特許検索と特許翻訳を指向したテストコレクションの構築研究

筑波大学大学院図書館情報メディア研究科准教授
藤井　敦

PROFILE 1998年東京工業大学大学院博士課程修了。博士（工学）。現在、筑波大学大学院准教授。2003年IPAから「天才プログラマー／スーパークリエータ」を受賞。自然言語処理、情報検索、音声言語処理の研究に従事。

fujii@slis.tsukuba.ac.jp

1 はじめに

情報検索や自然言語処理などの言語情報処理に関する研究では、「情報要求」、「言葉の意味」、「感情」といった、厳密な定義が極めて困難な概念を研究の対象としている。しかし、科学や工学における一つの研究分野として言語情報処理を位置付けるためには、問題の定式化や評価において、学問として要求される水準を満たす必要がある。すなわち、学術研究としての実証性、客観性、再現性が求められている。

事実、言語情報処理の研究において評価の重要性が増している。提案した手法の有効性を評価実験によって証明し、さらにその評価に対する信頼性について考察しなければ、高水準の国際会議や論文誌に採択されることは難しくなってきている。

そこで、複数の研究者が共有できる評価基盤としてのベンチマーク＝テストコレクションが重要性を増している。テストコレクションは大規模でかつ再利用可能であることが好ましい。このようなテストコレクションを組織的に構築するために、評価ワークショップという活動形態が存在する。評価ワークショップでは、複数の研究グループが協調と競争を通して問題設定、テストコレクション、評価方法を開発していく。

筆者らは、国立情報学研究所（NII）が主催する評価ワークショップ「NTCIR」において、特許情報処理を対象としたテストコレクションの構築研究を行っている。特許検索は長い歴史を持つ商用アプリケーションである。しかし、言語情報処理において特許が対象とされることは稀である。特許請求項の記述形式が日常言語と異なり、また請求内容の解釈に法律知識が必要なために、研究者にとって特許は馴染みが薄いためである。他方において、近年は知的な創造の成果を活用して産業の国際競争力を強化する動きがある。そこで、特許を研究対象として扱いながら、特許情報処理の関連技術を発展させ、その成果を社会に還元することには意義がある。

本稿は、NTCIRワークショップにおける筆者らの研究活動とその成果について解説する。

2 NTCIRワークショップにおける活動の概要

ある発明が特許として成立し、その権利が消滅する過程では様々な調査が行われる。調査の目的に応じて、性質の異なる特許検索が必要になる。代表的な調査として、技術動向の調査や特許庁の審査官が行う実体審査などがある。

調査の目的によって、調査対象やシステムに要求される性能（先願特許を1件でも見つければよいのか、それとも関連する特許を網羅的に見つけるのか）などが異なる。そこで、汎用的なテストコレクションを構築することは容易ではない。

NTCIRは1年半の周期で開催されるワークショップである。ただし、研究発表だけの場ではない。オーガナイザから提供されたデータを用いて、参加者が共通の「タスク」を実行し、互いのシステムを比較評価するための

場である。タスクには、情報検索、質問応答、自動要約などがある。

筆者らは、NTCIRワークショップにおいて「特許検索タスク」を運営し、1年半ごとに目的を段階的に変化させながら、様々な特許検索テストコレクションを構築した。

1回のワークショップは概ね以下の手順で行う。

(1)文書データの配布（オーガナイザ → 参加者）
(2)課題の作成と配布（オーガナイザ → 参加者）
(3)検索結果の提出（参加者 → オーガナイザ）
(4)検索結果の評価（オーガナイザ → 参加者）
(5)成果報告会（オーガナイザ、参加者）

こうした一連の活動を通して、最終的に以下の情報を含むテストコレクションが構築される。

・検索課題： ユーザの情報要求に関する記述
・文書集合： 検索対象
・適合判定： 各検索課題に対する正解文書一覧

NTCIRワークショップの参加者は情報検索や自然言語処理の研究者であり、特許検索の専門家ではない。学術研究と実システム開発のバランスを保つためには、特許に対する参加者の知識を深める必要がある。そこで、特許業界の専門家（特許庁や日本知的財産協会の関係者、弁理士など）によるチュートリアルを複数回企画した。

NTCIR-3では技術動向調査を目的とした。 NTCIR-4とNTCIR-5では無効資料調査を目的とした。NTCIR-5では、文書単位の検索に加えてパッセージ（段落）単位の検索も行った。検索以外の目的として、NTCIR-4では「特許マップの自動生成」、NTCIR-5では「Fタームを用いた特許分類」も行った。NTCIR-6では米国特許庁（USPTO）から発行された特許を対象とした検索を行った。

表1にNTCIR-3〜6の概要を示す。表1の「文書集合」に示したように、回を重ねるたびに文書データの規模を段階的に増やしていった。他方において、文書データの規模が大きくなると適合判定の負荷が大きくなる。NTCIR-3では日本知的財産協会の専門家が適合判定を

表1　NTCIR-3〜NTCIR-6の概要

<table>
<tr><th></th><th>NTCIR-3</th><th>NTCIR-4</th><th>NTCIR-5</th><th>NTCIR-6</th></tr>
<tr><td>調査目的</td><td>技術動向調査</td><td colspan="3">無効資料調査</td></tr>
<tr><td rowspan="2">文書集合</td><td colspan="3">日本公開公報</td><td rowspan="2">日本公開公報
10年分,米国
特許10年分</td></tr>
<tr><td>2年分</td><td>5年分</td><td>10年分</td></tr>
<tr><td rowspan="2">適合判定</td><td colspan="2">知財の専門家</td><td colspan="2" rowspan="2"></td></tr>
<tr><td></td><td colspan="2">特許庁審査官（拒絶の引例）</td></tr>
<tr><td>その他の
サブタスク</td><td></td><td>特許マップ
自動生成</td><td>Fターム分類
パッセージ
検索</td><td></td></tr>
</table>

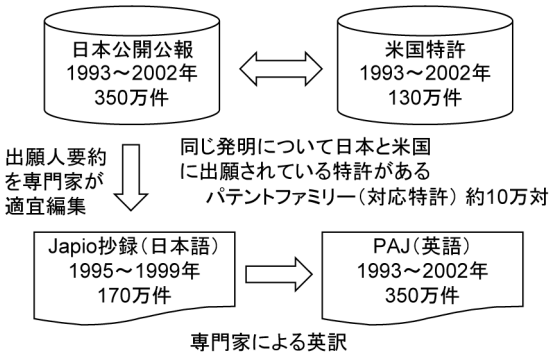

図1　NTCIRで配布している特許データ

行った。しかし、NTCIR-4からは特許庁に拒絶された特許を検索課題として利用し、その特許を拒絶する根拠となった別の特許（引例）を正解として用いることで適合判定の負荷を削減した。米国特許を対象とした検索では、検索課題の特許で引用されている特許を正解として利用した。そのため、引用文献は削除した上で検索課題として利用した。また、特許抄録データを訓練データとして配布した。

NTCIR-3〜6の成果によって、現在NIIから配布しているデータの関係を図1に示す。具体的には、「日本公開公報10年分」、「Japio抄録」、「PAJ」、「米国特許」で構成されている。Japio抄録は日本公開公報の出願人要約を専門家が適宜編集した和文抄録である。PAJはJapio抄録を専門家が翻訳した英文抄録である。米国特許はUSPTOから発行された特許である。さらに、日本公開公報と米国特許には同じ発明について日本と米国に出願された対応特許（パテントファミリー）が存在する。

NTCIR-3～6で構築したテストコレクションは、NIIと覚書を交わせば研究目的で利用することができる[1]。

NTCIR-3～5における活動の詳細は、Japio 2006 Year Book[2,3]を参照されたい。NTCIR-6の米国特許検索では、引用関係を文書間のリンク構造と見なして、テキスト検索とリンク解析を統合した検索手法[4]が提案された。また、海外論文誌において特許情報処理に関する特集号を企画した[5]。当特集号は特許情報の検索、分類、マイニングに関する優れた研究論文を掲載しており、NTCIR特許検索タスクに参加した研究グループの成果も報告されている。

3 NTCIR-7特許翻訳タスク

NTCIR-3～6における特許検索タスクを通して、特許検索と特許分類に関する大規模なテストコレクションを構築し、さらに種々の知見を得ることができた。本稿執筆当時は、NTCIR-7のタスク参加者を募集するための準備中である。NTCIR-7では、特許情報処理に関する新たな挑戦として、「特許翻訳タスク」と「特許マイニングタスク」を行う。ここでは、特許翻訳タスクについて説明する。

特許翻訳には、機械翻訳の研究開発という学術的な意義がある。また、外国特許の検索や特許情報の翻訳といったサービスにつながる点において産業上の価値がある。

近年、統計的な機械翻訳（Statistical Machine Translation: SMT）の技術が急速に発展している。SMTは、原言語と目的言語の対訳テキストから単語や句の単位で翻訳に関する統計モデルを事前に学習する。そして、翻訳対象の文が入力されると、事前に学習したモデルに従って単語や句の単位で目的言語に翻訳する。さらに、目的言語として自然な語順に並べ替える。図2にSMTの概要を示す。

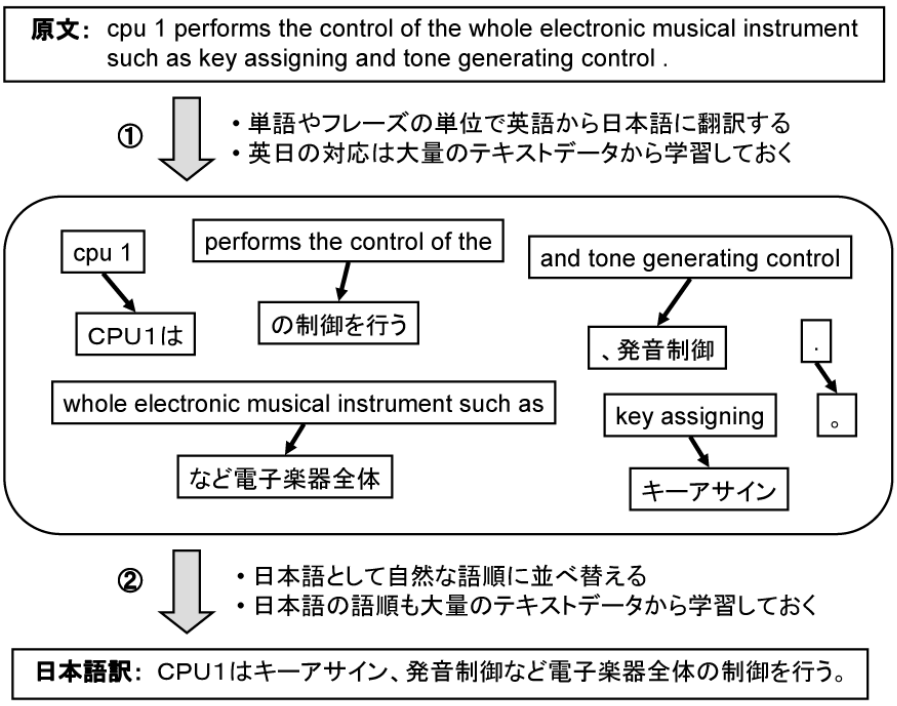

図2　統計的な機械翻訳の概要

SMTが発展している理由は、原言語と目的言語の対訳テキストが大量に入手可能になったことである。また、計算機の性能が向上したために、大量のテキストから統計モデルを効率的に構築することが可能になったためである。

図1に示したように、NTCIR-3～6の成果によって、日本語と英語の対応特許を研究目的で利用することが可能になった。筆者らは、この対応特許から日本語と英語の対訳文を約180万対収集している。この対訳文は日本語を対象とした既存の対訳テキストを凌駕する規模であり、日本語を対象としたSMT研究の発展に貢献することが期待できる。筆者らの実験では、英語とフランス語のSMTに匹敵する翻訳精度が得られている[6]。

4 おわりに

NTCIR-3～6で行った特許検索タスクの成果とNTCIR-7で進行中の特許翻訳タスクについて解説した。特許検索タスクで構築したデータが統計的な機械翻訳に有用なデータであることが分かり、NTCIR-7の特許翻訳タスクへと発展した。

特許情報処理の研究では、特許情報に関する知識や大量の特許データを入手するために、評価ワークショップにおけるチームワークが有用だった。今後もNTCIRにおける活動を通して特許情報処理の発展に貢献していきたい。

謝辞

特許検索タスクの運営は、岩山真准教授（東京工業大学／日立製作所）、神門典子教授（国立情報学研究所）と共同で行いました。特許翻訳タスクの運営は、山本幹雄准教授（筑波大学）、内山将夫氏（NICT）、宇津呂武仁准教授（筑波大学）と共同で行っています。

参考文献

[1] http://research.nii.ac.jp/ntcir/index-ja.html
[2] 藤井敦. NTCIRにおける特許検索テストコレクションの構築研究. Japio 2006 Year Book, pp.102-107, 2006.
[3] 岩山真. 特許マップ自動作成を目指した評価ワークショップ. Japio 2006 Year Book, pp.108-111, 2006.
[4] Atsushi Fujii. Enhancing Patent Retrieval by Citation Analysis. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.793-794, 2007.
[5] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Introduction to the special issue on patent processing. Information Processing & Management, Vol.43, No.5, pp.1149-1153, 2007.
[6] 内山将夫, 山本幹雄, 藤井敦, 宇津呂武仁. 特許情報を対象とした機械翻訳 ―共通基盤による評価タスクを目指して―. 電子情報通信学会技術研究報告, NLC2007-23, pp.133-138, 2007.