特許情報を対象とした言語横断検索システムの開発

樋口 重人[†]　福井 雅敏[†]　藤井 敦[††,†††]　石川 徹也[††]
[†]（財）日本特許情報機構
[††] 図書館情報大学
[†††] 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

近年の国際化の波に沿って，国内外の企業が海外での特許権利化を目指した海外出願重視の傾向にある．そこで，日本における特許検索サービスでは，海外出願された特許を検索する日本人利用者と，国内出願された特許を検索する海外利用者を考慮することが重要である．しかし，外国特許の検索では言語の違いが大きな障壁になり，利用者に過度の負担を強いることになる．

本研究では上記問題に対処するために，言語横断検索手法に基づく特許検索システムの研究開発を行った．本システムは日英を双方向に検索可能である．言語横断検索（Cross-Langauge Information Retrieval: CLIR）は，質問言語と異なる文書を検索するための技術である．インターネットを介して氾濫する多言語文書を有効に利用することなどを背景にして，近年盛んに研究されている．

本研究では，藤井らによって提案された種々の手法[2, 7, 8]に基づいてシステムを開発した．言語横断検索には様々な手法があるものの，当手法には本研究の目的にとって好ましい複数の特長がある．

まず，利用者の質問を対象言語に翻訳してから検索を行う「質問翻訳型」なので，質問翻訳機能を実装すれば，既存の単言語向け特許検索システム（エンジン，データベース等）をそのまま利用でき，開発コストが安価である．質問翻訳機能は，単語辞書や特許文書から抽出した統計情報を用いることで比較的容易に実装できる．

また，言語横断の特許検索においては，日々増え続ける新語を的確に翻訳することが重要である．この問題に対しては，複合語対訳から単語対訳を抽出して辞書を拡張する手法や翻訳辞書に未登録の単語を音韻的に等価な外国語に変換する翻字処理で対応することができる．

以上の機能は，著者らの先行研究[3, 6]で部分的に実装した．さらに今回は，新語に対する問題をより確実に解決するために，日英特許文書から対訳を自動抽出する機能を新たに導入した．

2 システム概要

本研究で開発した言語横断特許検索システムの構成を図1に示す．この図において，実線はオンライン処理，破線はオフライン処理を表す．本システムが対象とする質問言語および文書言語は日本語と英語である．

本章ではオンライン処理を中心に説明する．図1右側の対訳抽出機能については3章で説明する．

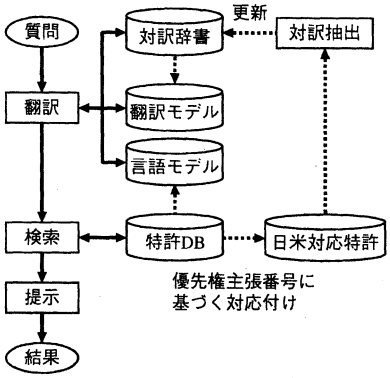

図 1: 言語横断特許検索システムの構成

利用者が入力した質問は，翻訳部によって文書言語に翻訳される．質問として，単語，複合語，文を入力することができる．質問が文の場合は，名詞（列）を単語や複合語として抽出し，以降の処理に利用する．そこで，日本語には，形態素解析システム「茶筌」[10]を，英語に対してはWordNet [1]の品詞情報を利用して名詞性単語を特定する．

質問は，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，一般に単一の語に対して複数の訳語が定義されており，訳語候補を全て用いると不要な特許文書が数多く検索されてしまう．そこで，対訳辞書から抽出した翻訳モデルと，検索対象の特許データベースから抽出した言語モデル（単語バイグラム）を利用して訳語の曖昧性を解消する[2, 7]．

本研究では，機械翻訳用に作成された（株）ノヴァの専門用語辞書を対訳辞書として利用した[1]．本辞書は19の専門分野で構成され（表1），合計約100万件の日英対訳を定義している．質問入力画面では，利用者が対象分野を明示的に選択するか，または全辞書を一括して翻訳に利用する．

表1: ノヴァ専門分野辞書の構成分野

航空・宇宙，バイオテクノロジー，ビジネス，化学，コンピュータ，土木・建築，防衛，地球環境，電気・電子，原子力・エネルギー，金融，法律，数学・物理，機械工学，医療・医学，金属，海洋・船舶，プラント，貿易

次に，検索部によって，翻訳された質問に関連する特許文書を検索する．現実の利用では，言語ごとに量や質の異なる特許が混在するデータベースが対象となる．しかし，研究開発の過程においては日英特許に対して均一な環境を用意した．具体的には，対訳関係にある日英特許公報（全文）と（財）日本特許情報機構で作成した特許抄録を用いた[2]．

抄録は，日英それぞれについて1995～1999年の5年分（約175万件）を収録している．公報は，資源の制約等の問題から5年分全てを対象とはせずに，1995～1999年の中でパテントファミリーを構成する日本と米国の公報（それぞれ約32,000件）を収録した．パテントファミリーについては3章で説明する．

各データベースは，オフラインで索引付けする．日本語文書は「茶筌」を用いて単語に分割し，名詞を索引語として抽出した．英語文書からは，WordNetの不要語リストと品詞情報を利用して，名詞を索引語として抽出した．翻訳された質問からも，同様の処理によって単語を抽出し，索引と照合する．また，本システムは確率型の検索手法[4]を用いて質問に対する各文書の適合度を計算し，適合度が高い文書から出力する．

最後に，提示部で検索文書を効果的に提示して利用者の閲覧効率を向上させる．そこで，ノヴァの機械翻訳システム「Transer」を用いて検索文書を利用者の母国語で提示する．また，文書分類などによって検索結果を概観できるような工夫についても検討している．

3 対応特許に基づく対訳抽出

パリ条約（4条）の優先権制度では，同盟国にした最初の出願に基づき，それと同一の発明について一定期間内に他の同盟国への出願を認めている．実際には，第2国への出願内容は最初の出願書類によって構成部分が明らかにされたものであればよいので，両者は完全に同一内容ではない．しかし，非常に類似した内容である．

このように，同一の発明に対して複数国に出願された特許のグループをパテントファミリーと呼び，パテントファミリーを構成する特許を対応特許と呼ぶ．

同一の発明を複数国に出願する方法には，優先権制度を利用する以外にも，各国への個別出願や国際出願がある．しかし，これらの方法で出願された特許に関しては，対応特許を特定することが容易ではない．それに対して，優先権制度に基づいて出願された場合は，特許に固有の優先権主張番号によって対応特許を機械的に特定できる．

2章で説明したように，我々は1995～1999年に日本で出願された特許175万件のうち，32,000件に関して米国の対応特許を特定し，収集した．

こうして収集したパテントファミリー（日米対応特許）は一種の対応付け多言語コーパスである．自然言語処理の研究では，対応付けされたコーパスから単語や複合語の対訳を自動抽出する手法が提案されている．そこで，これらの手法を応用してパテントファミリーから新語の対訳を自動抽出し，質問翻訳に用いる対訳辞書を更新する機能を実装した（図1）．

既存の対訳抽出の研究では，文単位の対応が付いたコーパスに基づく手法が多い[5, 9]．しかし，本研究が対象とする日米対応特許は文対応が付与されていない．そもそも，日本語と英語では本質的に文単位の対応付けが困難である．そこで，特許の文書構造に基づいて，文書全体よりも細かな単位で対応付けを行った．

特許には「発明の名称」「請求項」「発明の属する技術分野」「要約」などの様々な項目（フィールド）がある．しかし，フィールドによっては，対応関係にある日米特許の片方だけに存在する場合もある．32,000件の対応特許を分析した結果，少なくとも「発明の名称」と「要約」は一貫して存在することが分かった．

そこで，まず「発明の名称」と「要約」からフィールド単位で日英対応を抽出する．次に，フィールドを疑似的に長い一文と見なし，既存の手法[9]を用いて対訳を抽出した．定量的な評価は行っていないものの，辞書更新に有効な対訳を抽出できる見込みを得ることができた．

4 おわりに

本研究では，外国で出願された特許の検索を容易にするために，言語横断検索を応用した特許検索システムを

[1] http://www.nova.co.jp/

[2] 英語抄録は特許庁からPAJ (Patent Abstracts of Japan) としてCD-ROMで配布されている．

開発した．本システムは，利用者が入力した質問を対象文書の言語に翻訳して検索を行い，さらに検索文書を利用者言語に機械翻訳することで閲覧を支援する．

また，特許文書に含まれる新語（辞書未登録語）に対処するために，パテントファミリーの対応特許に基づいて対訳を自動抽出した．対応特許は今後も定常的に出願される確かな見込みがあり，本手法は実用的である．

今後は試験運用等を通してシステムをさらに拡張することや（財）日本特許情報機構が運用している特許検索サービス「PATOLIS」との連携を検討している．

謝辞

専門用語辞書および Transer 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] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[2] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, (To appear).

[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. In *ACM SIGIR Workshop on Patent Retrieval*, 2000.

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[5] 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.

[6] 石川徹也, 藤井敦, 樋口重人, 福井雅敏. 特許出願中「発明の名称: テキスト検索システム」.

[7] 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.

[8] 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.

[9] 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.

[10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム『茶筌』version 2.0 使用説明書. Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999.