

氏 名	植 村 俊 亮 うえ むら しゅん すけ
学位の種類	工 学 博 士
学位記番号	論 工 博 第 820 号
学位授与の日付	昭 和 50 年 11 月 25 日
学位授与の要件	学 位 規 則 第 5 条 第 2 項 該 当
学位論文題目	電子計算機による自動索引の研究

論文調査委員 (主 査) 教授 坂井利之 教授 清野 武 教授 大野 豊

論 文 内 容 の 要 旨

この論文は、電子計算機を使った索引の自動作成に関するもので、具体的な応用目的にそった多様な索引自動作成方式と、漢字かなまじり文の大規模なデータに適用した事例を内容とし、2部からなっている。

第1部は「自動索引方式の研究」と題し、書中の字句や事項を簡便に引き出せるように、一定の方式にしたがって編集するという、本来の意味での索引を、自動的に作成する方式を検討していて、10章よりなっている。

第1章は序論で、この研究の構成、自動索引研究の位置づけ、目的を明らかにしている。

第2章は KWIC (Keyword in Context) 索引について、それが自動索引手法の原点であり、現実的な唯一の手法で文献速報用索引に有効であるということを述べ、しかし日本では日本語特有の障害のために普及していない事情を説明している。

第3章は自動索引システム「ルイジュ」と題し、この研究過程で開発した索引作成プログラムの各種サブシステムを詳述している。この集積である「ルイジュ」は、計算機に無関係なソースプログラム・ライブラリ形式で、利用者はその目的に応じて適当なサブシステムを使用できる。サブシステムごとに、その目的・原理・用途・特徴について詳述してある。例えば日本語 KWIC 索引、索引行自動編集 KWIC 索引、拡張用例索引、句索引、単語語尾配列などである。また「ルイジュ」システム構成も示してある。

第4章は漢字かなまじり文からの見出し語自動抽出に関するもので、わかち書きをせずに普通に書かれた題名のなかから、索引の見出し語を自動的に抽出する手法を論じている。漢字、カタカナ、ひらがな、その他の字種および漢字列の構成規則に基いた手続き的な手法を開発している。

第5章、第6章は索引の見出し語の配列に関する内容で、英和辞典、国語辞典などにおける見出し語配列の原則を分析し、これを計算機で自動化するための分類用キー構成手法を考察している。6章は語尾配列に着目して、その意義と効果を示している。

第7章は索引行情報の圧縮を、英文を対象にして考察している。一行の文字数の制限からくる単語の切捨てを防ぎ、文脈を保つため単語中の母音の省略や略語化を試みている。

第8章は索引データ作成をめぐる諸問題、第9章は索引出力媒体としてのマイクロフィッシュ、第10章は「ルイジュ」システムプログラム要覧について述べている。

第2部は「大量日本語データによる自動索引の研究」と題し、第1部の基礎研究に立脚して、約75万行におよぶ漢字かなまじり文用例索引を自動作成した実験を内容としていて、9章よりなっている。

第1章は序論で、日本語大規模自動索引の意義と目的について述べ、第2章は、この実験に用いた国立国語研究所の言語データについての説明、第3章はこの国研データ自動索引計画のシステムと計画についての説明である。

第4章は漢字符号とその変換についての技術的、実際的問題について論じ、相互に符号変換可能な汎用のシステム採用について詳述している。

第5章は、大量索引データの処理についての実際的な問題とその解決策について述べ、第6章から第8章までは自動索引実験での重要項目について詳述している。すなわち、索引自動作成の過程を段階毎に具体的に示すこと、索引検索システムとして磁気テープに記録された大量の用例索引の中から必要な部分を手軽に自動検索するシステム（木かげ）、ならびに漢字マイクロフィッシュ索引の3つである。大量のデータ、その出力の形式と媒体などの選択について実際に扱ってみたときに生ずる問題点が詳しく述べられている。

第9章は結言である。

論文審査の結果の要旨

自然言語の電子計算機による自動処理のうちでも、自動索引はその内容からも実用性からも最も要望が高いものの一つである。

筆者は既に諸外国の科学技術論文で実際に採用されているキーワードならびに KWIC (Keyword in Context) に着目し、これを利用者からみたときの改良と拡張の方法、さらに日本語に対して適用する際の文字の入力、出力の問題、日本語漢字かなまじり文の処理の問題について、実用に耐える程度の量と質の開発研究を行なって、これをまとめたのが本論文である。

日本の国語辞典には単語の意味は記してあっても、その用例の文章を載せたのが少なく、しかもアルファベット言語のように、文章中で単語が独立して自明な存在でなく、漢字かなまじり文ではわかち書きの必要性があるなど、電子計算機への漢字入力のほかに様々な問題があって、本格的な日本語文章の自動処理の研究は甚だしい立遅れの状態にあった。

筆者は、このような環境の下にあって、主として日本語現代文章（75万単語）のデータを対象として、KWIC、拡張用例索引、句索引、単語語尾配列などを含む一貫した目的向き、手続き手法による自動索引システムを開発した研究をまとめたのが本論文であって、その主な成果をあげると次の通りである。

1) 電子計算機による自動索引の各種目的向き索引作成サブシステムのプログラムの集積である「ルイジュ」を開発した。これは計算機に無関係なソースプログラム・ライブラリ形式で作成され、利用者は目

的に応じて、日本語 KWIC 索引、索引行自動編集 KWIC 索引、拡張用例索引、句索引、単語語尾配列などのサブシステムを任意に選ぶことが出来る。特に拡張用例索引、句索引、単語語尾索引には、筆者のアイデア、工夫が多く盛り込まれていて、しかも結果として利用者に便利で有効な特徴を含んでいる。直ちにどのような電子計算機にも導入可能な形で開発された「ルイジュ」は、今後の日本語の計量言語学的研究、文献検索や索引作成において実際的にも利用される各種方式の端緒となるものである。

2) 漢字かなまじり文からの見出し語自動抽出の実際的な方法を考案し、これを大量のデータについて実験を行ない、KWIC サブシステムに応用して、索引自動作成には十分な精度で、自動わかち書きが出来ることを示した。原理的には漢字かなまじり文の字種および漢字列の構成法則に基いた手続き的な手法であるが、今後の計量言語学研究、実用計画に指針と勇気を与えるものである。

3) 索引データ作成にあたって、漢字入力の問題点と既存あるいは他のデータカードとの変換可能性、データ作成時の誤り発見の方法、出力が多量になるので出力用紙でなくマイクロフィッシュにするなど、システムの考慮を行ない、開発途上の問題点を解決した。また利用者の立場から磁気テープに記録された用例索引を、本を見開いて要所を探すのに似せた「木かげ」で、漢字ディスプレイ装置に表示させるようにした。

以上のように、この論文は電子計算機による自動索引、特に日本語漢字かなまじり文の大量のデータに対して開発研究を実施し、この分野の研究をかなり進展させ、学術上、工業上寄与するところが少なくない。

よって、本論文は工学博士の学位論文として価値あるものと認める。