多言語特許検索システム PRIME

樋口　重人*1　牧田　光晴 *1　藤井　　敦*2,3　石川　徹也*2
*1（株）パトリス
*2 図書館情報大学
*3 科学技術振興事業団 CREST

1. はじめに

経済のグローバル化によって，特許検索の世界でも従来の国内の特許出願データを検索するだけでなく，外国の出願に関しても検索することが重要になってきた．日本人利用者は海外データを，外国人利用者は日本国データを検索可能にしなくてはならない．また最近の特許検索においては特別な知識がなく使えるようなインターフェースが好まれる傾向にある．これは専門家のサーチツールだったものが一般の研究者でも自由に使えることが求められてきた結果である．そこで我々は藤井ら[1) 2)]が提案した手法を用いて，多言語特許検索システム「PRIME」（*Patent Retrieval In Multi-lingual Environment*）を開発した[3) 4) 5)]．PRIMEは，検索キーワードを検索対象の言語に翻訳することで多言語検索を実現する．

特許に用いられる技術用語は単語を組み合わせて創造する語が多い．これらの技術用語を的確に翻訳する為には対訳辞書の定常的な更新が必要である．

自然言語処理の研究では，対訳関係にある多言語コーパス（例文集）から単語や句の対訳を自動抽出する手法が提案されている[5) 6)]．しかし，一般的に対訳コーパスの入手や作成は高価である．そこで本研究は，優先権主張制度に基づいて出願された特許から複合語対訳を自動抽出する手法を取り入れた[7)]．

また特許検索は膨大な件数を検索しなくてはならない．目的の特許に辿り着くまでには各種の検索用語を投入し検索を繰り返すことも多々ある．そこで検索結果の集合をクラスタリングすることにより絞込みが出来るように試みた．これは検索結果の中で技術用語を用いて分類していく機能である．研究者が使用することを考慮し，わかりやすい表示を行うことを目的とした．

以下，2章でシステム概要を説明し，3章で評価を，4章で今後の課題を述べる．

2. システム構成

2.1 システム概要

本研究で開発した多言語特許検索システム PRIME の構成を図１に示す．実線はオンライン処理，破線はオフライン処理を表す．

搭載データは，日本特許抄録 1995 年～1999 年の 5 年分（約 175 万件），PAJ（Patent Abstracts of Japan）を同期間，同件数である．

上記の年範囲において優先権主張制度に基づいて出願された特許を特定し，その元となる米国での登録公報を抽出し搭載した．これらは日本公開公報，米国特許登録公報で約 32,000 件づつである．

利用者が質問を入力すると翻訳部において目的言語に翻訳される．本システムでは日英，英日の翻訳を実現している．質問には単語，複合語が入力でき，日本語の場合は形態素解析システム「茶筌」[注1)]を，英語に対しては WordNet [注2)]の品詞情報を利用して名詞性単語を抽出する．

その後，対訳辞書を引きながら単語や複合語の単位で翻訳する．しかし，一般に一つの語に対し複数の訳が存在する為，訳語候補の全てを採用するとノイズが増大する．そこで対訳辞書から抽出した翻訳モデルと検索対象の特許 DB から抽出した言語モデル（単語バイグラム）を使用して翻訳の曖昧性を解消する．

注1）　http://chasen.aist-nara.ac.jp/index.html.ja
注2）　http://www.cogsci.princeton.edu/~wn/

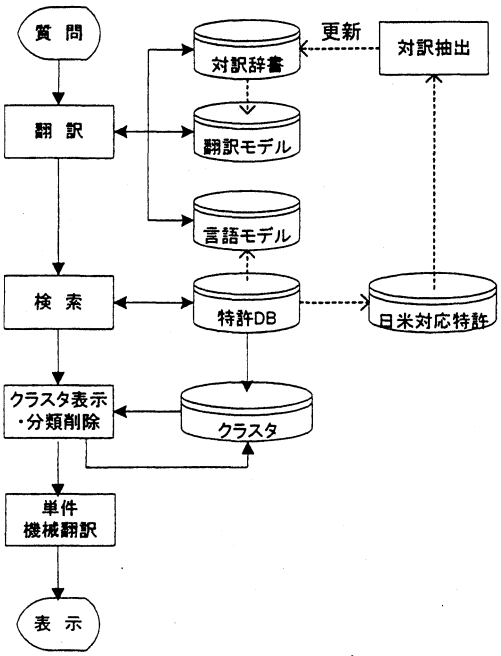

図１：多言語特許検索システム PRIME の構成

先に我々は日米の優先権主張に基づく特許出願の対を対訳コーパスと位置付け，複合語の新語の自動抽出を試みた[7]．ここでは約 32,000 件の対応特許を対象に評価実験を行い，年間数千件の新語を半自動的に取得できる見通しを得た．

次に，翻訳された語を用いて検索を行う．ここでは形態素解析処理後の単語により索引ファイルを事前に作成し，検索に利用する．

検索結果を得た後，クラスタリングを行い分類の様子を表示する．そこで不要な分類は削除を行い，再度クラスタリングを行う．絞り込まれた結果の中から単件表示を行い機械翻訳によってユーザーの母国語で表示する．

2.2 クラスタリング

特許文書を検索，分類する際には，国際特許分類(IPC)などが使われる．しかし，分類体系を把握しておかないと駆使することは難しい．そこで我々は検索の結果，その文書群に含まれる特徴的な用語により分類表示を行うことにした．この方式であれば，事前の予備知識を必要とせずに分類が可能になる．また最初の検索の結果，ヒット分件数が多い場合でも，それらの結果を分類することで不要な文書群を取り除くことができる．

クラスタリングには GETA[9]を使用した．特許 DB を索引付けする際に出来る中間ファイル（文書に含まれる単語のみを集約したファイル）を利用する．これを GETA の入力ファイルとし，あらかじめ単語の頻度統計を集計しておく．

PRIME で翻訳・検索した時，検索結果の文書群が特定される．これらの文書 ID と上述で作成しておいた文章に含まれる単語ファイルを GETA に引渡し，クラスタリング表示を行う．

クラスタリングする数は 3～10 個と可変とし利用者が選択できるようにし，不要文書群を検索結果から削除出来るようにした．「検索結果の表示」→「クラスタリング表示」→「不要文書群の削除」を繰り返すことにより検索結果を絞り込め，目的の特許文書に行き着くことが出来る．

以下に画面遷移の実行例を示す．図 2 は検索結果が得られた画面である．

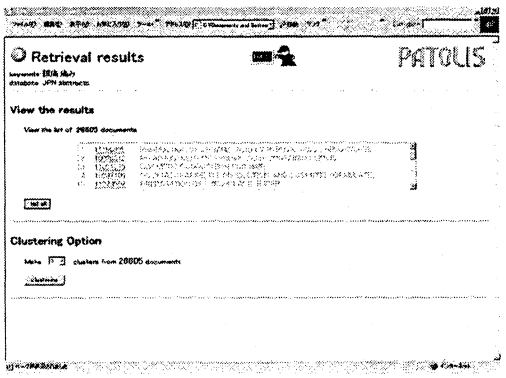

図２：検索結果が得られた画面例

図 3 は図 2 の結果を 3 つのクラスタに分類した結果である．この 3 クラスタの中で不要な文書群があった場合，それを取り除いて再度クラスタリングを行う．この繰り返しを行った後，単件表示，機械翻訳結果を表示する．図４では機械翻訳結果を示している．

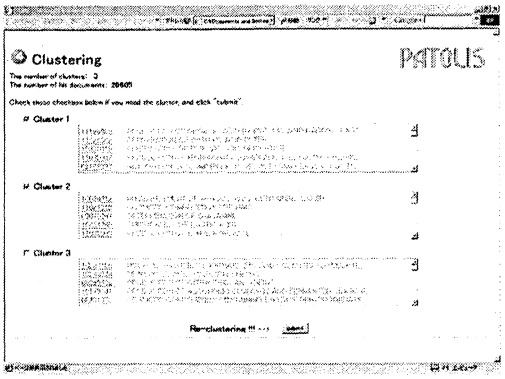

図３：クラスタリングを行った画面例

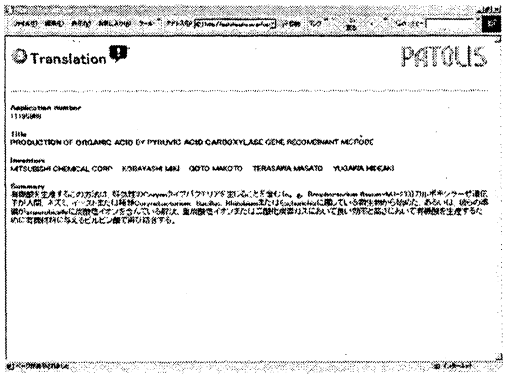

図４：単件を選択し機械翻訳を行った画面例

3. 評価

3.1 翻訳部の評価

PRIME の検索質問の翻訳機能について評価を行った．先の研究において辞書に掲載されていない日英対訳語を抽出することが出来た[7]．これらの語を用いて翻訳率を測定した．ここで使用した語数は 5951 語である．表 1 に例を記載する．

結果として日英で 94.2%，英日で 92.8%の正解率であった．

日本語	英語
ＡＴＭネットワーク	ATM network
くさび部材	wedge member
はんだドット	solder dot
アーチワイヤースロットライナー	archwire slot liner
アシルアセトアニリド系イエロー画像	acylacetanilide yellow image
イオン注入機	ion implanter
ウェーハ温度	wafer temperature
エアバッグモジュール	air bag module
エチレンコポリマー	ethylene copolymer
エネルギー感受性材料	energy sensitive material
エポキシ樹脂組成物	epoxy resin composition
エラー訂正コード	error correction code
エンジン性能	engine performance
オーバーレイ画像	overlie image
オルガノクロロシラン	organochlorosilane
カスコード電流ミラー	cascode current mirror

表１：評価に使用した対訳例

まず日本語から英語への翻訳について説明する．上述の 5951 語を正解の対訳語とし PRIME に翻訳させ，翻訳結果と 5951 語の正解語を人手により照合した．正解語数は 2718 語（45.7%）であり，異表記や省略形などの正解とみなせる訳語は 2886 語(48.5%)であった．日本語のままであったり，明らかな誤訳などは合計 347 語（5.8%）であった．

異表記や省略形などは，「semi-conductor」のハイフンの有無や「AMP=amplifer」であった．

日本語→英語への翻訳正解率

1	正解	◎	2718語	45.7%
2	異表記等	○	2886語	48.5%
3	未翻訳	×	67語	1.1%
4	誤訳	×	280語	4.7%
	合計		5951語	100%

表２：日英翻訳率

次に英語から日本語への翻訳結果について説明する．日英と同様に上述の 5951 語を PRIME に翻訳させ結果を人手により照合した．正解は 1532 語（25.7%），異表記（ひらがな，カタカナなど含む）は 3989 語（67%）であった．翻訳できず英語のまま

であったもの，明らかな誤訳は合計で430語(7.2%)であった．異表記の例は「rear side」を「後側」と訳すか「リア側」と訳すかなどであった．

英語→日本語への翻訳正解率

1	正解	◎	1532語	25.7%
2	異表記等	○	3989語	67.0%
3	未翻訳	×	398語	6.7%
4	誤訳	×	32語	0.60%
	合計		5951語	100%

表3：英日翻訳率

日英，英日の双方向で90%以上の正解率を出せたことにより商品化などの応用に期待が出来る．
ユーザーが訳語を追加出来るようにすることで，正解率の向上が可能となる．

4. 今後の課題

PRIMEの機能として，母国語による質問入力，翻訳，他国語のデータ検索，一覧表示，検索結果群のクラスタリングによる分類，検索結果単件の機械翻訳を実現した．今後は言語の種類を拡張する予定である．

5. 謝辞

対訳辞書及び機械翻訳システムは（株）ノヴァの許諾を得て使用させて頂きました．この場を借りて深謝致します．

参考文献

1) 藤井敦，石川徹也．技術文書を対象とした言語横断情報検索のための複合語翻訳．情報処理学会論文誌，Vol. 41，No. 4，pp. 1038-1045，2000.

2) 藤井 敦，石川 徹也．質問翻訳と文書翻訳を統合した日英言語横断情報検索．電子情報通信学会論文誌，Vol. J84-D-II，No. 2，pp. 362-369，2001.

3) Masatoshi Fukui，Shigeto Higuchi，Youichi Nakatani，Masao Tanaka，Atsushi Fujii and Tetsuya Ishikawa．Applying a Hybrid Query Translation Method to Japanese/English Cross-Language Patent Retrieval．ACM SIGIR Workshop on Patent Retrieval，2000.

4) 樋口重人，福井雅敏，藤井敦，石川徹也．特許情報を対象とした言語横断検索システムの開発．言語処理学会第7回年次大会発表論文集，pp. 445-447，2001.

5) 北村美穂子，松本裕治．対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol. 38，No. 4，pp. 727-736，1997.

6) Frank Smadja，Kathleen R．McKeown and Vasileios Hatzivassiloglou．Translating Collocations for Bilingual Lexicons: A Statistical Approach．Computational Linguistics，Vol. 22，No. 1，pp. 1-38，1996.

7) 福井雅敏，樋口重人，藤井敦，石川徹也．日米対応特許コーパスを用いた対訳抽出手法．情報処理学会研究報告 2001-NL-145，pp. 23-28，Sep. 2001.

8) Shigeto Higuchi，Masatoshi Fukui，Atsushi Fujii，and Tetsuya Ishikawa．PRIME: A System for Multi-lingual Patent Retrieval Proceedings of MT Summit VIII，pp. 163-167，Sep. 2001.

9) （株）日立製作所，東京工業大学，北陸先端科学技術大学院大学，文部省国文学研究資料館
汎用連想計算エンジンの開発と大規模文書分析への応用
情報処理振興事業協会 「独創的情報技術育成事業」第19回IPA技術発表会
2000年10月11-12日
http://www.ipa.go.jp/STC/dokusou.html