「情報管理研究」, vol. 23, no. 3, 1992. 9., pp. 19～44

新聞記事 自動索引에 관한 考察

A Study on Automatic Indexing System for Newspaper Articles

趙 先 熙*
(Cho, Sun Hee)

抄 錄

최근 國內 대부분의 新聞社에서 CTS시스템을 도입함에 따라 記事全文이 컴퓨터에 입력되는 장점을 고려한 自動索引 시스템의 필요성이 대두되고 있다.

본 연구에서는 先行研究와 國內外 事例들을 통해 신문기사 자동색인 시스템의 問題點과 앞으로의 展望을 고찰하였다.

키 워 드

新聞記事 自動索引 시스템, 自動索引, 新聞記事 데이터베이스

ABSTRACT

As most of the domestic newspaper companies are adopting CTS system, the need for automatic indexing system, which can transfer the full-text into a computer, is sharply expanding. In this research, I tried to analyse problems and prospects of the automatic indexing system through various examples and studies conducted by other analysts previously.

KEYWORDS

Newspaper indexing, Automatic indexing, Newspaper database, Machine indexing

* 서울신문사 조사부.
Seoul Newspaper Research Dept.

Ⅰ. 序　　論

지금까지 新聞은 新聞記事만이 가질 수 있는 여러가지 특성들로·인해 현대 사회에서의 情報傳達 媒體로 중요한 역할을 수행하여 왔다. 新聞이 1930~1940년대 라디오 放送의 성공과 1950년대 등장한 텔레비전의 영향력에도 불구하고 讀者들을 계속 확보할 수 있는 것[1]은 자료로서의 記事가 一般情報와는 달리 매일 매일을 생생히 엮어가는 世界像을 반영하는 생동감 넘치는 산 역사로서[2] 多樣性, 眞實性, 興味性 외에도 記錄性과 保存性이라는 특징을 지니고 있기 때문이다.[3]

그런데 情報化社會에서의 정보요구는 점차 다양해져 情報傳達媒體일뿐 아니라 蓄積媒體로의 역할[4]을 요구하고 있으며, 일각에서는 뉴 미디어의 영향을 크게 받을 분야의 하나로 新聞事業을 지적[5]하고 있다. 뉴 미디어로 주목받는 CATV, 텔리텍스트, 비디오텍스 등이 時間的, 空間的 制約[6]을 갖는 新聞市場을 잠식할 수 있다는 전망인 것이다.

日本에서는 "전자적인 신문제작을 중심으로 速報와 檢索機能은 뉴 미디어와 데이터베이스로 보완하는 종합정보 산업체로"라는 슬로건이 큰 新聞社마다 등장하였는데,[7] 이것은 과학기술시대에 정보환경의 변화에 부응하려는 自救策으로서 우리나라 신문의 당면과제이기도 하다.

新聞社가 도입할 수 있는 뉴 미디어로는 텔리텍스트, 비디오텍스, CATV에 의한 뉴스제공, 프레스 팩스, 데이터베이스에 의한 정보 제공, 그리고 電子新聞 등을 들 수 있다.[8] 國內에서도 이미 많은 新聞社가 CTS시스템[9]의 도입에 따라 프레스 팩스로 전국 동시 印刷製作을 하고 있으며, 電子新聞을 서비스하

---

1) 이두영, "언론사 조사부의 미래상", 「調査硏究」, 3號, (1990), p. 81.
2) 李寬基, 「新聞記事情報銀行」, 서울 : 우정출판사, 1986, pp. 3~4.
3) 박계숙, 「신문기사 색인 데이터베이스 설계 : 동아일보를 대상으로」, 석사학위 논문, 연세대학교 도서관학과 대학원, 1982, pp. 3~4.
4) 서경주, 「언어학적 분석기법에 의한 신문기사 자동색인 시스템 설계에 관한 연구」, 석사학위 논문, 숙명여자대학교 도서관학과 대학원, 1990, p. 1.
5) 「정보화사회와 언론」, 서울 : 한국언론연구원, 1987, p. 8.
6) 이성훈, "언론사 뉴 미디어 출현 시대적 요청, 전자신문은 무료라는 인식없애야", 「신문과 방송」, 1990. 12), p. 18.
時間的 制約이란 新聞은 하루에 한번씩만 발행되며, 뉴스 발생 현장에서 印刷 및 配達까지 뉴스의 생명인 速報性이 떨어진다는 것을 말한다.
空間的 制約이란 인쇄매체인 新聞이 무작정 증면할 수는 없는 紙面의 限界를 가지고 있다는 것이다. 그외에도 일방적으로 전달하는 일방적 미디어로의 제약이 있다.
7) 「정보사회와 언론」, p. 233.
8) 이성훈, 앞의 기사, p. 19.
9) 납을 이용하는 HST에 대응하는 뜻으로 사용되는 용어로 컴퓨터와 전자사식기를 연결한 조판 방법을 말한다. Computerized Typesetting System의 약자이다.

고, 또 아울러 데이터베이스 구축에도 많은 관심을 기울이고 있다. 新聞記事 데이터베이스는 모든 자료가 機械可讀型으로 記錄, 貯藏되기 때문에 앞으로 有線, 人工衛星 등 뉴 미디어에 의한 電子配布 시스템을 통해 이용을 극대화 할 수도 있는 전망이다.

그런데 新聞記事 데이터베이스의 개발이나 電子新聞의 檢索에서 중요한 역할을 담당하는 것이 바로 효율적인 自動索引 시스템이다.

索引은 지식 기록물의 지적 내용과 물리적 위치에 대한 정연한 안내[10]로서 매일 수백건씩 나오는 新聞記事 처리에는 自動索引이 필수적인 것이다. 기존 국내에서 시도되었고, 또 지금까지 보편적으로 사용되는 색인방법은 통제어에 의한 手作業 索引으로 기사색인의 목적 외에도 제작지원용 클리핑 자료의 분류, 정기간행물 색인, 사진자료 및 슬라이드의 분류·정리 등에서 행해지고 있다. 그러나 내용이 다양하고 수명이 짧은 新聞記事 색인어를 수작업으로 일관되게 유지한다는 것은 사실 많은 어려움이 있으며, 시소러스 작성도 투자한 費用, 努力, 時間에 비해 實用化되고 있는 것이 없는 상태이다.

본 연구에서는 CTS로 記事全文이 컴퓨터에 입력되는 장점을 고려한 自動索引 시스템에 대해 선행연구와 국내외의 사례들을 통해 문제점과 앞으로의 전망을 고찰하고자 한다.

新聞記事는 간결한 표현이 많고, 단어의 중복을 피한다는 특성을 고려해서 통계적 기법보다는 언어학적 기법을 적용하는 것이 효율적이라고 판단하였으며, 따라서 최근까지 제시된 自動索引의 언어학적 기법의 이론들을 살펴보고, 國內 및 日本新聞에서 실제로 구현된 시스템을 분석하였다.

본 연구의 목적은 현재 가동중인 시스템의 분석을 통해, 향후 그 단점들을 보완할 수 있는 시스템 구현에 도움이 되도록 하는 데 있다.

Ⅱ. 理論的 背景

1. 自動索引의 槪念

정보를 분석·가공하여 축적하고, 검색하는 것은 索引의 작성과 활용에 관

---

10) Donald B. Cleveland, Ana D. Cleveland., 「색인 및 초록 작성법」, 서울 : 구미무역 출판부, 1990, p. 29.

2. 言語學的 분석 기법

索引語 抽出을 자동적으로 수행하기 위하여 확률, 통계적 기법을 응용하는 것 외에 언어 자체에 내재하고 있는 특성을 분석해 주제의미를 구별하는 言語學的 分析方法이 있다.[24]

言語學的 技法은 문헌의 의미분석에 기초한 것으로서, 즉 단어의 문법적 이용(構文)이 문헌내용의 결정에 도움을 줄 수 있다는 假定에 기초한 것이다.[25]

自動文獻處理, 특히 자동색인 분야가 발전하기 위해서는 언어학적 지식이 뒷받침되어야 하며, 자연언어의 구조가 이해되어야만 가능하다는 것이 학자들의 공통된 견해이다.[26]

言語學의 여러 하부 영역들 중, 정보학과 밀접한 관련을 가지고 있는 분야는 단어의 구조를 연구하는 語彙論(Morphology), 한 문장의 단어들을 전치사와 같은 구조적 단위들로 묶어주는 構文論(syntax), 그리고 언어의 표현과 실제 세계에서의 그에 대한 대응물(object)과의 관계를 연구하는 意味論(semantics) 등이 있다.[27]

현재까지 自動索引 분야에 이용된 言語學의 영역은 주로 語彙論과 構文論 등이다. 語彙的 段階의 기법으로는 KWIC 색인에서 볼 수 있는 것과 같이 불용어사전을 이용한 불용어제거 기법이 있으며, 구문적 기법에서 단어의 구문적 범주를 위해 단어사전을 사용하는 방법이 여기에 해당한다.[28]

不用語除去 技法은 KWIC 색인에서와 같이 단독적으로 사용되거나 다른 색인기법과 함께 사용된다. 이 기법은 텍스트 내의 각 단어를 분리한 다음 불용어, 즉 전치사, 접속사, 조사, 관사와 같은 기능어 및 기타 공통적으로 사용되는 고빈도 단어를 제외한 나머지 단어를 모두 索引語로 선택하는 방법으로,[29] 가장 단순한 어휘적 단계로 볼수 있다.

言語學的 方法의 주류를 이루는 構文分析 技法은 특정한 구문적 기능을 수행하는 단어나 단어구가 문헌의 내용을 나타낸다는 가정 아래 이러한 구문단위를 식별하는 작업을 의미한다.[30]

---

24) 이태영, "색인 시스템 內의 自動化에 대한 考察",「도서관」, vol. 41, no. 2, 1986, p. 14.
25) 허미숙,「지식 베이스를 이용한 자동색인 시스템에 관한 연구」, 석사학위 논문, 연세대학교 도서관학과 대학원, 1991, p. 6.
26) R. F. Simmons, "Automated Language Processing", *ARIST*, 1, 1966, p. 137.
27) G. Salton, M. J. MaGill, *Introduction to Modern Information Science*, New York : McGraw－hill, 1983, p. 59(서경주, 앞의 논문에서 재인용, p. 13).
28) Salton, op. cit., p. 260(정영미, 앞의 책, p. 147에서 재인용).
29) 정영미, 앞의 책, p. 147.
30) Spark Jones, "Automatic Indexing", *Journal of Documentation*, vol. 30, no. 4, 1974, pp. 393～432.

構文分析 技法은 부분적인 문장분석에서부터 완전한 문장분석에 이르기까지 문장분석의 단계가 다양하다. 부분적인 문장분석 기법으로는 구두점이나 전치사, 접속사, 조사 등을 단서로 해서 문장을 문법적으로 분석하여 전치사구나 명사구 등의 한 단어로 처리되는 單語群을 찾아낸 다음, 이 가운데 빈번히 나타나는 단일어나 복합어를 색인 선택하는 방법들이 있다.[31] 美國의 DDC(Defense Documentation Center)가 개발한 MAI(machine-aided indexing)는 單語辭典(recognition dictionary)과 함께 색인어로 선택될 단어구의 76개 구문 형식을 수록한 形式辭典(format dictionary)를 이용하고 있다.[32]

보다 완전한 구문분석 단계는 어의적인 처리만 하지 않을 뿐, 句에서부터 節에 이르는 거의 완전한 문장분석 단계이며, 더욱 복잡하고 수준높은 구문분석은 컴퓨터에 내장된 文法과 語義辭典을 이용해서 완전한 문장분석을 행하는 것이다.[33] 그런데, 실제로 自動索引에 있어서 완전한 문장분석은 그 복잡성에 비추어 큰 효과가 기대되지 않고 있으므로, 이 수준의 索引 시스템은 별로 찾아볼 수 없다. 완전한 文章分析은 오히려 質問應答 시스템이나 自動飜譯 분야에서 요구되며, 이 분야에서 주로 연구되고 있다.[34]

構文論的 技法을 사용한 自動索引은 주로 단어의 구문적 형태, 명사구를 이용하여 색인어를 추출하는데,[35] 명사구는 심리학적, 구문론적 측면에서 문헌내용의 지표로 중요한 역할을 수행한다.[36]

최근 연구자들은 情報檢索 시스템의 성능이 단순히 중요어(keyword)를 사용하는 대신 核心句節(key phrases)를 사용함으로써 크게 향상될 수 있음을 확신하게 되었다.[37), 38)]

---

31) 정영미, 앞의 책, p. 148.
32) P. H. Klingbiel, "Machine-Aided Indexing of Technical Literature", *Information Storage and Retrieval*, vol. 9, no. 2, 1973, pp.79~84.
33) 정영미, 앞의 책, p. 166.
34) Ibid.
35) 최원태, 앞의 논문, p. 7.
36) R. K. Waldstein, op. cit., pp. 12~13(최원태, 앞의 논문에서 재인용, p. 7).
37) Joel L. Fagan, "Automatic Phrase Indexing for Document Retrieval : an Examination of Syntactic and Non-syntactic Methods", *Proceedings of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York : the Association of Computing Machinery, 1987, pp. 91~101(홍승의, 「신문기사의 자동색인 시스템에 관한 연구」, 석사학위 논문, 연세대학교 도서관학과 대학원, 1990, p. 13에서 재인용).
38) R. Tong, L. Appelbaum, V. Askman and J. Cunningham, "Conceptual Information Retrieval Using RUBRIC", *Proceeding of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York : the Association of Computing Machinery, 1987(홍승의, 앞의 논문에서 재인용, p. 14).

이상에서 기술한 言語學的 分析技法들은 주로 英語圈을 중심으로 연구되었는데, 이들 연구의 개요를 살펴보면 다음과 같다.

W. D. Climenson[39] 등은 自然言語 구조의 규칙들을 형식화하여 이러한 구조에서 허용된 구성요소들을 인식함으로써 意味分析과 필요하다면 통계적 분석까지 해낼 수 있는 言語分析 技法을 제안했다. 즉, 文獻內에 있는 한 문장 요소들의 구문적 기능을 나타내기 위해 構文分析 프로그램을 개발하였다. 구문분석을 마치는데 필요한 단계들은 컴퓨터가 文章을 읽고 지나가는 과정상에서 구분된다. 초기에는 각 문장을 문장 성분 단위로 다시 쓴 후에 괄호로 구분될 수 있는 특정 형태의 語句를 찾는다. 이런 과정들을 거치면서 괄호로 묶어진 전체 어구를 단일 개념으로 추출한다.

대표적인 自動索引 시스템인 스마트(SMART)시스템은 자연언어로 표현되어 있는 문헌과 질의문을 처리하여 질의문에 대한 응답으로 질문에 가장 가까운 문헌을 식별해내는 完全 自動文獻 시스템이다.[40] 동의어사전 혹은 시소러스는 중요한 단어의 어간들을 개념번호로 대치시켜서 어떠한 개념번호가 주어지더라도 그와 관련된 상위어, 하위어, 동등어 등을 참조할 수 있게 해주며, 불용어사전(negative dictionary)은 출현빈도는 높으나 별 의미를 갖지 못하는 불용어들을 제거시켜준다. 구사전(phrase dictionary)은 어떤 분야의 주제를 나타내는 명사 및 전치사 句에 대응하는 2~4개의 단어나 개념의 열들로 구성된다. 이들 구사전들은 통계적 구사전과 구문론적 구사전으로 나눠질 수 있는데, 前者는 句성분들의 동시 발행 특성만을 고려하는 것이고, 後者는 인식되어지는 句성분에 관한 사항들 뿐만 아니라 句가 인식되어지기 위해 필요한 통사의존 관계에 관한 정보도 포함한다.[41]

M. Dillon과 A. S. Gray가 개발한 FASIT(fully automatic syntatically based indexing system)[42]는 語義分析은 하지 않으면서도 문장내의 주요구들을 식별해서 유사 동의어군으로 묶어주는 시스템이다. 이 시스템은 의미있는 단어구

---

39) W. D. Climenson, et al., "Automatic Syntax Analysis in Machine Indexing and Abstraction", *American Documentation*, 1961, pp. 178~179(안현수, "한글문헌의 자동색인에 관한 실험적 연구", 「情報管理學會誌」, 3권 2호, 1986, p. 113에서 재인용).

40) Gerald Salton, *Automatic Information and Retrieval*, New Jersey : Prentice-Hall, 1971, p. 9(안현수, 앞의 記事, p. 113에서 재인용).

41) Gerald Salton, *The SMART Retrival System*, New York : McGraw-Hill, 1971, p. 117 (안현수, 앞의 記事).

42) M. Dillon, A. S. Gray "FASIT : A Fully Automatic Syntactically Based Indexing System", *JASIS*, vol. 34, no. 2, 1983, pp. 101~105.

를 색인어구로 선택하는 槪念選擇 段階와 이들을 표준화 시키는 槪念集團化 段階로 이루어진다.

효율성 평가를 위해 250개의 문헌과 22개의 질문으로 실험을 한 결과 시소러스와 語幹을 이용한 索引 시스템보다 우수한 것으로 나타났다.

T. Maeda 등이 제시한 '의미있는 명사구'를 선정하는 索引 시스템에서는 2,400개 정도의 단어를 수록한 사전과 단어 끝부분 일치 규칙[43]을 이용하여 각 단어의 구문적 카테고리를 결정한 뒤, 특정한 패턴에 맞는 단어나 단어구를 명사구로 판정, 색인어로 선정하였다. 辭典에는 불용어, 부사, 동사와 같이 무조건 삭제해야 할 단어와 일정한 조건에서만 삭제할 단어(특정한 형용사와 명사), 그리고 어떤 경우에도 삭제해선 안될 단어의 세부류의 단어가 수록되어 있다.[44]

P. H. Klingbiel에 의해 개발된 MAI(machine-aided indexing system)[45]는 두개의 상이한 사전과 두개의 서로 다른 일시적 記憶場所로 구성되어 있다. 索引過程은 먼저 색인하고자 하는 문헌을 단일어로 구성되어 있는 認識辭典(recognition dictionary)과 비교해서 받아 들일 수 있는 단어들은 일시적 기억장소에 넣는데, 여기에 단어열들이 축적된다. 이런 단어열들은 形態 레지스터(format register)에 N, AN, ZZZ 등과 같은 기호 형태로 나타내며, 이 사전의 내용들은 색인어로서의 적절성을 평가받기 위하여 形態辭典과 대조된다. 형태사전과의 대조에서 일치가 되면 예비색인어로 출력이 되며, 이러한 索引作業은 句讀點 등이 나타나면 자동적으로 정지하게 된다.

K. P. Jones와 C. L. M. Bell은 접사 제거방법과 변형단어들에 대한 정보제공 방법, 복합어를 단일어로 통제하는 방법, 복합어인 경우 복합어를 이루는 특정 단일명사를 입력하고, 단일명사 좌우의 단어가 갖는 첫 철자를 位置役割(positional roles)로 부기하는 방법 등을 써서 명사구를 추출하는 MORPH 시스템을 개발하였다.[46]

---

43) 예 : y로 끝나는 단어는 부사이다.

44) T. Maeda, et al., "An Automatic Method for Extracting Significant Phrases in Scientific or Technical Documents", *IPM*, 16, 1980, pp. 119~127(정영미, 앞의 책에서 재인용, p. 167).

45) P. H. Klingbiel, "Machine-Aided Indexing of Technical Literature", *Informatiom Storage and Retrieval*, 9, 1973, pp. 477~478(안현수, 앞의 記事에서 재인용, p. 114).

46) K. P. Jones and C. L. M. Bell, "MORPHS-an Intelligent Retrieval System", *ASLIB Proceedings*, vol. 38, no. 3, 1986, pp. 71~79(허미숙, 앞의 논문, p. 8에서 재인용).

계를 격관계로 표현하여 색인 표목을 구성하였다. 自動索引 모듈에서는 격관계이외에 키워드 색인도 작성하였으며, 각 색인은 檢索 모듈에서 이용자의 탐색 목표에 적합한 檢索技法과 함께 사용되도록 하였다.[60]

정준민은 제9차 全國新聞放送通信調査記者會 세미나에서 색인어의 특성을 고유명사와 그 밖의 어휘로 구분하고, 선정된 색인어 집합에 고유번호를 부여, 고유번호로 통제된 색인어들의 집합을 계층적 집락기법에 의해 생성된 분류번호를 해당 기사에 할당함으로써 관련기사를 檢索하는 방법을 발표하였다.[61]

최원태[62]는 自動索引의 연구 경향을 다음과 같이 분류하여 제시하였다.

① 어미사전, 불용어 목록 등의 간단한 구조의 사전을 사용하는 것이 아니라 단어의 構文的, 意味的 특성에 따른 辭典을 사용한다.

② 索引語의 구문형태는 單一語보다는 名詞句를 선호한다.[63]

③ 意味 네트워크의 데이터 구조 및 지식 표현 기법을 응용하여 색인어를 추출한다(人工知能의 방법 이용).

④ 索引語 선정의 기준에 이용자의 피드 백을 응용한다.[64]

⑤ 索引語의 문법적, 의미적인 역할과 기능을 표현한다.

Ⅲ. 新聞記事 自動索引 시스템의 現況

1. 新聞記事 自動索引의 필요성

情報化社會, 脫工業化社會라는 용어들이 생겨난 후 새로운 뉴 미디어 시대가 시작된지도 벌써 20년이 지났다. 기술문명은 1970년대 이후 기하급수적으로 발전하여 情報의 傳達手段이 혁명적으로 발달되고 있는데, 第5世代 컴퓨터와 人工知能이 바로 그것이다.[65]

---

60) Ibid.
61) 정준민, "신문기사 색인의 새로운 시도", 「제9차 전국신문방송통신조사기자 세미나 보고서」, 1992, pp. 7~21.
62) 최원태, 앞의 논문, p. 9.
63) Herold Borko, "Automatic Indexing : A Tutorial", *ACM SIGIR Forum*, vol. 1, 1981, pp. 9~13(최원태, 앞의 논문에서 재인용).
64) Gerald Salton, "A Blueprint for Automatic Indexing", *ACM SIGIR Forum*, vol. 16, 1981, pp. 30~38.
65) 「정보화사회와 언론」, p. 188.

人工知能에 대한 연구는 1960년대 말경부터 본격적으로 시작되었으며, 사람이 소유하고 있는 知能(intelligence)을 컴퓨터가 가질 수 있도록 하는 것이 궁극적인 목적이다. 즉, 言語를 이해하고 學習能力을 가지며, 推論을 하고 問題를 해결하는 것과 같은 인간이 가진 지능을 가지고 이러한 일들을 할 수 있는 知能型 컴퓨터 시스템(intelligent computer system)을 설계하는 것이 그 목적이다.[66]

人工知能은 CTS, 데이터베이스, 비디오텍스 등 新聞社의 情報處理 機能을 획기적으로 끌어올릴 電子媒體에게 절대적인 영향을 끼치게 될 것으로 주목되고 있다.[67]

情報化社會는 현재 선진 사회에서 새로운 단계에 접어들고 있는데, 産業社會가 1980년대 후반과 1990년대에 걸쳐 知識化社會의 양상을 띠고 있는 것이 바로 그것이다. 知識化社會란 지식처리 테크놀러지를 응용해 모든 지식과 정보를 데이터베이스化하는 사회이다.[68]

言論媒體에서도 요즘 知識化社會임을 입증하는 DB저널리즘이라는 새로운 용어가 출현하고 있다.

DB저널리즘[69]이란 DB(데이터베이스)라는 용어와 저널리즘이 융합된 새로운 報道方式을 지칭하는 것으로, 온라인 데이터베이스에서 情報를 探索, 記事를 작성하고, 한층 더 나아가서는 政府나 商業用 컴퓨터의 기록 테이프를 입수, 그것을 편집국의 컴퓨터에 연결시켜 데이터를 분석·결합하거나, 서로 매치시켜 새로운 데이터를 만들어 記事化하는 것을 의미한다. DB저널리즘의 역사가 오래된 美國의 新聞들은 이제 論調나 速報 차원의 경쟁과는 다른 데이터베이스를 이용한 경쟁을 벌이고 있다.[70]

新聞記事 自動索引 시스템은 각 新聞社가 大型 컴퓨터로 CTS 시스템 전체를 운영하는 綜合情報 시스템을 계획하면서 電子新聞과 데이터베이스 구축의 필요성에 의해 논의되기 시작했다.

記事를 색인 처리하지 않은 電子新聞은 記錄性이라는 측면에서 신문지면과 다를 바 없으며, 모든 記事가 색인만 되어 있다면 몇십만건의 뉴스가 모여 있

---

66) 김영환, "정보검색에 미치는 인공지능의 영향", 「인공지능소식」, 9호, 1988. 2, p. 2.
67) 「정보화사회와 언론」, p. 188.
68) Ibid.
69) 유경희·강명구·원영희, "특집 : DB저널리즘", 「신문과 방송」, 제258호, 1992. 6, pp. 2~8.
70) 앞의 글, p. 4.

다섯째, 우리말의 標準化도 이루어지지 않고, 띄어쓰기 규칙도 제대로 지켜지고 있지 않은데, 이것을 가장 많이 지키지 않는 것이 바로 新聞[82]이라는 점이다. 新聞은 한정된 紙面에 가능한 한 많은 정보를 전달하기 위해 최소한의 띄어쓰기만을 하는데, 이것은 自動索引 시스템에서 키워드의 식별을 어렵게 한다.

이상 살펴본 것같이 新聞用語는 특히 수명이 짧고 사회의 유행에 따라 동일한 주제어도 다른 표현으로 사용되므로 시소러스의 활용도 염두에 두어야 할 것이다.

英語文章의 경우 거의 완벽한 自動索引이 이루어질 수 있으나, 한글문장의 경우 한글의 특성상 完全 自動索引은 어려움이 많다.[83]

電算處理를 위한 한글의 특성을 살펴보면 다음과 같다.[84]

첫째, 國語는 어근에 派生接詞나 어미가 붙어서 단어를 이루는 添加語이다.[85] 즉, 모든 문법적 형태소(문장안에서 체언 기능을 보여주는 助詞, 用言의 활용어미 같은 것들)는 반드시 어간이나 어근 뒤에 온다. 이 점은 굴절어인 영어와 크게 다른 한글의 특징으로 전산처리를 위한 形態素 分析에 중요한 점이다.

둘째, 조사, 어미, 접사가 매우 발달되어 있다.

셋째, 體言은 격조사를 취하나 性(gender), 數(number)의 구별이 없다.

넷째, 韓國語의 각 품사들 중에서 문헌의 주제어가 될 수 있는 경우는 명사, 대명사, 수사를 포함하는 체언이 단독으로 쓰인 경우거나, 이들이 서로 결합하여 복합어구를 형성하는 경우이다.[86]

다섯째, 전치사, 관사, 관계대명사가 없다.

여섯째, 語順이 비교적 자유롭다.

2. 新聞記事 自動索引 시스템의 現況

(1) 國內의 新聞記事 自動索引 시스템

국내에서는 유경희가 1977년 新聞記事 索引 시스템으로는 처음으로 KWIC

---

82) 柳京熙, "情報社會에의 對應", 「신문과 방송」, 1983. 9, pp. 55.
83) 조성호, "신문기사 데이터베이스 컴퓨터 자동색인(하)", 「신문과 방송」, 241, 1991. 1, p. 33.
84) 서경주, 앞의 논문, p. 27.
85) 남기심 · 고영근, 「표준 국어문법론」, 서울 : 탑출판사, 1987, p. 21.
86) 안현수, 앞의 글, p. 116.

색인 시스템을 제시하였다. 이는 원래의 KWIC 색인 시스템을 약간 변형시킨 것으로, 신문기사 내용을 6하 원칙중에서 '언제'를 뺀 5하 원칙에 입각, 40자 이내(한 記事當 천공카드 2장분)로 한 문장형태로 요약해 참조 코드(年, 月, 日, 面)와 함께 컴퓨터에 입력시킨 뒤 KWIC 방식에 따라 출력시키는 自動索引 시스템이다.[87] 여기서는 색인생산 결과 마련된 데이터베이스를 후에 컴퓨터에 의한 情報檢索에 활용하기 위해 助詞는 모두 띄어서 입력시키되, 출력시는 붙여 주도록 하는 방법이 제시되었다.

東亞日報社에서는 국내에서는 유일하게 1920년 창간호부터 1962년까지의 新聞記事 索引集을 발행[88]하며 얻은 경험을 토대로 스포츠 기사에 대한 실험적인 自動索引 시스템을 설계한 바 있다.

新聞記事 데이터베이스는 美國의 뉴욕 타임스가 1970년대 초반에 구축한 이래로 美國, 日本 등의 선진국에서는 이미 商用化되고 있는데, 우리나라의 경우 新聞記事 데이터베이스와는 구별되는 뉴스 速報媒體로 中央日報의 JOINS, 每日經濟新聞의 MEET가 있으며, 과거 기사를 검색할 수 있는 新聞記事 데이터베이스로는 韓國經濟新聞社에서 개발한 韓國PC通信의 HITEL(舊 한경KETEL)과 韓國言論硏究院의 KINDS(舊 KPI-NEWSBASE : 언론종합정보은행)가 있다.[89]

그러나 國內에서도 대부분의 新聞社가 CTS 시스템의 도입과 함께 新聞記事 데이터베이스 구축에 대한 중·장기계획을 가지고 있으며, 현재 구체적인 작업에 착수한 곳으로는 中央日報와 朝鮮日報가 있다. 中央日報는 HAIRS를 이용한 新聞記事 自動索引 시스템을 개발하여 연내 시스템 가동을 목표로 실험중이며, 朝鮮日報는 개발단계인 것으로 알려지고 있다.

본 연구에서는 실제로 新聞記事 自動索引 시스템이 가동되고 있는 HITEL과 KINDS에 대해 검토하였다.

1) 韓國PC通信의 HITEL(舊 한경KETEL)

韓國經濟新聞社에서는 1988년 국내 처음으로 「한경KETEL」이라는 電子新聞으로 뉴스 速報를 시작하면서 自動索引 시스템을 채택하였다. 한경KETEL

---

87) 柳京熙, "傳達에 쫓겨 蓄積에 소홀한 新聞記事", 「신문과 방송」, 76, 1977. 3, pp. 26~33.

88) 金兌益, "컴퓨터를 이용한 기사색인", 「한국조사기자회 현장수첩」, 창간호, 1992, p. 35.

89) 卞春洙, "言論公用記事 데이터베이스(NEWSBASE)의 활용 : 新聞記事 데이터베이스의 現況과 展望", 「제8차 全國新聞放送通信調査記者 세미나집」, 1991, p. 25.

KINDS의 自動索引 시스템에서는 키워드가 全文을 대상으로 하여 자연언어 그대로 색인어로 선정되는데, 불용어 사전을 통해 불용어가 제거된 후 품사별로 형태소사전에 의한 분석을 거쳐 키워드가 추출된다. 自動索引 이전에 手作業으로 구축한 단어사전과 동의어 파일, 불명어 파일을 유지하고 있으며 최근 日本의 中日(주니치) 新聞의 시소러스를 토대로 수정작업한 시소러스를 첨가했다. 복합어의 경우 세 단어까지 조합가능하도록 하였고, 인명의 경우 형태소 사전에 색인전문가가 등록을 해주고 있다. 불명어로 처리된 신조어 등도 색인전문가가 담당하고 있다. 현행 自動索引 시스템은 手作業과는 86%의 일치율을 보이며, 재현율은 98% 정도인 것으로 자체 평가되고 있다. 그러나 재현율이 높다보니 불필요한 색인용어가 많다는 점, 특히 記事全文을 대상으로 하기 때문에 색인어의 부피가 커지고 있다는 점, 불명어로 처리되는 신조어나 새로운 동의어, 새로운 약칭의 처리는 색인담당자의 부담이 된다는 점, 키워드로 추출된 단어와 주제와의 관계가 정확하지 않은 점 등이 短點으로 지적되고 있다.

(2) 日本의 新聞記事 自動索引 시스템

美國과 日本 등 선진국에서는 신문기사 자동색인 시스템에 관한 연구가 활발하여 이미 상용중인 많은 데이터베이스들에서 이용되고 있음을 앞서 지적한 바 있다. 특히, 英語의 경우 98%의 정확률까지 나타내고 있다는 연구도 발표되고 있다.

본 연구에서는 英美圈보다는 우리나라와 언어구조 및 신문 편집체계가 유사한 日本의 新聞記事 自動索引 시스템을 대상으로 자동색인의 시행사례를 검토하였다.

1) 日本經濟新聞의 Needs-IR[93]

索引語는 제목으로부터 기사 앞부분 200자 이내의 범위를 추출 대상으로 하여 주제별로 구성된 통제어사전(키워드집)을 바탕으로 통제어로 색인된다. 이 과정에서 불용어가 제거되며, 단어 단위로 형태소가 분해된 후 다시 형태소 합성을 통해 키워드와 부합되는 어휘를 색인어로 추출한다. 따라서 색인어 체제는 통제어사전에 있는 회사, 단체, 인명, 품목, 업계, 항목, 지역, 칼럼명, 기사분류, 보조 키워드로 구성된다. 이러한 과정은 索引 프로그램에 의해 자동처

93) 金兌益, 앞의 글, p. 38.

리되며, 그 결과 추출된 색인어에 의해 색인자가 手作業 編輯過程을 통해 점검·수정한다.

2) 아사히 新聞의 HIASK[94]

색인어는 제목으로부터 기사 앞부분 13행 이내의 범위를 1차 추출 대상으로 하며, 이외에 색인자가 필요하다고 생각하는 주요 부분에 대해 수작업으로 설정한 범위를 추출 대상으로 하여 自然言語로 색인된다.

이 과정에서 單語辭典에 의해 문장을 분석하여 명사로 분석된 단어를 색인어로 선택하며, 복합어는 모든 조합형태를 색인어로 한다. 이러한 과정은 索引 프로그램에 의해 자동처리되며, 그 결과 추출된 색인어에 대해 색인자가 手作業 編輯을 통하여 점검하고 수정한다. 또한 분류어 형태의 미니 시소러스를 이용하여 記事의 주제, 종류, 국가코드, 기사종류로 구성된다.

이 시스템은 자연언어 색인방식을 중심으로 하고, 분류표(미니 시소러스)를 통해 이를 보완하고 있지만, 일반적 의미의 시소러스가 없기 때문에 동의어와 상, 하위 개념에 대한 統制槪念이 없다. 또한 複合語에 의한 모든 조합형태를 색인어로 선택하여 불필요한 조합형태까지 색인어로 저장하게 된다.

3) 요미우리 新聞의 YOMIDAS[95]

색인어는 제목으로부터 기사 앞부분 400자 이내의 범위를 추출대상으로 하며, 自然言語 형태 그대로 색인어로 선정된다. 여기서 먼저 불용어를 제거한 다음, 남는 단어를 단어사전을 통해 형태소를 분해한 후, 이들을 색인어로 추출한다. 복합어는 모든 조합 형태를 그대로 추출한다. 또한 시소러스 형태의 관리어사전을 이용하여 동의어를 설정하여 색인한다. 이러한 색인과정은 索引 프로그램에 의해 자동처리되며, 그 결과 추출된 색인에 대해 색인전문가가 수작업 편집을 통하여 점검하고 수정한다. 또한 특정한 종류의 수치정보와 주제분류 및 기사종류를 색인자가 편집하여 색인어로 추가한다.

索引語體系는 모든 분야를 포함하는 자연언어와 주제분류, 기사종류, 수치정보로 구성된다.

단일신문만을 처리하는 시스템인 관계로 HIAS와 마찬가지로 자연언어 색인방식에 의한 색인어 파일의 확장이 전체 시스템에서 크게 문제가 되지 않는다. 시소러스를 통해 자연언어 색인의 문제점인 檢索의 再現率 低下를 보완했

---

94) Ibid.
95) Ibid.

지금까지 新聞記事를 대상으로 한 自動索引 시스템에서는 言語學 理論을 도입하여 본문에서 색인어를 추출하려는 시도가 있었으나, 실제로 구현된 것은 불용어 제거기법 등을 이용한 초보적 단계에 머물고 있다. 그러나 人工知能 技法을 응용하려는 목적하에 자연언어 처리를 위한 한국어의 구문적 구조에 관한 연구는 현재 많이 진전되고 있으며, 構文分析과 意味分析을 병행한 시스템들도 제시되고 있는 실정이다.

기본적으로 자연언어처리 기술을 색인어 추출 시스템에 도입하려는 목적[97]은 언어학 정보를 이용해서 적절한 색인어를 찾고자 하는 데 있다. 言語學 情報는 구문, 의미 등에 걸쳐 다양한 형태로 존재한다. 그러나, 단순한 言語情報만으로는 文脈을 파악한다거나, 특정단어가 그 문맥에서 갖는 중요성 등을 찾아내는 것은 쉽지 않다.

좀더 정확한 索引을 위해서는 言語情報 뿐만 아니라 知識 베이스와 推論體系까지도 도입해서 문맥구조를 생성해서 각 후보 색인어의 중요성 혹은 관계성을 계산하여야 한다.[98]

현재 人工知能의 응용은 인간이 생활하는 모든 영역을 대상으로 하고 있으며, 人工知能의 핵심요소는 자연언어로서 知識情報를 컴퓨터에 이해시키려는 데 있다. 즉, 자연언어로 표현된 지식을 컴퓨터에 入力, 知識 베이스化(knowledge base)하여 컴퓨터에 대한 전문적인 지식 없이도 모국어로 컴퓨터를 사용할 수 있고, 기계번역, 컴퓨터와의 대화, 自動 프로그램 등을 가능하게 하는 것을 의미한다.[99]

知識基盤 시스템은 개념들간의 포함관계 및 유사관계들을 명확히 명시하여 사용할 수 있으므로 文獻檢索 시스템에서 큰 효과를 얻을 수 있다. 이러한 접근방법은 사용자 질의어의 이해를 쉽게 할 수 있고, 인간인 전문가와 비슷하게 적합한 文獻을 檢索하게 한다.[100]

비록 國內의 新聞記事 自動索引 시스템은 시작단계에 불과하지만, 예상되는 문제의 합리적 해결방안을 위해서는 人工知能 技法의 도입이 필요한 시기라고 생각된다.

---

97) 최기선, 앞의 글, p. 98.
98) Ibid.
99) 李楨賢, 「한국어 처리를 위한 구·절 문법과 질문응답 시스템」, 박사학위 논문, 인하대학교 전자공학과 정보공학전공, 1988, p. 3.
100) 김영환, "정보검색에 미치는 인공지능의 영향", 10.

〈參 考 文 獻〉

- 金庚煥, "情報化社會와 新聞編輯", 「신문연구」, 1975, 봄, p. 39.
- 김영환, "정보검색에 미치는 인공지능의 영향", 「인공지능소식」, 1988. 2, 9호, p. 10.
- 김영환, 「한글·한자·영어 혼용문의 자동색인 시스템」, 석사학위논문, 한국과학기술원, 1983.
- 金兌益, "컴퓨터를 이용한 기사색인", 「韓國調査記者會 현장수첩」, 창간호, 1992, pp. 35~38.
- 남기심·고영근, 「표준국어문법론」, 서울 : 탑출판사, 1987.
- 박계숙, 「신문기사색인 데이터베이스 설계 : 동아일보를 대상으로」, 석사학위논문, 연세대학교 도서관학과 대학원, 1982.
- 박상규, 「신문기사 정보관리 시스템의 현대화에 관한 고찰」, 석사학위논문, 연세대학교 행정대학원 언론홍보 전공, 1988.
- 卞春洙, "言論公用記事 데이터베이스(NEWSBASE)의 新聞記事 데이터베이스의 現況과 展望", 「제8차 全國新聞放送通信 調査記者會 세미나집」, 1991, p. 35.
- 사공철, 「情報檢索論」, 서울 : 亞細亞文化社, 1982.
- 서경주, 「언어학적 분석기법에 의한 신문기사 자동색인 시스템 설계에 관한 연구」, 석사학위논문, 숙명여자대학교 도서관학과 대학원, 1990.
- 안현수, "한글문헌의 자동색인에 관한 실험적 연구", 「情報管理學會誌」, 3권 2호, 1986, pp. 110~116.
- 오택섭, "韓國新聞 조사부의 실태와 문제점", 「신문연구」, 41호, 1986, 여름, pp. 8~11.
- 柳京熙, "情報社會에의 對應", 「신문과방송」, 1983. 9, pp. 55.
- 柳京熙, "傳達에 쫓겨 蓄積에 소홀한 新聞記事", 「신문과 방송」, 76, 1977. 3, pp. 26~33.
- 유경희·강명구·원영희, "특집 : DB저널리즘", 「신문과 방송」, 258, 1992. 6, pp. 2~8.
- 유경희 · 신현웅, "한국어 KWIC 색인 시스템에 관한 연구", 「한국정보과학회 학술발표논문집」, 1981, pp. 99~107.
- 李寬基, "新聞記事情報銀行", 서울 : 우정출판사, 1986.
- 이두영, "언론사 조사부의 미래상", 「調査研究」, 3호, 1990, p. 81.
- 이성훈, "언론과 뉴 미디어 출현 시대적 요청, 전자신문은 무료라는 인식 없애야", 「신문과 방송」, 1990. 12, pp. 18~19.
- 이영주, 「자동색인을 위한 한국어 형태소분석 알고리즘에 관한 연구」, 석사학위논문, 연세대학교 산업대학원 전자계산전공, 1987.
- 李楨賢, 「한국어처리를 위한 구·절 문법과 질문응답 시스템」, 박사학위논문, 인하대학교 전자공학과 정보공학전공, 1988.
- 이태영, "색인시스템內의 自動化에 對한 考察", 「도서관」, vol. 41, no. 2, 1986, p. 14.

- _____, 「정보화사회와 언론」, 서울 : 한국언론연구원, 1987.
- 정영미, "우리말 정보자료를 처리하는 지능형 정보검색 시스템의 설계", 「情報管理學會誌」, 8권 2호, 1991, pp. 3～31.
- 정영미, 「정보검색론」, 서울 : 정음사, 1987.
- 정준민, "신문기사 색인의 새로운 시도", 「제9차 全國新聞放送調査記者會 세미나 보고서」, 1992, pp. 7～21.
- 조성호, "신문기사 데이터베이스 컴퓨터 자동색인(상)", 「신문과 방송」, 제240호, 1990. 12, pp. 89～93.
- 조성호, "신문기사 데이터베이스 컴퓨터 자동색인(하)", 「신문과 방송」, 제241호, 1991. 1, pp. 33～37.
- 최기선, "구문 및 의미분석을 통한 한국어 자동색인", 「情報管理學會誌」, 8권 2호, 1991, pp. 96～107.
- 최기선, "한국어 정보처리와 지능형 자동색인", 「한국정보학회 정보관리 강좌」, 1991, p. 101.
- 최원태, 「격문법을 이용한 자동색인 및 탐색확장에 관한 연구」, 석사학위논문, 연세대학교 도서관학과 대학원, 1986.
- 한국언론연구원, "조사보고서 : 90년대 언론과 독자, 제4회 미디어의 영향과 신뢰도 조사", 「신문과 방송」, 240호, 1990.12, p. 14.
- 허미숙, 「지식 베이스를 이용한 자동색인 시스템에 관한 연구」, 석사학위논문, 연세대학교 도서관학과 대학원, 1991.
- 홍승의, 「신문기사의 자동색인 시스템에 관한 연구」, 석사학위논문, 연세대학교 도서관학과 대학원, 1990.
- Donald. B. Cleveland, And Cleveland, 「색인 및 초록법」, 서울 : 구미무역출판부, 1990.
- Gerald Salton, "A Blueprint for Automatic Indexing", *ACM SIGIR Forum*, vol. 16, 1981, pp. 30～38.
- P. H. Klingbiel, "Machine－Aided Indexing of Technical Literature", *Information Storage and Retrieval*, vol. 9, no. 2, 1973, pp. 79～84.
- M. Dillon, A. S. Gray, "FASIT : A Fully Automatic Syntactically Based Indexing System", *JASIS*, vol. 34, no. 2, 1983, pp. 101～105.
- J. R. Driscoll, et al., "The Operation and Performance of an Artificially Intlligent Keywording System", *Information Processing & Management*, vol. 27, no. 1, 1991, pp. 43－54.
- R. F. Simmons, "Automated Language Processing", *ARIST*, 1, 1966, p. 137.
- Susan Artandi, "Machine Indexing Linguistic and Semiotic Implication", *JASIS*, vol. 27, no. 4, 1976, p. 236.