論文内容の要旨

博士論文題目 Cross-language Information Retrieval, Document Alignment and Visualization
— A Study with Japanese and Chinese
(多言語間情報検索 — 文書対応付けとその視覚化 —)

氏名 Hasan Md Maruf

(論文内容の要旨)

インターネットと電子図書館が出現により、多言語情報の急増し、これらの情報のインデックス化及び検索を行うために、文書を洗練した形式で表現することは重要な課題である。近年、電子的に利用できる情報の量は段階的に増加しており、英語以外の情報 (アジアやヨーロッパの言語の情報) についても急増の一途をたどっている。このような背景において、ヨーロッパ言語における多言語間検索についてはいくつかの研究が報告されているが、日本語-中国語間の言語横断検索 (CLIR) はこれまでのところ初期的な段階に留まっている。本研究では、書き言葉において用いられる漢字が表意文字であるという点でヨーロッパ言語とは全く違う日本語と中国語の二言語を対象とした。

本論文では、漢字と漢字から導出される意味を用いて日本語と中国語の検索とインデックス化をするいくつかの方法を研究し、漢字に基づいた日本語-中国語の検索に関する枠組みを提案した。この枠組みは本質的に柔軟なベクトル空間モデルに基いている。特異値分解 (SVD) は潜在的意味インデックス化 (Latent Semantic Indexing) の概念と語彙とを結びつけることができる基本的な手法であるが、提案モデルは、この SVD のようなベクトル空間の次元を圧縮する手法と結合することが容易である。本論文では、このような手法との結合について実験を行った。

この研究の最終的な目的は、自然言語処理 (NLP) および情報検索 (IR) 技術の成果を容易に利用することが可能な日本語、中国語の情報検索 (IR) と横断言語検索 (CLIR) のための適当な枠組みを明かにすることである。本研究が提案する漢字意味に基づいたベクトル空間モデルは、境界分析 (segmentation) や形態素解析といった NLP 技術の発展や、質問拡張、関連性フィードバックといった IR 技術の発展に基づいて、モデルを容易に拡張し、適応することが可能である。また、真の多言語化への拡張として、日本語と中国語の漢字意味をヨーロッパ言語の語幹 (word-roots) に対応付けることが可能である。

同時に、本研究では、二ヶ国語文書の対応付けに関する取り組みを行い、文書対応付けを用いて自動的に日本語-中国語の対称コーパス (comparable corpus) を作成するアルゴリズムを提案した。最後に、二ヶ国語文書集合に対して概念クラスタリングの実験を行い、多言語情報を取り出す上での、文書のクラスタリングによる可視化の有効性を評価した。

Keywords: 多言語情報処理, 横断言語検索, 概念検索, 文書対応付け, 概念クラスタリング

氏名	Hasan Md Maruf
----	----------------

(論文審査結果の要旨)

平成13年7月23日に開催した公聴会の結果を参考に平成13年9月4日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

Hasan Marufは、本博士論文において、日本語-中国語間の言語横断検索 (CLIR) を対象とし、両言語に共通の書き言葉表記文字である漢字を媒介として、単言語を対象とした自然言語処理および情報検索技術がどのように有効に利用可能かを明かにした。

本論文では、漢字と漢字から導出される意味を用いて日本語と中国語の検索とインデックス化を行うためのいくつかの方法を検討し、漢字に基づいた日本語-中国語の検索に関する枠組みを提案した。ここでの方法は、情報検索で広く用いられているベクトル空間モデルを用いた。漢字をインデックスに用いるだけでは、たとえ類似の文字を用いる言語でも言語間検索を有効に行うことは難しい。そこで、特異値分解 (SVD) に基づく潜在的意味インデックス化 (Latent Semantic Indexing) という考え方を利用し、日本語中国語両者の漢字空間を同時に圧縮することにより、漢字が持つ意味的な情報をより柔軟に利用する手法を提案した。これを、単漢字や漢字 bi-gram などのインデックスに利用し、その効果を明かにした。

本研究では、自然言語処理 (NLP) および情報検索 (IR) 技術の成果を容易に利用することが可能な日本語、中国語の情報検索 (IR) と横断言語検索 (CLIR) のための適切な枠組みを明かにすることを最終的な目的としている。本研究が提案した漢字意味に基づいたベクトル空間モデルでは、境界分析 (segmentation) や形態素解析といった NLP 技術の発展や、質問拡張、関連性フィードバックといった IR 技術の発展に基づいて、モデルを容易に拡張し、適応することが可能であることを示した。

本研究では、さらに二ヶ国語文書の対応付けに関する取り組みを行い、文書対応付けを用いて自動的に日本語-中国語の対称コーパス (comparable corpus) を作成するアルゴリズムを提案した。また、二ヶ国語文書集合に対して概念クラスタリングの実験を行い、多言語情報を取り出す上での、文書のクラスタリングによる可視化の有効性を評価した。

近年注目されている言語横断検索の中でも特殊な関係にあるアジア言語対に関する早期の研究であり、漢字インデックスを利用することの有効性を示したということによって有意義な結果を示したと言える。よって、自然言語処理および情報検索の分野において高い貢献があると評価する。