HOKKAIDO UNIVERSITY

Title	Study on a Cross-language Search Method for the Japanese TV Guide [an abstract of dissertation and a summary of dissertation review]
Author(s)	Kiselev, Denis
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11761号
Issue Date	2015-03-25
Doc URL	https://hdl.handle.net/2115/58903
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Kiselev_Denis_review.pdf, 審査の要旨



Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　Kiselev Denis

審査担当者　主　査　教　授　荒木 健治
副　査　教　授　山本 強
副　査　教　授　長谷山 美紀

学位論文題名

Study on a Cross-language Search Method for the Japanese TV Guide
（日本語 TV ガイドにおける異言語間検索手法に関する研究）

本研究の目的は, 日本語 TV ガイドにおける検索の既存手法を以下に述べるように改善するための新たな手法を提案し, その有効性を確認することである.

日本語 TV ガイドを分析したところ,TV ガイドに含まれているローマ字文字列の内, 約 64% は英単語であり, 相当数の英単語が出現することが明らかとなった. 一方,Google 日本のカスタマイズサーチを日本語 TV ガイドのウェブサイトにおいて使用してみると, 英単語が検索対象にされていないと考えられる. なお,「Yahoo テレビ　Japan」というウェブサイト上の検索システム等を利用した検索でも, 英単語が対象外であることがわかる. このため, 当然, 検索結果として必要となる英語の情報が検出できていない. 本問題の解決に向けて, 提案手法では本来の検索語とともに, 検索語に対して WordNet から抽出した英語パラフレーズを, 検索に利用している.

上述した Google 日本や,「Yahoo テレビ　Japan」の検索システムでは, 検索語と, それに意味的に関連がある他の語を検索において使用するクエリ拡張が行われていないと考えられる. 実際に Google 日本で, そのカスタマイズサーチ機能を日本語 TV ガイドのウェブサイトに対して使用してみると, 同義語等が検出されていないので, クエリ拡張が行われていないと推測される. このようにクエリ拡張を行わなければ, 当然, 検索語句とは同じ文字ではないが同じ意味を持つ文章の抽出ができない. この問題の解決策として, 提案手法では元々の検索語と共にその日本語のパラフレーズが検索に利用されている. パラフレーズの中には日本語 WordNet の同義語, およびカナ, ローマ字で表記された検索語が含まれている.

「Yahoo テレビ　Japan」上の検索システムのような既存のものは, 検索語句を単に文字列として扱い, 単語単位で分離しないと考えられる. また, 文章において検索語が様々な語順や, 間に本来の検索語と違う単語が存在することによって, 望ましい検索結果が得られない場合がある. この解決策として, 本提案手法では有限状態オートマトン (a finite-state automaton, FSA) を利用している. FSA の実装によって, 文章に出現しているすべての検索語がすべての可能な語順で把握できるようになった. さらに長い文章を処理する際の高速化手法として, 並列処理を利用している. このことによって検索時間が約 95% 減少した.

日本語 TV ガイドに含まれる文字列を分析した結果, ローマ字文字列は約 10% であることがわかった. さらにこの文字列において, 約 64% のトークンが英語 WordNet に記載されている英単語と一致した. この点は,TV ガイドの日本語だけではなく, 英語も対象にできる異言語間検索の提案手法を開発する必要性を示している.

入力された検索語に対して WordNet から日本語と英語の同義語が抽出され, 検索語とともに検索に用いられる. また, 検索語の平仮名, カタカナ, ローマ字表記も生成されて検索に利用される. 結果的に元々のクエリの他に, その英訳, 日本語同義語, カナとローマ字表記を用いて TV ガイドの文章が検索される.

提案手法の評価には, アンケート調査によって得られた検索語句が用いられている. この評価において提案手法は,5 つのベースライン手法より良い性能を示した.F 値が, 提案手法では 0.718, 検索語句拡張を行わないベースラインでは 0.432 となり, 約 66 ポイントの検索効率の向上を示している. このことは日本語 TV ガイド検索における提案手法の有効性を示していると考えられる.

さらに, 前述した既存の TV ガイド検索システムにおいても, 提案手法と同じような英訳, 同義語, カナ表記のクエリ拡張を用いた異言語間検索を使用することにより, 性能が向上すると考えられる.

これを要するに, 著者は, 検索語句拡張により日本語 TV ガイドにおいて異言語間検索を行う新たな手法を提案し, 評価実験によりその有効性を示したものであり, 自然言語処理工学の発展に貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.