。
 今回の検証の基礎数値を元に概算算出すると、全文献を翻訳するのに、以下のサーバ台数ならば、約 1 か月程度で可能と考えられる。

表 10.1-3 コンテンツ翻訳方式に対する再翻訳予測(サマリ)

言語	ローカル 文献数	再翻訳 期間	想定 CPU スペック	必要台数 (概算)	備考
中国	約 200 万件	1ヶ月	Intel Core2 Duo 3GHz 相当×2CPU	6～7 台	4 多重処理を活用
韓国	約 150 万件			1～2 台	

[見積前提]

ファミリが存在する文献は、英文抄録や WPI でサーチできるため、アクセス手段の無いローカル文献に絞って蓄積することを前提としている。

- 多言語への概念検索の導入はまだ早い。まずは、現行クラスタ検索の踏襲が必要。

概念検索は、全文検索に比べると、ブラックボックスな面が多いのが実情である。加えて、現在の翻訳精度が要求レベルに達していない為、概念検索用のチューニングをしても、全体的な効果は薄いと考えている。

また、審査利用するためには、現行機能からデグレードすることは許されない。このため、まずは現行クラスタ検索の機能踏襲を最優先に、今後要件定義が必要である。

なお、多言語横断検索における、現行機能の踏襲可否の状況は、『9.2.5(5)1 現行システム機能の踏襲』を参照。

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	10-4

(3) 総論

- 韓国は実用可能。中国は注力分野を絞って、完成度の高いシステム提供を目指すべき。

英文抄録や WPI は「抄録」のため、実施例レベルの検索をすることができない。このため、多言語横断検索システムというものは、今後必要不可欠なツールと考えられる。

しかし、平成 26 年の検索系最適化までに、全分野にわたって中国語翻訳の精度を高めることは非常に難しいと思われる。これは、全分野の辞書を整備することが困難なためである。

このため、中国の技術優位性が高い「伝統的医薬品 (A61K 等)」など、分野を絞って辞書整備を行い、特定分野ではあるが、完成度の高いシステムを提供することが重要と考えている。(中国・韓国の技術優位性が高い分野は、『図 1.1 3 中国文献の IPC 分布グラフ』『図 1.1 4 韓国文献の IPC 分布グラフ』を参照。)

- 初期の多言語横断検索は、ローカル文献に対して利用するのが良い。

多言語横断検索は、WPI や英抄と異なり、全文あるいは要約、実施例等の構成部を対象とした検索が可能であることが大きなメリットである。ただし、全文を翻訳・蓄積するには、大幅に時間がかかる(特にコンテンツ翻訳方式)。

また、ファミリー文献が存在する文献は、英文抄録や WPI などを活用することで、精度の良い人手翻訳(英語)の結果でサーチすることができるが、ファミリー文献が存在しないローカル文献はサーチする手段がないため、多言語横断検索技術がますます必要である。

このため、初期の段階では、ローカル文献のみを対象とし、ファミリー文献が存在する文献は、段階的に蓄積し、検索可能とすることも検討すべきである。

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	10-5

10.2 結果と考察における課題と解決策

検証 1 から検証 4 を実施する中で発見された課題のサマ리를『表 10.2-1 結果と考察における課題一覧』に示す。

表 10.2-1 結果と考察における課題一覧

項番	カテゴリ	課題
1	文献特性	① 実施例の検索ができない。
2	翻訳方式	① 検索エンジンの中韓対応が実用レベルにない。
3	翻訳技術	① 中国文献の翻訳精度が低い。 ② 自動辞書メンテナンスでの登録語数が意外と少ない。 ③ 専門用語の分野の切り分け方が粗い。 ④ 出願人名で検索できない。 ⑤ 異表記をサポートできない。
4	検索技術	① 概念検索はブラックボックス。 ② 訳語候補展開で、訳語候補の日本語訳が同じ場合、適切か否かを判断できない。
5	審査関連情報の活用	① IPC の付与基準が日本と異なる場合がある。
6	翻訳精度・検索精度を高める工夫	① 現行検索機能は全て実現できるか？ ② 文献照会のスピードが担保される必要がある。 ③ 動作を検索できない。 ④ ローカルで簡単に修正できる辞書を用意する。 ⑤ 図面と文献は同一画面で表示すべき。 ⑥ 原文と日本語訳は同一画面で表示すべき。 ⑦ 検索対象国での審査状況や引用状況を表示したい。
7	蓄積性能	① コンテンツ翻訳方式は、蓄積や再蓄積に時間がかかる。
8	その他	① 翻訳文献の証拠性 ② UNICODE では、日中韓の似ている漢字が同一コードに割り当てられていることがある。

検証 1 から検証 4 を実施する中で発見された課題と解決策を以下に示す。

表 10.2-2結果と考察における課題と解決策(1/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
1	文献特性							
1-1	実施例の検索ができない。 → 『9.1.2 中韓の文献特性』	原文のXMLデータに、実施例のタグが公報XML規約(DTD)には存在しているが、 ①韓国:存在しているがほとんど使われていない。 ②中国:今回のテストデータ(2007年公報)上には存在しない。		●	◎	SIPO、KIPO に、XML データに実施例のタグを入れるよう働きかける。	△	長
2	翻訳方式							
2-1	検索エンジンの中韓対応が実用レベルでない。 → 『9.2.1 翻訳方式』	①日本のベンダは、基本的に日本語対応。外国語は英語対応が限度。 ②外国ベンダ製品を採用する場合には、チューニングを研究者レベルの人が行う必要がある。		●	◎	海外ベンダの研究者も含めた、システム導入/構築が必要。 ※コンテンツ翻訳方式であれば、本課題は発生しない。	△	長
3	翻訳技術							
3-1	中国文献の翻訳精度が低い。 → 『9.2.2(1)翻訳精度』	①中国語は構文解析が難しい ②外来語を漢字表記するため、異表記が多い。 ③翻訳エンジンとしての辞書の登録語数が、現在の語数では足りない	●		◎	形態素解析及び構文解析の技術を継続して研究する。	△	長
3-2			●		◎	辞書の登録語数を増加する。	○	長
3-3	辞書自動メンテナンスでの登録語数が意外と少ない。	適否判断を実施した人数が少なかった	●		◎	大規模な人的リソースにて対応する。	○	長
3-4	→ 『9.2.2(3)辞書自動メンテナンス』	適否の判断を、該当分野の専門家でないメンバが行ったため、時間がかかった。	●		◎	その分野の専門家にて適否判断を実施する。 ※但し、適切な要員の確保、専門家ゆえのコスト増を抑える対策の検討が今後必要となる。	○	長

凡例:

翻訳技術/検索技術 ●・・・該当、空欄・・・該当しない 業務上の重要度 ◎・・・全分野に影響、○・・・一部分野あるいは一部文献に影響、－・・・業務への影響は無し
実現難易度 ○・・・実現が可能、△・・・実現が難しい 時系 短・・・短期対応、長・・・長期対応

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	10-7

表 10.2-3結果と考察における課題と解決策(2/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
3-5	専門用語辞書の分野の切り分け方が粗い。 → 『9.2.2(3)辞書自動メンテナンス』	適切な粒度で分けないと、特定分野の翻訳精度が低下する虞がある。	●		◎	FI ハンドブックの関連分野を参考にする等、適切な切り分け方を検討する。	○	長
3-6	出願人名で検索できない。 → 『9.2.2(2)辞書』	固有名詞は、辞書に登録されていないとうまく翻訳できないため、必要。	●	●	◎	出願件数上位 1 万件程度の出願人の固有名詞辞書を用意する。	○	短
3-7	異表記をサポートできない。 → 『9.2.2(2)5c)異表記・統制語の活用』	特に、中国語の異表記は、無限に存在する可能性がある。全てを把握することが難しいため必要。	●	●	◎	<異表記を減らす:統制語> IPCの代表的な語等は、審査官を交えてそれぞれの言語に対する統一的な語を検討する。	△	長
3-8			●	●	◎	<既知の異表記展開:異表記テーブル> 検索時に関連語を抽出するためのテーブルを用意する。	○	長

凡例:

翻訳技術/検索技術 ●・・・該当、空欄・・・該当しない 業務上の重要度 ◎・・・全分野に影響、○・・・一部分野あるいは一部文献に影響、－・・・業務への影響は無し
 実現難易度 ○・・・実現が可能、△・・・実現が難しい 時系 短・・・短期対応、長・・・長期対応

表 10.2-4結果と考察における課題と解決策(3/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
4	検索技術							
4-1	概念検索はブラックボックスで、何を検索したのかわからない。 →『9.2.3(2)概念検索』	再現性が保てないと、特許審査に適さないため。		●	◎	<検索条件のホワイトボックス化> 概念検索によって拡張された、実際に検索に使用する語を画面上に表示し、追加・削除ができるようにする。	○	短
4-2	訳語候補展開において、訳語候補の日本語訳が同じ場合、適切か否かを判断できない。 例：トラックの中国語訳 ① 卡车(truck: 自動車のトラック) ② 跑道(track: 競技場のトラック) ③ 磁道(track: 磁気テープのトラック) →『9.2.3(3)訳語候補展開』	モデル検証での日本語訳表示では、同じ訳が表示された場合に、違いが判断できないため、意味合いの違いが瞬時にわかるような工夫が必要。	●	●	◎	理論的には、文例集を用意して提示することで判断可能。 但し、全ての辞書語に文例を登録することは、作業コスト面からも実現が困難。	△	長
5	審査関連情報の活用							
5-1	IPC の付与基準が外国と日本で異なり、ノイズや漏れが生じる恐れがある。 →『9.2.4(1)書誌情報(IPC)』	特に、新しい分野は付与基準が各国で異なっている可能性が高い。また、過去分が IPC8 以前の場合、付与基準のズレは更に大きいと推測される。		●	○	細かいレベルの IPC で、当該分野の隣接分野を OR 条件で検索することで、IPC の揺らぎを吸収する。	△	短
5-2						MCD に蓄積された分類データを活用する。	○	短

凡例:

翻訳技術／検索技術 ●・・・該当、空欄・・・該当しない 業務上の重要度 ◎・・・全分野に影響、○・・・一部分野あるいは一部文献に影響、－・・・業務への影響は無し
実現難易度 ○・・・実現が可能、△・・・実現が難しい 時系 短・・・短期対応、長・・・長期対応

表 10.2-5結果と考察における課題と解決策(4/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
6	翻訳精度・検索精度を高めるための工夫							
6-1	国内文献の現行検索機能は全て実現できるか？ → 『9.2.5(5)1)現行システム機能の踏襲』	—				<p><コンテンツ翻訳方式> 以下は、システム的には実現可能だが、国内文献検索と同等の効果は得られない可能性あり。 ①近傍検索(語順指定あり) DB 上の翻訳精度が、日本語の語順と一致していないと、検索されない虞がある。</p> <p><キーワード翻訳方式> 以下は、システム的に不可能。 ①スクリーニング性能 スクリーニング時に翻訳処理が入るため。</p> <p>以下は、国内文献検索と同等の効果を得られない可能性あり。 ②全文検索 検索ワードの翻訳時に、訳語の揺れがある。 但し、訳語候補展開機能で防止可能。 ③ヒットワード反転 日中辞書と中日辞書の内容は同一でないため、日→中→日を行うと、1対N対Mになり、検索時ワードと関連の無い語まで語まで反転する虞あり。 ④近傍検索(語順指定あり) DB 上の翻訳精度が、日本語の語順と一致していないと、検索されない虞がある。</p>	○	短
			●	◎			△	長

凡例:

翻訳技術/検索技術 ●...該当、空欄...該当しない 業務上の重要度 ◎...全分野に影響、○...一部分野あるいは一部文献に影響、—...業務への影響は無し
 実現難易度 ○...実現が可能、△...実現が難しい 時系 短...短期対応、長...長期対応

表 10.2-6結果と考察における課題と解決策(5/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
6-2	文献照会のスピードが担保される必要あり。※キーワード翻訳方式のみの課題。コンテンツ翻訳方式は既に翻訳しているため、JP 公報と同様の性能。 → 『9.2.5(5)1)現行システム機能の踏襲』	翻訳後に表示するため、時間がかかる。長大文献は、時間が分オーダーになることもあった。 HW 増強や、キャッシュ機能の用意等である程度短縮可能だが、限界はある。	●		◎	①タグ語とに翻訳→表示 全文翻訳だと遅いので、翻訳をタグ毎に実施し、タグ毎に表示。 ②翻訳キャッシュ機能 一度翻訳したらキャッシュ保持し、毎回翻訳することを避ける。 但し、実際の審査時に毎回同じ文献をみる事は少ないと思われるため、キャッシュヒット率は低く、本質的な解決策にはならない。	△	長
6-3	動作を検索できない。 → 『9.2.5(5)1)現行システム機能の踏襲』	動作の検索は、国内文献でも難しい。 国内文献での検索方法は以下。 ・F タームを使用する。 ・近接演算子(近傍検索)を使用する。 ・関連する単語を複数記載する。 中韓文献は、IPC のみ付与されているため、F タームの様な多観点での検索ができない。		●	○	近傍検索を多言語横断検索で実現する。	○	短
6-4	ローカルで簡単に修正できるような辞書を用意し、審査官自身が少し操作するだけで次回から正しく翻訳される機能が必要。 → 『9.2.5(5)2)検索精度・翻訳精度を高めるための工夫』	スクリーニングの際に、辞書への用語追加や蓄積文献の再翻訳のタイムラグを短縮するために必要。	●		◎	システムに例えば、以下のような機能を加えることで可能と思われる。 ・辞書修正機能。 ・サーバへ辞書をアップロードして、翻訳時に辞書指定する。	○	短
6-5	図面と文献を同一画面で表示すべき → 『9.2.5(5)2)検索精度・翻訳精度を高めるための工夫』	図面に対する説明等を読む際に、参照する必要がある。	●		◎	図面と文献を同一表示する UIを開発する。 (例:フレーム分割やタブ切り替え等)	○	短

凡例:

翻訳技術/検索技術 ●・・・該当、空欄・・・該当しない 業務上の重要度 ◎・・・全分野に影響、○・・・一部分野あるいは一部文献に影響、－・・・業務への影響は無し
実現難易度 ○・・・実現が可能、△・・・実現が難しい 時系 短・・・短期対応、長・・・長期対応

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	10-11

表 10.2-7結果と考察における課題と解決策(6/6)

項番	課題	推定される原因	カテゴリ			解決策	実現性	
			翻訳技術	検索技術	業務上の重要度		難易度	時系
6-6	原文と日本語訳を同一画面で表示すべき → 『9.2.5(5)2)検索精度・翻訳精度を高めるための工夫』	誤訳があった場合、正しい訳の調査に、原文と日本語訳の対比を行うため、必須。			◎	原文と日本語訳を同一表示する UI を開発する。 (例: フレーム分割やタブ切り替え等)	○	短
6-7	対象国での審査状況や引用文献情報を表示する。 → 『9.2.5(5)2)検索精度・翻訳精度を高めるための工夫』	芋づる式に、引用文献を発見できる可能性があるため必要。 但し、中韓共に、引用文献データは公開していない。		●	◎	システム的には情報のリンクを作るだけで対応可能だが、元データとして、中韓の経過情報や引用文献情報の入手が必要。	○	長
7	蓄積性能							
7-1	コンテンツ翻訳方式は蓄積／再蓄積に時間がかかる。	コンテンツ翻訳方式は、蓄積前に翻訳処理をするため、時間がかかる。	●		◎	サーバ台数を増やして翻訳することで、できるだけ時間を短縮する。	○	短
8	その他							
8-1	翻訳文献の証拠性	機械翻訳の精度が上がったとしても、その正確性を担保する術が無いので、引用文献として拒絶できない。		●	◎	法律に絡むため、現時点では、解決策を導くことができない。	△	長
8-2		最終的には人手の翻訳文献が必要だが、翻訳文献の証拠性を確保するためには、クエリと原文、日本語訳のどこが対応しているかがわかる必要がある。	●		◎	①翻訳精度を上げる。	△	長
8-3			●		◎	②日本語訳のヒットワード反転+日本語訳、英抄、原文を比較しやすい UI を実装し、クエリと原文の対応関係がわかるようにする。	○	短
8-4	UNICODE で、日中韓の似ている漢字が同じ文字コードに割り当てられている (CJK 統合漢字。例: 「単」と「単」、「机」と「机(「機」の簡体字))	—		●	◎	日中韓で、UNICODE の統一マッピングテーブルを用意する。	△	長

凡例:

翻訳技術／検索技術 ●・・・該当、空欄・・・該当しない 業務上の重要度 ◎・・・全分野に影響、○・・・一部分野あるいは一部文献に影響、—・・・業務への影響は無し
 実現難易度 ○・・・実現が可能、△・・・実現が難しい 時系 短・・・短期対応、長・・・長期対応

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	10-12

10.3 本検証方法における課題と解決策

10.3.1 結果

本検証を通じて判明した、検証方法における主要な課題を『表 10.3-1 検証方法課題』に示す。

表 10.3-1 検証方法課題

項番	カテゴリ	概要	影響箇所
1	全文検索	一部機能を全文検索モードに対応しなかった。	ユーザ検証全般
2	ユーザ検証の 検証課題	正解の存在しない検索課題を選択した可能性がある。	ユーザ検証全般
3		概念検索に向かない分野があった。	ユーザ検証全般 ・電子・電気系分野
4	業者検証の 正解データ	検索課題と正解データが一字一句同じ内容	9.2.3(4)検索対象範囲
5		検索課題と正解データが 1 対 1	9.3 基準値の提示

10.3.2 考察

『10.3.1結果』で挙げた本検証方法における課題に対する解決策を述べる。

(1) 全文検索

1) 一部機能を全文検索モードに対応しなかった

モデル検証では、ユーザ検証と業者検証の検証方法を統一する目的で、「自然文検索」をメイン¹に各機能の効果を確認した。このため、一部、全文検索モードにおいて実装しなかった機能があった。

このため、通常、審査に使用している全文検索でのユーザの体感が、一部機能において確認できなかった。

解決策としては、以下が考えられる。

- ✓ 全文検索モードに機能を実装し、ユーザの体感を確認する

a) 全文検索モードに機能を実装し、ユーザの体感を確認する

業者検証(大量の検索課題)で全文検索は実施できないため、ユーザ検証は自然文検索で実施していただく必要がある。しかし、体感的に、機能の有効性を確認していただくためにも、全文検索モードに全機能を実装し、自由に使えるようにする必要がある。

¹業者検証では、キーワードを検討できないため、第一請求項を使用した自然文検索を実施した。

(2) ユーザ検証の検証課題

1) 正解が存在しない可能性

モデル検証では、検索課題に対する正解の有無を確認できなかった。これは、システムの、正解と判断できるデータは、「ファミリー文献」程度しか考えられず、また、『7.4.1(2)検索課題抽出の条件』を満たす文献に「ファミリー文献」が存在しなかったためである。

結果、検索課題に対して「正解が必ず 1 つは存在する」と言えず、今回の蓄積文献内に正解が存在しない検索課題があった可能性がある。

解決策としては、以下が考えられる。

- ✓ 検索課題と正解のセットを作成する

a) 検索課題と正解のセットを作成する

国内文献の検索課題と外国文献の正解セットを作成すると、「必ず 1 つは正解がある」という前提となり、大規模な検索課題を流す業者検証とユーザの体感がマッチする可能性が高い。このため、検索課題と正解のセットを作成する必要がある。

2) 概念検索に向かない分野

モデル検証では、公平性を保つために、検索課題を無作為に抽出した。しかし、検証を通じて、①プロトコル系分野等、処理の流れを表すような発明は、F タームを使用できないため、検索精度があまり良くないということが判明した。また、②G セクションや H セクション等の電子・電気系分野においては、(A)技術的特徴が一般的な言葉で記載されていること、(B)用語の統制が成されていないことから、検索精度があまり良くないということが判明した。

モデル検証システムは、概念検索をメインとしたため、このような問題が発生した。

解決策としては、以下が考えられる。

- ✓ 検証システムに近傍検索を実装する

a) 検証システムに近傍検索を実装する

国内文献において①プロトコル系の分野の案件を審査する際、F タームを使用する他にも、近傍検索を使用している。また、②一般的な用語や統制されていない語で記載されている文献群を検索する際には、近傍検索を用いて、文献を絞り込むことが有効である。このため、近傍検索を用いることで、一定の検索精度向上は見込めると考えられる。

但し、語順を指定した近傍検索は、多言語横断検索で、著しい効果は望めないため、ノイズが混ざり可能性がある。(詳細は、『9.2.5 翻訳精度・検索精度を高めるための工夫』を参照。)

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	10-14

(3) 業者検証の正解データ

モデル検証では、正解データにファミリー文献を使用した。このため、①検索課題と正解データが、一字一句同じ内容で、②検索課題と正解データが1対1となってしまう。

1) 検索課題と正解データが一字一句同じ内容

検索課題と正解データが一字一句同じ内容であると、正解データ中の重要でない語(「発明」や「装置」等)が多数ヒットし、正解データのスコアが高くなり、検索結果の順位が必要以上に高くなることがある。このため、審査で得られるはずの効果が業者検証で得られないことがあった。

2) 検索課題と正解データが1対1

本来、審査では、本願に対する類似文献を1件だけ探すのではなく、複数件の正解を探すものである。しかし、業者検証の検索課題と正解データが1対1であると、審査と業者検証に不一致が起こり、効果的な検証ができなくなる虞がある。

解決策としては、以下が考えられる。

- ✓ 検索課題に対する正解データが引用関係である大規模な文献集合を作成する
- ✓ EPO の引用情報を利用して検索課題と正解データが引用関係の文献集合を作成する

a) 検索課題に対する正解データが引用関係である大規模な文献集合を作成する

例えば、①外国特許庁に出願している国内文献を検索課題とし、②検索課題のファミリー文献が引用した文献を正解データとすることで、作成することが可能である。

但し、SIPO、KIPO ともに、引用情報を公開していないため、引用情報を入手することが先決である。

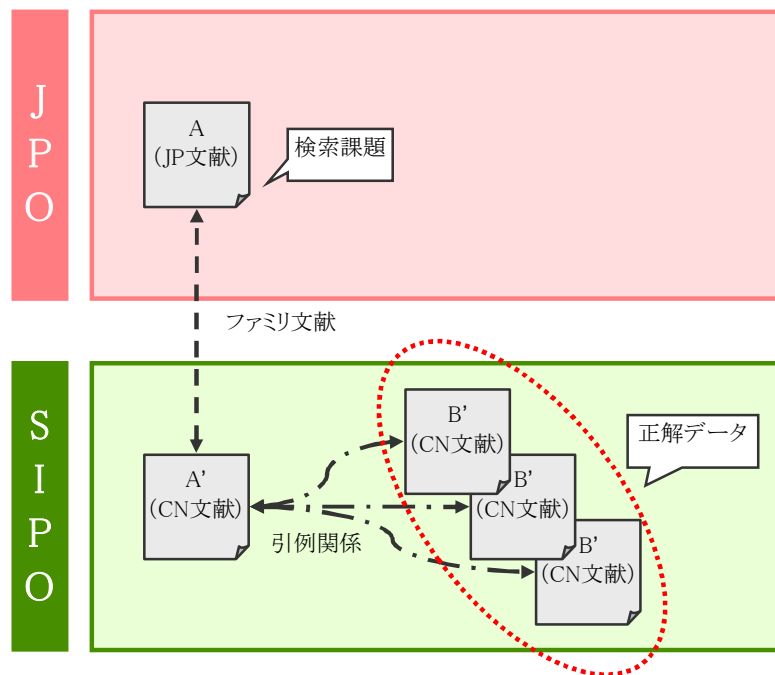


図 10.3-1 検索課題と正解データが引用関係

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	10-15

b) EPO の引用情報を利用して検索課題と正解のセットを作成する

EPO の引用情報を利用して、①EPO にも出願している国内文献を検索課題とし、②SIPO に出願している、①の引用文献のファミリーを正解データとすることで、正解データが複数存在するセットを作成できる可能性がある。

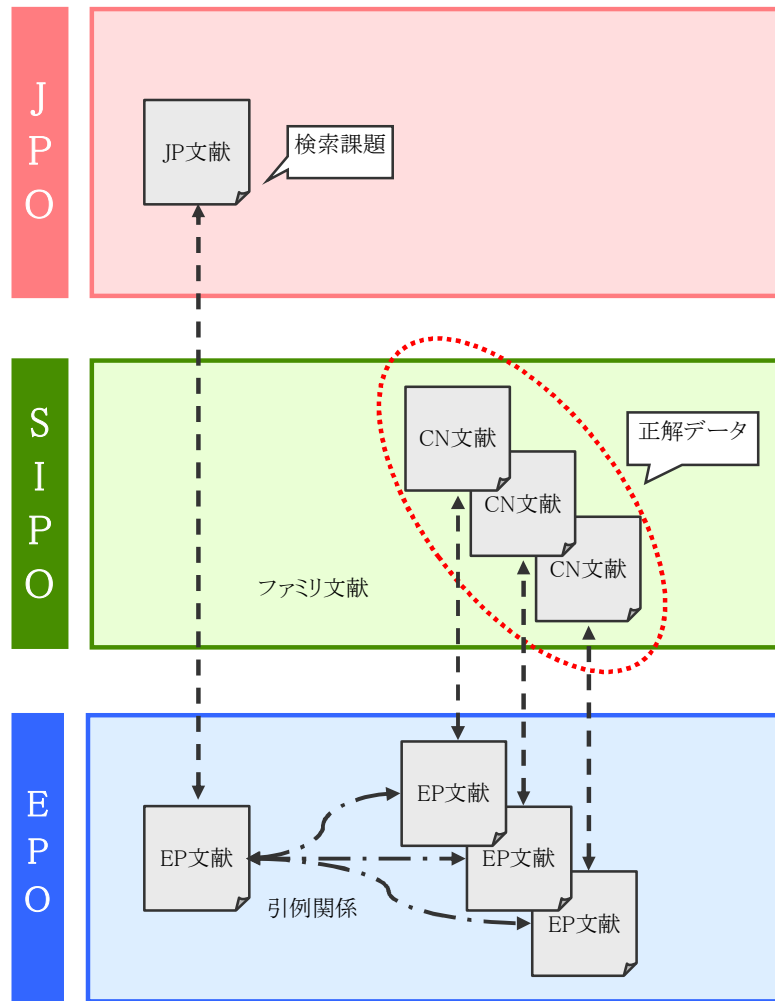


図 10.3-2 検索課題と正解が引用関係 (EPO 経由)

11 検索系最適化における多言語横断検索システムの構成案

導入システムへの要件を纏め、想定される構成案（コンテンツ翻訳方式、キーワード翻訳方式）について記述する。

11.1 システム要件

11.2 構成案

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	11-1

11.1 システム要件

11.1.1 要件決定の観点

検索系最適化における多言語横断検索システムの構成を検討するためには、機能要件だけではなく、運用、保守性、コスト面等々、その他要件も踏まえる必要がある。

このため、以下の観点から多角的な分析を行う。なお、本観点は、特許庁殿にて各種更改サーバに対して実施している観点及び、総務省が発行している「情報システムに係る政府調達の基本指針¹」も参考とした。

表 11.1-1 要件決定の観点

項番	観点	内容
1	信頼性	ハードウェア、ソフトウェア等の信頼性(冗長化等)。
2	性能	レスポンス性能と、品質的な性能(検索精度、翻訳精度)。
3	運用・セキュリティ	定常的な運用作業である、バックアップ、リカバリ、文献蓄積。また、当該システムの保守性についても整理。
4	処理方式	業務要件としての「オンライン処理」、及び文献蓄積等の「バッチ要件」の2点を整理。
5	インフラ設計	CPU/メモリ、サーバ台数等々、インフラ的な要件。
6	拡張性	データ量の増加、トランザクション量の増加、対象言語の追加への対応。
7	移行	公報のデータ移行(原文、日本語訳)と、システム移行(他社間インタフェース)について整理。
8	コスト	導入コスト、ランニングコストの傾向について整理。

また、上記観点を実現するシステム構成としては、「コンテンツ翻訳方式」と「キーワード翻訳方式」の2つが存在する。このため、次ページ以降にて、各要件に対して2つの方式毎の実現性等々も整理する。

¹調達手順のオープン化、コスト最小限化、競争力の促進を狙った指針書(http://www.soumu.go.jp/s-news/2007/070301_5.html)

11.1.2 システム要件

新検索システムでの多言語横断検索システムの実現に向け、必要となるシステム要件を整理する。以下に必要とされるシステム要件の概略を記載。また、要件に対する実現性について、○×で評価。

表 11.1-2 システム要件(全体概要) 1/2

項番	観点	コンテンツ翻訳方式	キーワード翻訳方式	
1	信頼性	— 処理方式による信頼性の実現性の差異は無し。	— 同左	
2	性能	検索	○ 現行 JP 公報と同一性能を確保可能。	○ 言語の違いはあるが、現行 JP 公報とほぼ同一と推測される。
		照会	○ 現行 JP 公報と同一性能を確保可能。	× 翻訳処理が必要なため、遅い。
		蓄積	△ 翻訳してから蓄積するため、翻訳時間が追加される。	○ 原文蓄積なので、言語の違いはあるが、現行 JP 公報とほぼ同一と推測される。
		翻訳精度	○ 中国語は更なる精度向上が必要。韓国語は、現時点でも実用レベルに近い。	△ ワードによる全文検索の場合、訳語の曖昧性が出てしまう。但し、訳語候補展開機能にて補正可能。
		検索精度	○ JP 公報と同等の性能確保も可能。	△ 言語毎の各種特性を考慮したチューニングが必要。
3	運用・セキュリティ	○ バックアップや文献蓄積に時間が掛るが、保守性は良い。	△ 各国言語に対応した検索エンジンの導入が必要なため、保守性は悪くなる可能性が高い。	
4	処理方式	○ 現行機能は、ほぼ踏襲可能。(機能、性能共に)	△ 日本語訳を利用するケースで難しい点がある。(ヒットワード反転の実現、スクリーニング性能の確保)	

凡例:○…実現可能、△…難しい面があるが回避可能、×…実現は非常に難しい。

表 11.1-3 システム要件(全体概要) 2/2

項番	観点		コンテンツ翻訳方式	キーワード翻訳方式
5	インフラ設計		△ 再翻訳時には、多数のサーバが一時的に必要な。	△ 各国言語に対応した検索エンジンの導入が必要。 (外国製エンジンが有力)
6	拡張性	リソース	○ データ増加には、蓄積用の翻訳サーバの増設で可能。 トランザクション増加には、JP 公報と同等の考え方で実現可能。	○ 同左。但し、データ増加に、蓄積用の翻訳サーバの増設は不要。
		言語	○ 対象言語の翻訳エンジンは片方向(原語→日本語)のみで良い。 また、人手翻訳済みの抄録等を蓄積して利用することも可能。	△ 翻訳エンジンは双方向が必要。 また、人手翻訳済みの抄録等を検索で利用することはできない。
7	移行		△ 日本語訳のデータ移行が必要。	○ 原文データのまま利用するため、国内公報のデータ移行と同等。
8	コスト		○ 複数台サーバでの再翻訳のために、翻訳エンジンのライセンス費が要。また、ディスク量(コスト)が2倍必要。但し、辞書が片方向整備(中→日)のみで良い。	△ 各国言語に対応した検索エンジンのライセンス費が必要。 また、海外製品だと、製品保守やトラブル発生時の対応コストが掛りやすい。

凡例:○…実現可能、△…難しい面があるが回避可能、×…実現は非常に難しい。

11.2 構成案

『11.1.2 システム要件』の各種要件を加味して、コンテンツ翻訳方式、キーワード翻訳方式のシステム構成案を記載する。

なお、構成案で提示したHWスペックは、あくまで今回の検証結果を元に算出した概算の予測値である。このため、正確なサイジングを行う為には、別途、性能測定を行って定義する必要がある。

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	11-5

11.2.1 構成案 1 ～コンテンツ翻訳方式～

(1) 構成案

前述の前提を元にしたコンテンツ翻訳方式の構成案を以下に示す。

検索対象を日本語に統一するため、シンプルなシステム構成で実現可能。但し、蓄積処理(特に再翻訳)に複数の翻訳サーバが必要となる点がネックである。

コンテンツ翻訳方式の方が、原文と日本語訳のデータを共に持つため、データサイズが大きくなると思われるが、実際は、全データサイズに占める割合は、イメージデータが主であり、翻訳方式によるデータサイズの差異はほとんど無い。

なお、EPO でもコンテンツ翻訳方式を採用しており、JP 等の公報が検索可能である。

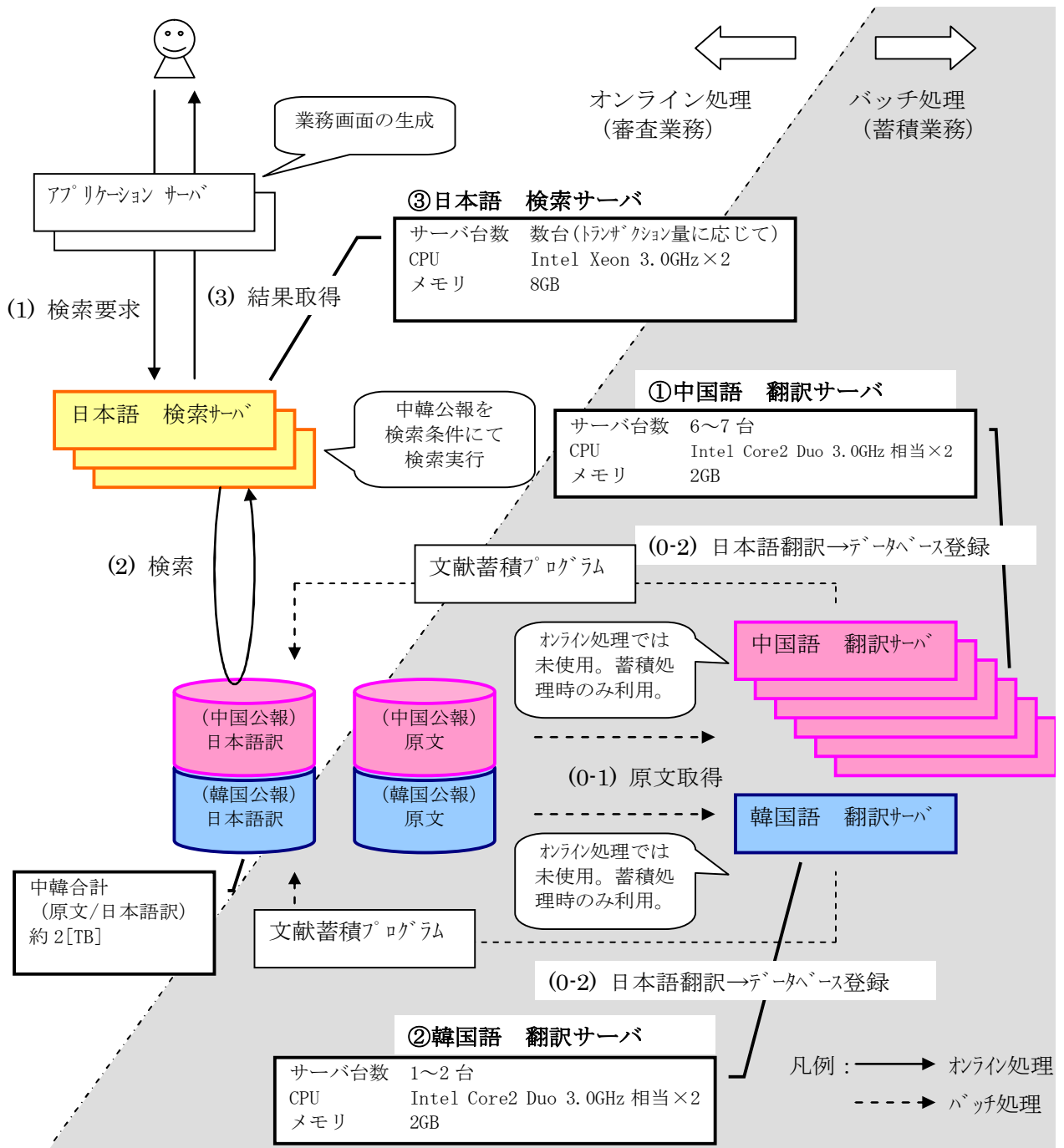


図 11.2-1 コンテンツ翻訳方式の構成案

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	11-6

11.2.2 構成案 2 ～キーワード翻訳方式～

(1) 構成案

前述の前提を元にしたキーワード翻訳方式の構成案を以下に示す。

コンテンツ翻訳方式に比べると、中韓それぞれの検索システムに検索要求を行い、アプリケーション層で取りまとめる流れである。翻訳サーバの精度向上の即時反映が可能だが、日本語訳のスクリーニング時に都度翻訳サーバを使った翻訳処理が必要となる。

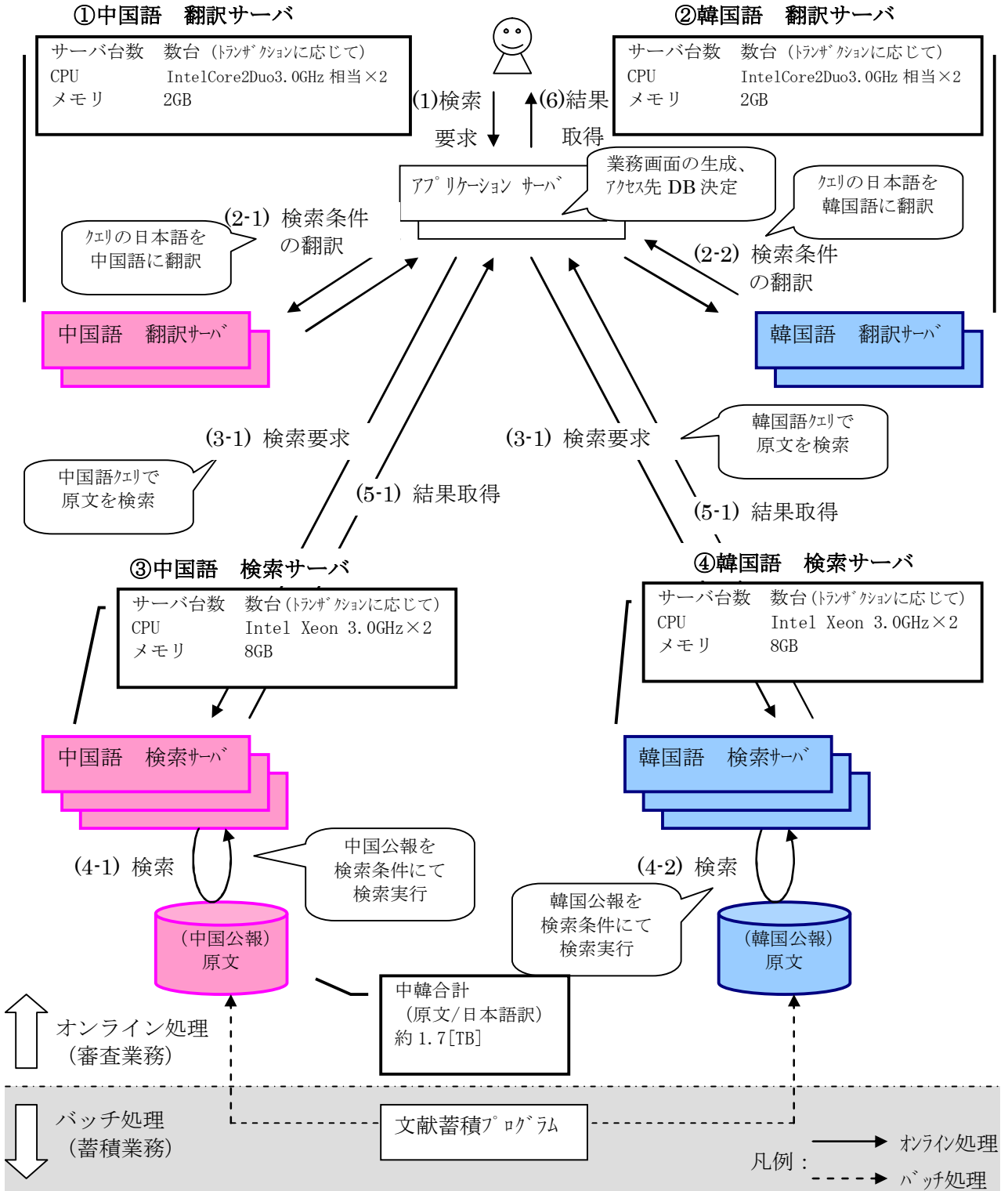


図 11.2-2 キーワード翻訳方式の構成案

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	11-7

12 総括

長期的に解決していくべき課題と多言語横断検索の将来的な展望について記述する。

12.1 今後の課題

12.2 将来的な展望

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（詳細版）	Rev.	頁
調査報告書		1.0	12-1

12.1 今後の課題

モデル検証で判明した課題の中で、特に長期的に対応していく必要のある課題を、『表 12.1-1 長期的に対応すべき課題』に示す。また、モデル検証で判明した全課題（短期的に解決する課題を含め）は、『10.2 結果と考察における課題と解決策』を参照。

表 12.1-1 長期的に対応すべき課題

項番	課題	解決策	担当区分	
			庁	ベ
1	中国語の翻訳精度が低い。	形態素解析、構文解析の精度を向上させる。	—	○
		公的機関、業界団体が連携し、辞書の登録語数を増加させる。	○	○
2	翻訳精度が向上しても、翻訳文献の正確さを担保できない。	簡便に外国文献を人手翻訳する環境を整備する。	○	—
		法改正も視野に入れて、調整する。	○	—

凡例:

担当区分: 庁…特許庁、ベ…ベンダ ○…担当、—…担当外

12.1.1 今後の課題(詳細)

(1) 中国語の翻訳精度が低い

モデル検証現在の、翻訳技術では、中国語翻訳を韓国語翻訳と比較すると、現状の翻訳精度では、審査に使えるレベルでないということが判明した。翻訳精度を向上するプロセスは、以下の 2 点である。

1) 形態素解析・構文解析

中国語は、『9.1.1(1)中国の言語特性』で述べた通り、単語の区切りが明示されない「膠着語」であるため、形態素解析が非常に難しい。また、「同表記異品詞」が多いこと等から、構文解析も非常に難しい。

しかし、形態素解析や構文解析の技術は、一朝一夕に向上するようなものではないため、ベンダが継続的に研究し、技術の向上に励む必要がある。

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（詳細版）	Rev.	頁
調査報告書		1.0	12-2

2) 辞書の登録語数

中国語は、『9.1.1(1)中国の言語特性』で述べた通り、異表記(同義語)が多数存在する。このため、辞書の用語登録数を多くする必要があるが、現在の登録語数は、英語翻訳と比べても圧倒的に少ない。

しかし、特許庁に必要な分野に特化した辞書をベンダが作成することは考えにくく、また、特許庁だけで、辞書を大規模に作成することも現実的ではない。このため、公的機関や業界団体が連携して、辞書の作成を推進していく必要がある。

尚、連携方法の例として、『図 12.1-1 公的機関と業界団体の連携(例)』のような方法が考えられる。

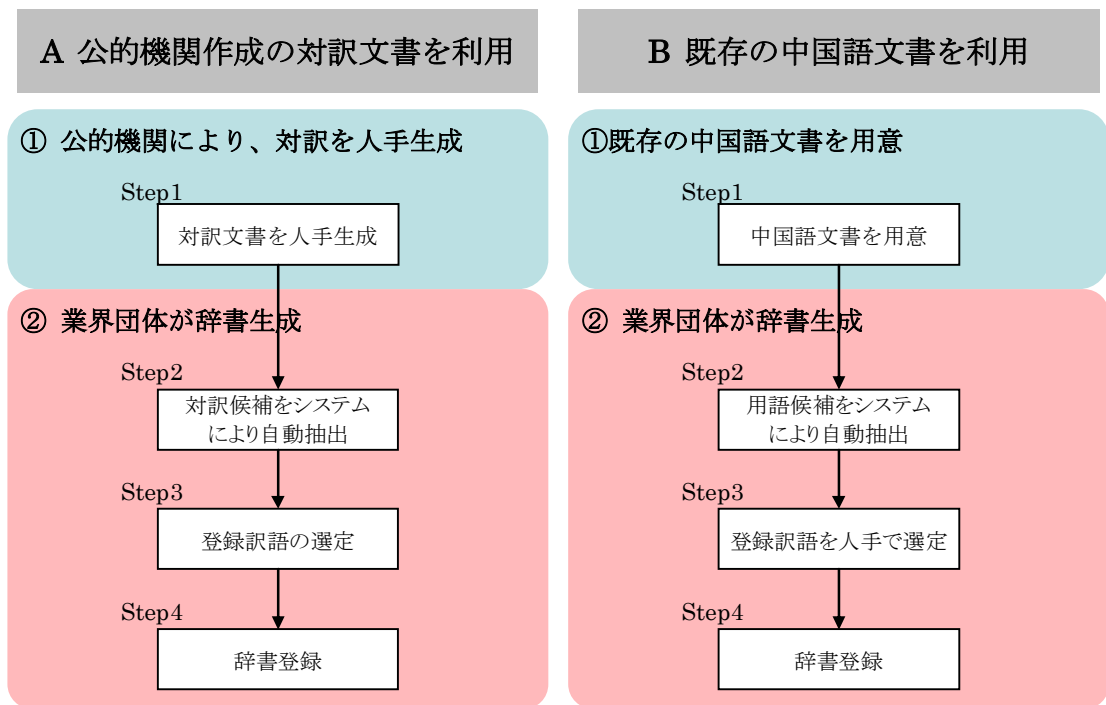


図 12.1-1 公的機関と業界団体の連携(例)

辞書の権利が課題となるが、整理標準化データのように、作成した辞書を対外提供ができれば良い。しかし、公的機関が作成した対訳を使用し、業界団体(複数ベンダがそれぞれ)が辞書を作成したとすると、どこの著作物となるのか不透明である。このため、対外提供が難しくなる。

また、翻訳エンジンの更改時に、辞書を引き継ぐことができなくなる可能性がある。これは、例えば辞書のライセンスを購入することで翻訳エンジンが変更しても使える、というようなライセンス形態を検討することで、解決できる可能性がある。

以上のように、平成 26 年の稼働前に 100 万語の辞書を作成する場合、平成 24 年中頃までには、課題をクリアにする必要がある。

(2) 翻訳文献の正確さを担保できない

今後、機械翻訳の精度が向上し、外国文献を引用文献とする際、機械翻訳文献の翻訳精度を担保する術がない。

このため、①人手翻訳を簡便に行える環境を整備する、②原文を引用できるような法令要件の解釈を整理する必要がある。

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 (詳細版)	Rev.	頁
調査報告書		1.0	12-4

12.2 将来的な展望

産業界のグローバル化が進む中、非英語圏も視野に入れた特許審査は、ますます重要性が高くなる。また、ファーストアクション短縮に向けて、現行の審査スピードを落とすことなく、非英語圏の文献のサーチが必要とも考えている。

また、今回のモデル検証では、多言語横断検索技術の両輪である、「機械翻訳技術」、「情報検索技術」ともに、多くの課題を洗い出すことができた。これらの課題は、解決不可能なものではなく、短期的、あるいは長期的に、計画性をもって対策を実行することで解決するものばかりであり、数年後には更なる機能向上が期待できる。

このため、日本語で外国文献を検索できる「多言語横断検索技術」は、今後の審査業務に必要なツールになる可能性を秘めており、日本国特許庁の「グローバル化対応力の向上」を推し進めるための強力な武器になると考えている。

弊社も「審査で使える多言語横断検索」システム実現に向けて、今後も継続的な研究開発を進め、特許庁の「グローバル化対応力の向上」の一助となるべく、日々邁進していく所存である。

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（詳細版）	Rev.	頁
調査報告書		1.0	12-5