学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　Kiselev Denis

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. 例えば,「Yahoo テレビ　Japan」に対して検索実験を行ったところ, 検索結果の約 81% が, 抽出された検索語の語順が本来の検索語句と異なったり, 本来の検索語の間に違う単語が存在していることがわかった. つまり,「Yahoo テレビ　Japan」等の検索方法により, 約 81% の検索結果が得られないことが推測される. 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. 有限状態オートマトンはデータマイニング等で多く利用されているが, 本提案手法では TV ガイド検索の目的で実装を行った.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.

本提案手法では, 検索結果を適切か否か分類する際, 語の出現数だけではなく意味素性分析も行う. この点が検索結果における分類の効率化に繋がると考えられる. なお, 提案手法による約 66% の検索効率化がマルチパラメータ評価によって示されている. 本提案手法のように, 上述のクエリ拡張と意味素性分析を併せ持った手法は, 他に類似している研究が見当たらず, このことが本研究の独自性

である.

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,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. つまり, 日本語 TV ガイドに含まれる英単語は, ローマ字表記の日本語より多く, 相当数存在するという結論が導き出される. 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.

提案手法の概要を以下に述べる. 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. 検索の際に, 文章に現れているこれらの語における, 全ての可能な語順や組み合わせを把握できるように FSA が実装されている. しかし, データ量が多いためにこの検索過程は遅くなる場合もある. そこで,TV ガイドを分離して, それぞれの部分を並列的に処理するという手法を取ることにより, 検索時間は 95% 減少した. このような方法により, ユーザーの待ち時間の負担が軽くなった.

前述した同義語やカナが生成される前に, 元々の検索語に対してストップリストを用いたフィルタリングと形態素解析が行われる. フィルタリングの際に検索語が, 助詞などのそれ自身では意味を持たない語を含むストップリストと照合され, 一致したものは検索で使用されない. 一部の語尾もこの時に排除される.

提案手法で使用されている形態素解析の役割は, 文字列を語単位で分離すること及び, 語において意味素性分析を行うということである. 意味素性分析の際に,「オブジェクト」(物体など) と「オブジェクトのプロパティ」(属性) という二種類の意味素性索引をそれぞれ語に付与し, その索引の検索結果が適切か否かの分類を行うことに利用される. 索引の付与は,MeCab という形態素解析ツールの品詞情報出力をもとにしている. つまり, 名詞には「オブジェクト」の索引が付与されて, 形容詞は「オブジェクトのプロパティ」の索引が付与される.

検索結果はグループ化されてから出力される. ディスプレイ上では, より適切な結果のグループが適切性の低いグループよりトップに近くなる. 上記の索引が付与されている語の組み合わせを含む結果は, 他の結果より適切だと判定される.

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. つまり,TV ガイドのウェブサイト等でも提案手法の実用性があるものと考えられる.