レファレンス事例の分析結果に基づいた論文検索システムの構築

難波 英嗣

広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: nanba@hiroshima-cu.ac.jp

あらまし 筆者はこれまでに，第一回 NTCIR ワークショップ情報検索タスクの検索課題およびレファレンス協同事例や Yahoo! 知恵袋の回答欄に学術論文データベースへのリンクがあるものを対象に，論文調査目的のレファレンス事例を分析している．本研究では，この分析結果に基づいて構築した論文検索システムについて述べる．

キーワード 論文検索, 質問分析，レファレンス

1. はじめに

学術論文検索では，情報要求を満たす論文を網羅的に探す必要がある．しかし，初学者にとって，それはなかなか容易ではない．そこで，著者らは，初学者が論文を検索する際，適切な検索語を推薦することで，検索作業を支援するシステムの開発に取り組んでいる．

このような検索支援システムを開発するには，「どのような情報要求に対し，検索者がどのようなプロセスで検索しているのか」を調査する必要がある．これまでに，著者は，論文調査目的のレファレンス事例を大量に収集し，情報要求およびプロセスを体系的に分析している[1]．本稿では，この分析結果に基づいて構築した論文検索システムについて述べる．

本論文の構成は以下のとおりである．次節では，論文検索および情報要求の分類に関する関連研究について述べる．3 節では，論文調査目的のレファレンス事例の収集および分析結果について報告し，4 節で本稿をまとめる．

2. 関連研究

2.1 学術論文を対象とした情報検索

学術論文を対象とした情報検索の研究は古くから行われており，Cranfield，Medlars，CACM，NPL，LISA など，数多くのテストコレクションが作られてきた．より最近の研究プロジェクトとして，TREC で行われた Chemical IR トラックがある[2]．このトラックの中で実施された技術サーベイタスクでは，化学のある分野の動向を知るために必要な論文と特許を検索することを目的としている．このタスクでは，例えばある化合物の画像ファイルおよび構造ファイルが検索システムの入力として与えられ，その化合物に関する特許と論文を検索することが求められる．

上述のプロジェクトは英語を対象にしているが，日本語論文を対象にしたものに，第一回および第二回 NTCIR ワークショップで行われた情報検索タスクがある[3, 4]．本研究では，NTCIR-1 情報検索タスクのデータを分析に用いる．

2.2 情報要求の分類

渡邊ら[5]は，QA サイトの質問を，以下の 5 つのタイプに分け，機械学習に基づく手法で，自動的に分類する手法を提案している．

- **事実**：事象の定義，真実，客観的な理由や手法を問う質問
- **根拠**：客観的な根拠，理由を問う質問
- **経験**：回答者の経験や体験がなければ回答できない質問
- **提案**：問題の解決方法を問う質問や情報提供を依頼する
- **意見**：推測，嗜好など，主観的に回答をしてよい質問

上述のものとは定義は若干異なるものの，林ら[6]も，渡邉らと同様に，質問を「事実」,「根拠」,「経験」,「提案」,「意見」の 5 つのタイプに分類している．本研究でも，QA コンテンツを分類するという点ではこれらの研究と共通するが，論文調査目的のレファレンス事例に対象を限定し，学術に特化したカテゴリを扱う点が異なる．

3. レファレンス事例の分析結果に基づいた論文検索システムの構築

3.1 レファレンス事例の分析

著者は，レファレンス協同事例データベース(レファ協)[1]において，回答欄に NII 学術情報ナビゲータ(CiNii)へのリンクがあるものは論文調査目的のレファレンス事例と考えて，3,032 件の質問と回答の対を収集して

[1] http://crd.ndl.go.jp/reference/

いる．この回答には，「回答プロセス」という項目がある．これは，司書が回答となる論文をどのような手順で見つけたのかをまとめたものである．筆者は，この回答プロセスを分析し，以下の 4 種類のカテゴリ(a)～(d)に分類している．

- **(a) より適切なクエリへの修正**：クエリ「頭蓋骨早期癒合症」を，より一般的な表現である「頭蓋骨縫合早期癒合症」や「頭蓋縫合早期癒合症」に修正
- **(b) 同義語，上位語拡張**：「玄米」を，その上位語である「コメ」に拡張して検索．あるいは，上位語を使う代わりに検索対象文書に付与されているカテゴリで結果を絞る．
- **(c) 書籍を対象にした検索**：「家庭で出た生ごみをダンボールに入れて堆肥にする方法」など，一般性の高い手順について調べる時には，書籍を対象．
- **(d) 属性語の拡張・省略**：「肝臓移植のドナーのデメリットについて書かれた本はあるか」という質問に対し，属性語である「デメリット」をクエリから除外して検索．あるいは，属性語「吸水率」を「吸水」や「浸漬」などに言い換えて検索．

本研究では，これらのカテゴリの一部を論文検索システムのモジュールとして実装した．次節では，これらのモジュールについて述べる．

3.2 論文検索システムの構築

(a) より適切なクエリへの修正

文書集合から同義語を自動的に収集する研究は数多く行われているが，その代表的な手法のひとつに統計的機械翻訳技術を用いるものがある．例えば，「自動翻訳」の英訳が“machine translation”，「機械翻訳」の英訳も“machine translation”であるとき，英訳が共通である「自動翻訳」と「機械翻訳」は同義語であると考えられる．この考え方に基づき，統計的機械翻訳技術を対訳コーパスに適用して得られた翻訳モデル(フレーズテーブル)から自動的に同義語対を得ることができる．ここで，もし，“machine translation”から「機械翻訳」への翻訳確率が“machine translation”から「自動翻訳」へのものより高ければ，「自動翻訳」よりも「機械翻訳」の方がより一般的な表現であると考えられる．

本研究では NTCIR-1[3]で提供された論文データにおいて，ひとつの論文に日本語と英語のキーワードが同数付与されているものを翻訳対とみなし，上記の考え方に基づき，より適切なクエリに修正するモジュールを構築した．なお，本研究では，ある日本語と英語のキーワード対の論文データベース中における頻度を翻訳確率の代わりに利用した．図 1 は，「自動翻訳」を入力とした時の本モジュールの出力結果である．

416	機械翻訳
2	機会翻訳
2	翻訳システム
2	*自動翻訳
2	翻訳
1	自動翻訳(機械翻訳)
1	(機械)翻訳
1	機能語表現
1	日英機械翻訳
1	英日機械翻訳

図 1 「自動翻訳」をより適切なクエリに修正した結果 (入力語と同じ用語の前には*が付与される)

図 1 において，「自動翻訳」の英訳として“machine translation”が論文データベース中に 2 回出現していた．一方，“machine translation”の和訳は「機械翻訳」の頻度が最も高く 416 回であった．このことから，「自動翻訳」という表現は「機械翻訳」の方がより一般的であると判断される．

(b) 同義語，上位語拡張

(a)で述べたモジュールを用いれば．検索クエリを同義語で拡張することができる．ただし，図 1 を見るとわかるとおり，低頻度語の中には「日英機械翻訳」のような「自動翻訳(機械翻訳)」の下位語や，そもそも同義語として不適切な「機能語表現」といった表現も含まれるため，一定の頻度以上の語句のみを利用するなどの処理が必要である．

上位語拡張に関しては，「A 等の B」といった定型表現パターンを用いてテキストデータベースから上位，下位関係にある用語を抽出する手法が提案されており[7]，実際に特許データベースから，この定型表現パターンを用いた辞書が構築され[8]，公開されている[2]．図 2 は「機械翻訳」の上位語の検索例である．各語句の後ろの数字はコーパス中の頻度を示している．例えば，「機械翻訳等のアプリケーション」という表現が特許データベース中に 21 回出現することを意味している．

[2] http://165.242.101.30/cgi-bin/ontology/search/search.cgi

アプリケーション(21)
自然言語処理(17)
処理(10)
技術(9)
言語処理(4)

図 2　「機械翻訳」の上位語の検索例

本節ではさらに，検索対象文書に付与されているカテゴリで結果を絞る手法について述べる.本研究では，文書分類技術を用いてあらかじめ検索対象の論文を分類しておき，この分類結果を用いて検索結果の絞り込みを行う．著者らは，学術論文を以下の 5 種類のカテゴリに分類する研究を行っている.

- 科研費カテゴリ:科学研究費助成事業の研究課題を分類するための体系．4 階層からなり，2015 年時点で，最下層で 319 カテゴリ存在している．[9]
- JST 科学技術分類表：JST によって考案された科学技術文献を分類するための体系．3 階層からなり，最下層で 784 カテゴリ存在している.
- 国際十進分類法:国際書誌学会によって考案された図書分類法.
- 国際特許分類：特許を分類するための体系で，5 階層からなり，2014 年時点で，最下層で 78,773 カテゴリ存在している．[10]
- Yahoo! 知恵袋カテゴリ：Yahoo! 知恵袋の質問−回答事例を分類するためのもの.3 階層からなり，2014 年時点で、最下層で 564 カテゴリ存在している．[11]

論文を 5 種類のカテゴリに分類する理由は，科研費カテゴリや JST 科学技術分類表は研究者，国際特許分類は企業などの開発者，国際十進分類法や知恵袋カテゴリは非専門家と，それぞれ，異なる利用者を想定しているためである．このように，学術論文を様々なカテゴリに分類しておけば，論文を色んな側面から探しやすくなるだけでなく，例えば，同じカテゴリに分類された論文と特許を用いて，学術界と産業界の技術の関係性を分析するといったことも可能になる.

文書分類に用いられる技術は，SVM，k 近傍法，ロジスティック回帰などの機械学習に基づく手法を用いるのが一般的であるが，近年では，DNN(Deep Neural Network)を用いた手法も使われるようになってきている．本研究では，DNN をベースとした手法である fastText[12]を用いて，入力された文書を上述の 5 種類のカテゴリに分類するシステムを構築している.

表 1 は，NTCIR-1 の論文データベース中のすべての論文を，その概要とタイトルを用いて科研費カテゴリに分類した結果をまとめたもの(分類された論文数の多い上位 10 カテゴリ)である．NTCIR-1 の論文データは情報系の論文が数多く含まれていることが知られているが，表 1 の結果からも確認できる.

表 1　NTCIR-1 の論文データを科研費カテゴリに分類した結果

論文数	科研費カテゴリ
31,090	知能情報学
29,924	通信・ネットワーク工学
28,430	建築構造・材料
15,188	電子デバイス・電子機器
14,933	都市計画・建築計画
9,725	教育工学
9,218	地盤工学
8,730	建築環境・設備
7,626	電力工学・電力変換・電気機器
7,407	計測工学

NTCIR-1 論文検索タスクの正解データの各論文を科研費カテゴリに分類した場合，1 トピックあたりどの程度カテゴリにばらつきがあるか調査した．NTCIR-1 論文検索タスクでは検索課題が 53 トピックあり，1 トピックあたりの正解文書数は平均で 36.01 件であった．1 トピックあたりのカテゴリ数は 5.77 であり，正解文書はある程度まとまったカテゴリに属していることがわかった．このことから，検索時のカテゴリによる絞り込みは，検索精度の向上が期待できると思われる.

4. おわりに

本研究では，これまでに，著者は，論文調査目的のレファレンス事例を大量に収集し，情報要求およびプロセスを体系的に分析した結果に基づいて構築した論文検索システムを紹介した.

謝辞

本研究の一部は科学研究費補助金(基盤研究(A))(研究課題番号:15H017214)の支援を受けて行われた．Yahoo!知恵袋データは，ヤフー株式会社から提供いただいた．論文データは，国立情報学研究所および科学技術振興機構から提供いただいた．論文検索テストコレクションは，国立情報学研究所主催の第 1 回 NTCIR ワークショップのものを利用させていただいた．ここに記して謹んで感謝の意を表する.

参　考　文　献

[1] 難波英嗣，“レファレンス事例の分析による論文検索に効果的な要素の調査”，第 8 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2016)，2016.
[2] M. Lupu, F. Piroi, J. Huang, J. Zhu, and J. Tait,

"Overview of the TREC Chemical IR Track", Proceedings of the 18th Text Retrieval Conference, 2009.

[3] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka, "Overview of IR Tasks", Proceedings of the 1st NTCIR Workshop Meeting, 1999.

[4] N. Kando, K. Kuriyama, and M. Yoshioka, "Overview of Japanese and English Information Retrieval Tasks", Proceedings of the 2nd NTCIR Workshop Meeting, 2001.

[5] 渡邊直人, 島田諭, 関洋平, 神門典子, 佐藤哲司, "QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討", DEIM Forum 2011, 2011.

[6] 林秀治, 山本和英, "質問意図による QA サイト質問文の自動分類", 電子情報通信学会技術研究報告, 思考と言語, 113(82), pp. 51-56, 2013.

[7] M.A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora", Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545, 1992.

[8] H. Nanba, "Query Expansion using an Automatically Constructed Thesaurus", Proceedings of the 6th NTCIR Workshop, pp. 414-419, 2007.

[9] 福田悟志, 難波英嗣, 竹澤寿幸, "要素技術とその効果を用いた学術論文の自動分類", 日本図書館情報学会誌, Vol.63, No.3, pp. 145-162, 2016.

[10] 難波英嗣, 竹澤寿幸, "2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類", 情報処理学会論文誌データベース, Vol.2, No.3, pp. 76-86, 2009.

[11] 重田識博, 難波英嗣, 竹澤寿幸, "論文データの Yahoo!知恵袋カテゴリへの自動分類", 第 8 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2016), 2016.

[12] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification", arXiv Preprint arXiv:1607.01759, 2016.