特許情報処理を指向したテストコレクションの構築：情報検索と自然言語処理の融合を目指して

藤井 敦（筑波大学），難波 英嗣（広島市立大学），岩山 真（東京工業大学/日立製作所），神門 典子（国立情報学研究所），内山 将夫（情報通信研究機構），山本 幹雄（筑波大学），宇津呂 武仁（筑波大学），橋本 泰一（東京工業大学）

E-mail: fujii@slis.tsukuba.ac.jp

情報検索や自然言語処理に関する技術を体系的に評価するためには，ベンチマークとして研究者が共有できる大規模なテストコレクションが必要である．本稿は，NTCIR ワークショップにおいて構築している特許情報処理テストコレクションについて，検索，分類，機械翻訳，マイニングの観点から解説する．検索では技術動向調査，無効資料調査，パッセージ検索を目的とし，分類では F タームに基づくカテゴリ分類を目的としたテストコレクションを構築した．現在進行中のワークショップでは，検索や分類で用いた日英特許情報を応用して，翻訳とマイニングに関するテストコレクションを構築している．翻訳では，対応特許（パテントファミリー）から抽出した文対応データをシステムの訓練や評価に用いる．マイニングでは，特許と技術論文を横断した技術動向分析を想定して，論文抄録に特許分類のカテゴリを付与することを目的とする．

Producing Test Collections for Patent Information Processing: Toward the Fusion of Information Retrieval and Natural Language Processing

Atsushi Fujii (University of Tsukuba), Hidetsugu Nanba (Hiroshima City University), Makoto Iwayama (Hitachi, Ltd. / Tokyo Institute of Technology), Noriko Kando (National Institute of Informatics), Masao Utiyama (National Institute of Information and Communications Technology), Mikio Yamamoto (University of Tsukuba), Takehito Utsuro (University of Tsukuba), Taiichi Hashimoto (Tokyo Institute of Technology)

E-mail: fujii@slis.tsukuba.ac.jp

To evaluate technologies for information retrieval and natural language processing systematically, sharable large test collections as benchmark data are needed. This paper describes the test collections for patent information processing at the NTCIR workshop from retrieval, classification, machine translation, and mining perspectives. For the retrieval task, test collections for technology survey, invalidity search, and passage retrieval were produced. For the classification task, test collections for the F-term patent classification system were produced. In the current workshop, applying the patent documents in Japanese and English that were used for the retrieval and classification tasks, test collections for machine translation and mining are being produced. For the translation task, sentence-aligned data extracted from patent families are used for training and evaluation purposes. For the mining task, aimed at the analysis of technology trends across patents and technical papers, the purpose is to categorize technical abstracts based on a patent classification system.

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．評価実験によって提案した手法の有効性を証明し，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．大規模でかつ再利用可能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，評価方法を開発する．

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 における筆者らの研究活動とその成果について説明する．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「研究課題（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，第 3 回 NTCIR ワークショップ（NTCIR-3）から「特許検索タスク」を開始し，NTCIR-6 まで行った．本稿執筆当時は，NTCIR-7 のタスク参加者を募集中である．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．2 章で特許検索タスクについて説明し，3 章と 4 章で 2 つの新しいタスクについてそれぞれ説明する．

2　特許検索タスク

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，特許検索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．筆者らは，NTCIR ワークショップが回数を重ねるたびに，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 [7, 8]．NTCIR-4 と NTCIR-5 では無効資料調査を目的とした [2, 3]．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 [5].

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 と 4 では日本知的財産協会の専門家が適合判定を行った．しかし，NTCIR-4 の一部，NTCIR-5，NTCIR-6 で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．

米国特許を対象とした検索タスク以外では，検索課題は日本語である．しかし，NTCIR-3 の検索課題は，英語，韓国語，簡体中国語，繁体中国語に人手で翻訳した．NTCIR-4 の検索課題は，無効化の対象となる請求項のみを英語，韓国語，簡体中国語，繁体中国語に人手で翻訳した．NTCIR-5 の検索課題は，無効化の対象となる請求項のみを英語に翻訳した．外国語に翻訳された検索課題と日本語の特許文書データを用いることで，言語横断の特許検索システムを評価することができる．

各回における特徴的な研究成果について説明する．NTCIR-3 では，新聞記事に掲載された技術内容に関する特許を検索することが目的であった．新聞と特許において単語の出現頻度が異なることを利用した Term Distillation [6] が提案された．また，オーガナイザ自身が実験を行い，特許検索に関する種々の知見を得ることができた [7].

NTCIR-4 では人間が適合判定した少数の課題を用いたのに対して，NTCIR-5 と 6 では，引例を正解とした大量の課題を用いた．異なる課題を用いた実験結果を比較した結果，参加チーム間の優劣には差がないことが分かった．すなわち，引例を正解として大量の課題を自動的に構築することに意義があった．他方において，同一チームにおいてパラメタ調整などの軽微な変更をしたシステムどうしを比較した場合には，課題の種類によってシステム間の優劣が変動することがあった [10].

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"></td></tr>
<tr><td></td><td colspan="3">特許庁審査官（拒絶の引例）</td></tr>
<tr><td rowspan="2">その他の
サブタスク</td><td rowspan="2"></td><td rowspan="2">特許マップ
自動生成</td><td>Fターム分類</td><td rowspan="2"></td></tr>
<tr><td>パッセージ
検索</td></tr>
</table>

NTCIR-5 の特許分類では，特許明細書が請求項や実施例などの項目で構造化されていることから，構造情報を用いた分類手法 [9] が効果的だった．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 [1] が効果的だった．

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].

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 [4]．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 特許翻訳タスク

特許翻訳には，機械翻訳の研究という学術的な意義がある．また，海外に出願された特許を検索したり，海外に出願するために日本語の特許を翻訳するための基盤技術になるため，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．

[1] http://research.nii.ac.jp/ntcir/index-ja.html

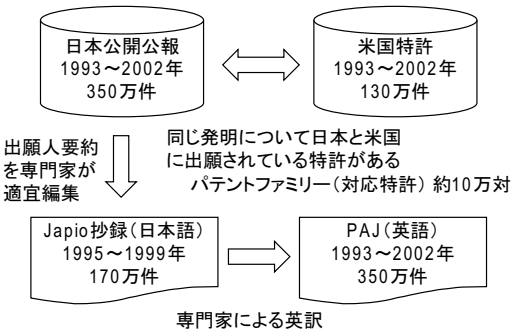

図 1: NTCIR で配布している特許文書データ

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．SMT が発展している理由は，原言語と対象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 万対収集した．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 の研究に貢献することが期待できる．BLEU という評価尺度を用いた場合には，英語とフランス語の SMT と同程度の翻訳精度が得られた [12].

特許翻訳タスクにおける参加者の目的は，日本語もしくは英語のテスト文をもう一方の言語に翻訳することである．テスト文の件数は千の単位で用意し，翻訳結果の評価として以下の 3 通りを予定している．

- 訳質による評価

  人手判定による評価と BLEU による自動評価
- 外因による評価

  言語横断特許検索の精度による評価

外因による評価では，特許検索タスク（2 章）で作成したテストコレクションを利用する．具体的には，人手で英語に翻訳された請求項を入力として，日本語の特許公報を検索する．言語横断特許検索を実現するために，英語の検索課題を日本語に機械翻訳し，日本語どうしの単言語検索を実行する方法がある．参加者は，検索課題を機械翻訳し，オーガナイザは共通の検索エンジンを用意する予定である．

4 特許マイニングタスク

4.1 背景と目的

近年，大学研究者自身が関連論文だけでなく関連特許について情報を検索したり，特許を出願したりする機会が増えている．2007 年 5 月に政府の知的財産戦略本部が発表した「知的財産権推進計画 2007」においても，大学研究における特許情報の重要性が謳われている．この計画で，大学研究者の利用を想定した特許・論文情報統合検索システムの整備が含まれていることからも，このような傾向は今後さらに強まっていくことが予想される．

特許と論文を検索するのは大学研究者に限った話ではない．例えば，特許庁の審査官は，出願された技術が特許権の取得に該当するかどうか判断するために，過去に同様の特許が出願されたり論文が発表されたりしていないか調査する．これは，一般に無効資料調査と呼ばれている．同様の調査は，サーチャーと呼ばれる専門の担当者が審査官による審査を経た出願技術を再調査し，競合する他者の権利を無効化するために民間企業の社内で行われることもある．そこで，NTCIR-7 特許マイニングタスクでは，特許と論文を対象にした検索や技術動向分析など，様々な目的に利用可能な言語処理技術の開発を目指す．

論文と特許を対象にしたタスクのひとつとして，本タスクでは日本語または英語論文抄録を特許分類体系のひとつである「国際特許分類」(International Patent Classification: IPC) に自動分類することを目的とする．特許を分類するタスクとして，NTCIR-5 と 6 における F ターム分類タスクが実施された．今回のタスクでは分類対象となる文書が論文に変わるため，特許と論文で使われる用語の違いについて新たに検討する必要がある．

特許では請求範囲を広く確保するため，一般性の高い特許用語を用いて記述する傾向がある．また，特許では学術用語よりも多様な表現が用いられることが多い．例えば,「機械翻訳」という学術用語に対する特許用語は「機械翻訳」の他にも「自動翻訳」「言

語変換」などがある．このため，単純に表層的な単語の一致度を用いる従来の分類モデルでは十分な分類精度が得られるとは限らない．本タスクでは，このような論文と特許の用語の使われ方の違いを吸収できる分類や検索のための基礎技術の確立を目指す．

4.2 関連研究

ジャンル横断検索や文書分類に関して，これまでにいくつかの先行研究がある．NTCIR-3 で実施された技術動向調査タスク [8] では，与えられた新聞記事と関連する特許を検索する課題が設定され，Itoh ら [6] は Term Distillation という手法を提案した．例えば,「社長」という単語は新聞記事中では高頻度で出現するのに対して，特許中では出現頻度が低い．このため,「一般的な用語ほど重要ではない」という考えに基づいて単語の重要度を計算する TF.IDF 等の手法を用いると，同じ単語でも新聞記事と特許では重要度が異なる。Itoh らは，単語の新聞記事集合中での出現頻度と特許中での出現頻度の違いを考慮して重み付けを行うことで，ジャンルを横断した文献の対応付けを行った．

特許と論文を横断的に検索するための研究として安善ら [13] の研究がある．近年，特許中で関連論文を引用し，また論文において関連特許を引用するケースが増えている．このような文書間の引用関係をたどれば，論文や特許と関連する文書を集めることができる．安善らは特許中で関連文献が引用される「従来の技術」という項目を解析して引用論文の書誌情報を抽出し，特許と論文間の引用関係を解析した．現状では，特許中の引用文献の中で論文が占める割合と論文中の引用文献の中で特許が占める割合は数パーセント程度であるため，あるテーマに関する特許と論文を網羅的に収集するためには引用関係をたどるだけでは十分ではない．

この問題に対処するため，釜屋ら [11] は論文用語を特許用語に自動変換する手法を提案した．例えば，論文用語「フロッピーディスク」を特許用語「磁気記録媒体」に自動変換する．釜屋らは，論文用語の特許用語への変換を実現するために，特許と論文間の引用関係に着目した．一般に，引用関係にある特許と論文は，同一トピック（分野）である可能性が高い．そこで，ある用語を表題に含んだ論文を収集し，それらと直接引用関係にある特許から特許のトピックを示す用語を抽出すれば，入力された論文用語に関連する特許用語の変換が実現できる．

以上は，文書のジャンルを横断した情報アクセス技術の一例である．特許マイニングタスクでは，参加者による文書分類技術やジャンル横断技術の提案を推奨している．

4.3 テストコレクションの構築方法

特許マイニングタスクでは日本語または英語論文抄録を特許分類体系のひとつである IPC に自動分類する．IPC の概要について説明する．IPC は，特許文献の技術内容によって上から順に「セクション」,「クラス」,「サブクラス」,「メイングループ」,「サブグループ」の 5 階層で構成されている．本タスクでは，最下層の「サブグループ」を論文抄録に付与することを目的とする．

次に，テストコレクションの構築方法について説明する．論文抄録を IPC に自動分類するタスクを実施するためには，正解データとして，論文抄録に IPC コードを付与する必要がある．国際特許分類第 6 版ではサブグループ数が 50,000 以上あり，論文抄録に人手でサブグループを付与する作業は，知識と経験を持った専門家でも作業に時間がかかる．

そこで，以下の手順で正解データの作成を行う．ある発明を論文等で発表すると，その時点で発明の内容は公知になるとみなされ，特許を受ける権利が失われる．しかし，日本では特許庁の承認を受けた学協会，団体での発表，出版であれば，発表後 6ヶ月の間に限り，特許出願を行うことができることが特許法第 30 条で規定されている．特許法 30 条が適用される特許には，特記事項として，論文を発表した論文誌名，発表した日付などが特許中に記載されるため，これを手がかりに効率的に正解データを作成することができる．例えば，以下のように記載される．

> 【新規性喪失の例外の表示】特許法第３０条第１項適用申請有り２０００年３月１４日 社団法人情報処理学会発行の「第６０回（平成１２年前期）全国大会講演論文集（４）」に発表

各特許に付与されている IPC コードを「新規性喪失の例外の表示」の欄に記載されている論文の IPC コードと見なすことで，正解データを効率的に作成する．特許マイニングタスクで用いる公開公報（1993 年〜2002 年）約 350 万件には,「新規性喪失の例外の表示」を含んだ公報が約 9,000 件存在する．これらから，発表日と論文誌名を抽出する．この欄には，論

文表題や発表者名が記載されることはほとんどない．しかし，多くの場合，特許の発明者と論文の発表者が同一（連名の場合は共通する人名がある）であるため，特許の発明者を論文の著者として抽出する．

次に，これらの情報を抄録データベースの書誌情報と照合し，対応付けを行う．抄録データベースとして，NTCIR-1 と 2 の言語横断検索タスクで用いられた約 46 万件の抄録データを用いる．ここで，ひとつの「新規性喪失の例外の表示」に記載された論文に対して，平均 6 件程度まで対応付けの候補となる抄録を絞り込むことができる．最終的に，これらの候補を人間が確認し，該当する抄録を特定する．抄録に IPC を付与するためには専門的な知識が必要である．しかし,「新規性喪失の例外の表示」に記載された論文と候補の論文が同一であるかどうかの判定は，専門知識がなくても可能である．

5 おわりに

NTCIR-3 から NTCIR-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と特許マイニングタスクについて解説した．NTCIR では，特許検索という古典的なアプリケーションを軸としつつも，情報検索と自然言語処理の融合を常に指向しながら研究を進めてきた．その結果，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．また，特許情報の検索，分類，解析を総合的に扱う特許マイニング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効果的だった．今後も特許情報処理を通して，情報検索と自然言語処理の発展に貢献していきたい．

参考文献

[1] Atsushi Fujii. Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 793–794, 2007.

[2] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Test collections for patent-to-patent retrieval and patent map generation in NTCIR-4 workshop. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1643–1646, 2004.

[3] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 671–674, 2006.

[4] 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.

[5] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp. 359–365, 2007.

[6] Hideo Itoh, Hiroko Mano, and Yasushi Ogawa. Term distillation in patent retrieval. In *Proceedings of the ACL-03 Workshop on Patent Corpus Processing*, pp. 41–45, 2003.

[7] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing & Management*, Vol. 42, No. 1, pp. 207–221, 2006.

[8] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano. Overview of patent retrieval task at NTCIR-3. In *Proceedings of the ACL-03 Workshop on Patent Corpus Processing*, pp. 24–32, 2003.

[9] Jae-Ho Kim and Key-Sun Choi. Patent document categorization based on semantic structural information. *Information Processing & Management*, Vol. 43, No. 5, pp. 1200–1215, 2007.

[10] Hisao Mase and Makoto Iwayama. NTCIR-6 patent retrieval experiments at Hitachi. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pp. 403–406, 2007.

[11] 釜屋英昭, 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山. 特許, 論文間の引用情報を用いた論文用語の特許用語への変換. 情報処理学会研究報告, 2007-NL-178, pp. 97–102, 2007.

[12] 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.

[13] 安善奈津美, 難波英嗣, 相沢輝昭, 奥村学. 論文データベースを統合した検索環境の構築. 言語処理学会第 12 回年次大会発表論文集, pp. 743–746, 2006.