情報管理研究 Vol. 14 No. 4 1981. 12.

自動索引의 動向과 逆說的 接近

加 藤 德 義 著
高 亨 坤 訳
(KORSTIC 資料管理室)

訳者註：本稿는 Library & Information Science Vol. 15, P.P. 165～180에 실린 加藤徳義의 自動索引の 動向と 逆説的アプローチ를 翻訳한 것으로서 紙面関係上 図表와 参考文献을 省略하였다.

1. 序 言

機械翻譯에 대한 研究가 지금까지의 莫大한 研究投資에도 不拘하고 그 對象인 自然語의 神祕 앞에 継續해서 挫折해온 것처럼 自動索引에 대한 研究도 決定的인 壁에 直面하고 있음은 실로 有感스러운 일이다.

한편 情報量의 增大는 이제 人力에 의한 索引能力의 限界를 威脅하기에 充分하며 情報處理의 現場에서 自動化에 대한 要望은 增大되고 있다. 또한 이러한 緊急性은 自動索引에 대해 보다 現實的인 方向을 要求하고 있다.

本稿에서는 지금까지 發表된 文献의 批評을 통해 앞으로 自動索引이 취해야 할 具體的인 方法의 考察을 試圖하였다. 여기서 提起된 것은 一가령 自動索引研究의 進展을 沮害하는 것과 같은 困難性을 그 對象으로 하는 自然語가 갖지 않는다면 實質的인 自動索引시스템은 存在하지 않을 것이라는 逆説的 發想이다. 그래서 處理對象을 圖書館資料가 아닌 病院内部資料 중 病理學報告書로 정하고 이 逆説的 假説을 展開해 보았다. 病理學報告書의 入力處理는 基本的으로 索引作業임에도 不拘하고 지금까지 自動索引研究者들이 論外로 하고 있지만 위에서 言及한 假説의 検證을 위한 適合한 自然語環境을 具備하고 있다.

2. 自動索引法

2.1 自動索引法의 概要

情報量의 增大에 對應하는 情報의 效果的 検索方法이나 壓縮技術의 開發 必要性은 두말할 나위도 없다. 情報의 壓縮에 대해서는 物理的·形態的 壓縮과 主題的·内容的 壓縮의 両面으로 생각할 수가 있다. 예컨대 하드카피로 存在하는 技術報告書를 마이크로·피시로 撮影하여 保存한다는 行爲는 情報의 物理的·形態的 壓縮이라 볼 수 있다. 또 어느 雜誌記事를 抄録한다는 作業은 그 原論文을 主題的·内容的으로 壓縮하는 것이다. 단 壓縮変換後에도 原論文의 情報의 總量을 전혀 줄이지 않고 抄録이라는 짧은 表現으로 変換할 수는 없다. 또한 主題名標目을 생각할 때 原情報를 内容的으로 追窮하여 하나의 標目을 附與한다는 意味에서도 同一하게 内容的 壓縮이라고 認定된다. 이 경우에도 主題名標目은 自然語에 비해 극히 単純한 構文體系를 갖고 簡潔度도 극히 낮으므로 壓縮도 同時에 損失情報가 많다.

各種의 内容的 壓縮에는 共通的으로 検索이라는 目的이 있다. 이러한 意味에서 情報의 内容的

壓縮은 그 壓縮度의 相異함을 不問하고 主題索引作業(以下, 本稿에서는 특히 限定하지 않는 限 索引은 主題索引을 意味한다)이라고 생각하는 것이 妥當하다.

索引作業은 Meadow가 指摘한바와 같이 語彙와 構文으로 이루어진 索引言語를 使用하여 파일 중에 蓄積된 데이터內容을 代表하는 일이다. 이러한 觀點에서는 抄録, 索引, 分類 등의 行爲를 索引作業으로 一元化하여 생각할 수가 있다.

Baxendale에 의하면 一般的으로 索引에는 다음 3가지 機能이 있다고 한다.

①藏書中에 包含되는 情報에 대해 濃縮된Key를 附與할 것

②文献의 著者와 그것을 探索하는 者와의 意味的 差異에 架橋가 될 것

③藏書中의 文献을 區別하기 위한 道具가 될 것

抄録은 上記의 첫번째 機能을, 分類는 세번째 機能을 主目的으로 한다. 이와 같이 各索引言語는 各各 上記機能의 하나 以上에 力點을 두고 있다. 索引言語의 各各에 의해 索引語의 呼稱도 여러가지이나 基本的으로 같은 文脈으로 捕捉되므로 여기에서는 一般的으로 「索引語」를 使用했고 個個 시스템의 例는 各其의 用語(예컨대 디스크립터 등)를 使用했다.

그리고 情報検索시스템의 經濟費用의대부분이 入力處理段階에서 使用되는 것은 잘 알려져 있으나 이 점에서 索引을 包含한 入力處理面에서 컴퓨터 導入에 의한 經濟的 貢献이 期待되고 있다.

人間에 의한 索引作成에는 2가지의 明白한 問題點이 指摘된다. 하나는 複數의 索引作成者間에서 發生되는 索引結果의 散逸됨이고 다른 하나는 索引作成者의 精神的 状況의 相異, 經驗에 의한 索引作業의 質的 水準 등에 의해 다른 索引結果를 招來하는 것이다.

이에 대하여 컴퓨터 프로그램으로서의 自動索引시스템은 어느 設定된 알고리듬 下에서는 그 알고리듬이 変更되지 않는 한 時間의 經過에 關係치 않고 同一文献에 항시 同一한 索引語를 附與한다. 自動索引은 이같은 均質性이 人間에 의한 索引보다 높은 信頼性을 保障한다.

그러면 索引의 自動化란 具體的으로 어떠한 것인가. 가령 完全한 自動索引시스템이存在한다면 그 시스템은 모든 手作業에 의한 索引作業過程을 自動的으로 行한다. 즉 그와 같은 理想的인 시스템은 自動으로 原文献을 判読하고 內容을 分析하여 文献의 파일에서 自然語의 自動分析에 의해 生成된 索引言語를 使用하여 自動으로 索引語를 附與하고 自動으로 質問을 分析하여 適合文献을 検索한다.

여기에는 基本的으로 2가지 制約이 있다. 첫째는, 自然語는 원래 動的 性質을 갖고 있으므로 어느 새로운 文献에 藏書가 追加될 경우 거기에는 전혀 予測할 수 없는 語彙나 意味가 包含되는 수가 있다. 따라서 文献에 使用된 自然語를 完全히 分析하여 定義할 수 있다고 保證할 수 없다. 둘째는, 現在의 技術로는 人間에 대한 인터페이스로서 印刷되거나 手書된 文献을 機械가 그대로 解読하는 段階로는 아직 이른 入力準備作業으로서 機械可読形態로의 変換의 必要性은 自動索引의 實用化를 위한 障害要因임을 否定할 수가 없다.

따라서 自動索引이란 機械可読文献을 自動的으로 分析하고 自動的으로 索引語를 割當하는 일이라고 定義하는 것이 實際的이다.

索引語를 選定할 때에 文献에 나타난 用語를 그대로 使用하는 自動索引시스템을 自動抽出索引이라 하는데 KWIC索引이 그 典型的 例이다. 이에 반하여 自動割當索引은 索引語로 統制語彙를 使用하는데 語彙統制機能은 通常 辭典 또는 디소러스에 의한다. 索引言語의 特性은 시스템 全體의 業績에 강하게 影響을 미치므로 이와같은 語彙統制装置의 作成方法은 自動·非自動을 不問하고 自動索引의 諸技術과 깊이 關係한다.

抄録이나 分類가 基本的으로 索引과 같은 文脈으로 捕捉된다는 立場에서 自動索引은 自動抄録이나 自動分類의 研究에서 배우는 것이 많다. 自動抄録은 實際로 文中의 內容을 表現하는 用語를 찾아 무엇인가의 得點計算基準에 따라서 自動的으로 文章을 選擇·抽出하는 것이다. 이 경우 用語의 重要度를 反映하도록 Term Weighting方法이 자주 使用된다. 또 自動分類와의 關係는 카테고리名을 어느 文献에 自動的으로 割

當한다는 作業은 그 文献을 카테고리에 分類하는 것과 같은 뜻이다.

自動索引의 性能評價라는 測面에서는 自己의 自動索引시스템을 MEDLARS의 人間에 의한 索引結果와 比較해서 말했던 Salton의 主張에 注目할 必要가 있다.

全自動의 文章分析·検索시스템은 伝統的인 手作業 文献索引시스템으로 얻은 結果보다 低質의 検索性能을 形成하고 있는 것처럼 볼 수는 없다. 手作業에 의한 索引 및 検索式設定은 그 索引作成者나 検索者가 蓄積藏書와 利用者의 要求를 完全하게 理解하고 있을 경우에는 例外的으로 훌륭한 結果를 나타내지만 이와같은 條件에 不合致할 경우에는 대단히 좋지않은 検索結果가 나오기도 한다. 反對로 自動시스템은 綱羅的인 入力데이터와 複雜한 分析方法들에 의해 대단히 나쁜 結果를 나타내는 일은 극히 적고 때로는 滿足할 만한 検索行動을 만들어 낸다.

Salton의 主張에서 代表되는 바와 같이 自動索引의 性能은 通常 手作業에 의한 것과 比較하여 評價되는 일이 많다. 그러나 Salton의 比較研究를 包含하여 一般的으로 比較를 위한 資料가 充分한 量이 아니고 個個의 比較研究가 各各 別個의 藏書를 對象으로 하는 것이 대부분이므로 一括的으로 논하기에는 無理가 많다.

이러한 點에서 1961年의 Borko의 研究와 같이 同一의 藏書에 대한 個個의 自動索引시스템의 比較検討가 必要하다.

2.2 自動索引의 試図

1960年代 初頭에 갑자기 活潑해진 自動索引의 試圖의 標的이 된 것은 1950年代末 IBM의 Luhn이 發表한 一聯의 研究였다. Luhn의 文章分析法은 基本的으로 文献에 나타난 用語의 統計的 特質, 즉 用語의 出現頻度에 의하고 있다. 또한 그는 基本技法을 進步시켜 클러스터(Cluster)化와 分野에 대한 語彙의 相異, 相對頻度 및 共出現頻度 등을 考慮한 方法에 대해 示唆하였다.

Luhn의 自動索引에 대한 貢献은 重要出現頻度語에 의한 文章選定過程에서 抽出한 것이다. 또한 그는 가장 基本的인 尺度로서 文献 i에 대한 用語 k의 出現頻度 f_i 와 用語 k의 藏書에 대한 總出現頻度 F 를 생각하고 다음과 같이 定義하였다.

$$F^k = \sum_{i=1}^{n} f_i{}^k$$

(단, n는 藏書中의 文献數)

이 式은 나중에 모든 自動索引시스템의 基礎가 되었다.

入力上의 困難을 피하기 위해 Luhn은 文献의 Full-Text 대신에 標題만을 使用하고 KWIC 索引을 만들어 냈다. Luhn의 索引技法은 索引語의 人爲的인 統制없이 著者의 用語에 따라서 文献內容을 機械的으로 操作할 수 있다는 可能性을 나타낸 것으로서 評價된다. 따라서 이 技法은 自動抽出索引으로 認定된다.

한편 1958年에 發表된 Swanson의 研究는 自動割當索引의 原型이라고도 할 수 있다. Swanson의 實験에서의 自動化는 단지 미리 分析된 Clue Words에 原文을 機械로 照合한다는 方法이 취해졌다. 여기서는 主題名標目에 몇 개의 Clue Word를 準備하고 文献中에 그 어느 것인가가 發見되면 単純히 對應하는 主題名標目을 割當하였다. 따라서 이 方式은 유니텀索引으로 代表되는 伝統的인 手作業索引方式의 単純한 機械化라고 생각된다.

Maron은 이와같은 伝統的 索引方法을 취하지 않고 確率論的 手法을 導入하였다. Maron의 立場에서는 藏書의 全體를 아는 일은 不可能하고 이같이 不確定한 領域에서는 어느 디스크립터가 發見되었다라고 하는 事實은 限定된 確率에 있어서만 對應하는 主題카테고리에 割當되는 可能性이 있다고 하였다.

이에 반하여 Borko는 因子分析法을 應用하고 Clue Word와 文献의 相關關係를 出發點으로 한 方法을 試圖하였다. 方法論的으로는 Borko의 實験이 그리 높게 評價되지는 않는다고 하나 前述한 바와 같이 다른 自動割當索引法(具體的으로는 Maron의 것)과 同一한 샘플로서 比較했다는 點을 注目해야 한다. 또한 이 比較의 結果 一般的으로는 因子分析法보다 確率論的 手法이 自動索引에 有效하다고 생각하게 되었다.

이외에 Borko의 方法과 類似한 形態의 研究

나 Baker의 潛在構造分析, Williams의 差別係數에 의한 索引語選定을 위한 閾置設定方法의 提案 등이 提示되었으나 1960年代初의 이들 研究는 모두 나중의 研究에 강한 影響力을 갖지 못했다.

1965年에 Damerou는 어느 文献에 包含되는 總語數와 같은 數의 無作爲한 用語의 集合을 생각했을때 어느 特定한 用語가 該當文献에 實際로 나타난 것보다 無作爲한 標本中에 많이 나타나지 않는 確率을 用語의 重要度의 基準으로 하였다. 이 確率分布는 幾何級數的 分布로서 正確히 計算할 수 있다고 되어있다. 標本이 充分히 크고 全體中에 各各의 用語가 적은 比率로 分布되어 있으면 二項分布 또는 포와송分布(Poisson's Distribution)로서 거의 가까와질 수 있다. 어떤 文献을 取擇했을 때 用語의 重要度 rank를 附與하기 위해 포와송標準偏差를 使用하라는 提案이었다. 이와 같은 觀點은 近年에 Bookstein과 Swanson의 用語의 出現分布에 대한 確率論的 모델에 継承되었다. 그들의 假說에 의하면 同一文献中에 集中하기 쉬운 用語는 索引語로서 有效하다. 어느 藏書에서 索引語가 될 수 있는 것과 없는 것 또는 內容을 表現하지 않는 것과 같은 用語와의 區別은 無作爲性에 대한 逸脫度를 말한다. 무엇인가의 統計的 尺度에 의해서 행하여 진다. 가령 非重要語가 多數의 文献 속에 無作爲로 分布되어 있다 하고 또 그分布型을 數學的으로 表現할 수 있다면 重要語는 逆으로 非重要語의 分布를 表現하는 數學的인 分布型이 어느 用語에 대한 實際의 分布를 説明하는데 어느 정도 不適當한가를 測定함으로써 選定할 수 있다.

Bookstin과 Swanson은 重要語란 그 用語가 表現한 內容이 藏書中에서 取扱되는 程度로 보아서 文献의 Class를 區別하는 것과 같은 用語를 생각하여 모든 重要語에 대해 2개의 文献의 Class를 定義하는데 꼭 2가지의 取扱法이 있다 하고 이것을 모델화하기 위해 포와송모델을 適用하였다. 즉 用語 w가 1文献中에 k번 出現하는 確率$P(k)$는

$$P(k)=\pi\frac{e^{\lambda 1}\cdot\lambda 1^{k}}{R!}+(1-\pi)\frac{e^{-\lambda 2}\cdot\lambda 2^{k}}{k!}$$

(단, λ_1, λ_2는 各各의 Class에 있어서 用語의 平均出現數, π는 한 쪽의 Class에 그 文献이 屬하는 確率)로서 表現된다.

Harter는 이 모델을 基礎로 하여 文献d에서 用語 w의 索引性(Indexability)의 尺度 또는 相對重要度 β를 抽出하였다.

$$\beta=P(d\cdot 1/k)+Z$$

그래서 Z는 各各의 Class의 重要度에 따른 效果의 尺度로서 다음과 같이 定義된다.

$$Z=\frac{\lambda_1-\lambda_2}{\lambda_1+\lambda_2}$$

이 β尺度는 人間에 의한 索引으로 割當된 索引語의 単純한 文献內出現頻度보다 항상 優秀한 結果를 나타낸다. β尺度는 特定藏書用 辭典을 作成하기 위한 充分한 索引性을 具備한 語彙의 選定에 有效한 道具라고 생각해도 좋다.

用語의 出現事實을 計量的으로 測定함으로써 키워드를 選定하고자 하는 統計的分析方法에서는 特定한 文脈에 대한 用語의 品詞種類를 識別하거나 特定한 意味를 決定하거나 하는 일은 어렵다. 그래서 言語研究 分野에서 여러가지로 研究되고 있는 構文分析이나 意味分析의 諸理論을 自動索引으로 應用하고자 하는 움직임을 볼 수 있다.

Artandi에 의한 프로젝트 MEDICO는 英文藥學情報의 自動索引 實驗에서 특히 링크의 自動生成에 特徵을 갖는다. 1개의 文章中에 共出現한다. 2개 以上의 用語는 그 文献中의 同一文脈에 同時에 屬한다는 假說에 의해 그들 用語의 特定한 連結關係를 自動的으로 表現하고 模糊性을 減少하고자 하는 것이다. 이 生成實驗의 結果 링크의 適正率은 約 70%이었으며 링크로 表示된 用語와 用語의 距離에 대해 適正이라 判斷된 組成의 平均이 3.71語이었고 不適正이라 判斷된 것은 7.08語라는 結果가 明白하게 되었다. 또 本實驗에서는 入力데이터로서 Full-Text의 경우와 抄録의 경우를 比較하여 文體가 길어짐에 따라 링크生成의 適合率이 低下된다고 指適되었다. 一般的으로 抄録文의 各 文章은 Full-Text보다 길기 때문에 語間距離가 긴 不適正한 確率이 높은 組成을 많이 生成한다고 생각할 것

이다.

Lockheed Palo Alto研究所에서는 9年間에 걸쳐 自動言語解析의 研究가 행해졌고 自動索引에 대한 構文分析法의 應用實驗도 試圖되었다. 同研究에서는 주로 語法의 分析에 의해 品詞의 決定이나 意味用法을 識別하는 등의 可能性이 試圖되었으나 그 중에서도 Word Government라고 呼稱되는 各 用語의 文法的, 意味論的 統治法의 概念을 採用한 것이 注目된다. 이 統治法의 採用에 의해 特定用語에 대한 品詞의 型, 用法의 型, 意味, 다른 用語와의 關係가 整理되었다. 여러 種類의 表를 使用하여 다음과 같은 構文的 또는 意味論的 分析이 상당히 可能하다고 하고 있다.

A.構文分析

①名詞句의 領域設定

②前置詞句修飾의 決定

③不定詞의 定義

④名詞·動詞의 區別의 模糊性 解消

⑤分詞用法의 模糊性 解消

B.意味分析

①統治語의 意味上의 模糊性 解消

②前置詞의 意味上의 模糊性 解消

③主語, 目的語, 述語動詞修飾語의 役割의 模糊性 解消

이외에 構文構造的 關係를 索引法으로 採擇한 것으로서 Hillman의 研究나 Syntol 프로젝트가 代表된다.

지금까지 概觀한 바와 같은 統計的, 構文構造的, 意味論的 諸方法의 集約으로서 이루어진 것이 SMART(Saltons Magical Automatic Retriever of Text)시스템이다. 이 自動文献處理 시스템에서는 주로 蓄積過程에서 構文分析的方法이 檢索過程에서는 統計的 方法이 準備되어 各各의 諸方法은 處理代案으로서 比較檢討가 可能하게 되었다. 文献中의 用語는 語幹 디소러스에 의해 語幹과 語尾로 分離되어 各語幹은 概念코드와 構文코드의 雙으로 換置된다. 語尾辭典과의 對照에 의해서 語尾에도 構文코드가 붙는다. 略語辭典 및 句構造辭典은 各各 名詞句, 前置詞句 등의 識別 및 文章의 木構造分解에 의한 句概念의 識別이 행하여 진다. 이와 같은 詳細한 文章解析을 取擇한 索引시스템의 例는 現在로서는 다른 곳에서 볼 수 없으나 性能의 觀點에서는 다른 시스템과의 比較檢討가 充分하지 않고 安易한 結論은 피해야 한다.

3. 自動索引法의 困難性과 方向転換

3.1 컴퓨터補助索引

概觀한 바와 같이 지금까지 상당한 數의 實驗이 행해졌지만 1960年代 初期의 試驗的인 諸研究와 比較해 보면 最近 10年間의 研究에 基礎的인 進步를 認定하기 困難한 것은 事實이다.그래서 完全한 自動索引시스템은 實現不可能한 것이 아닌가 하는 悲觀的 結論을 내리는 경우가 많다.

人間의 知的 作業을 컴퓨터·프로그램에 의해서 換置하려고 하는 研究는 一般的으로 너무나 複雜하고 難解하기 때문에 Man-Machine 시스템으로 指向하는 傾向이 있으나 自動索引도에의 없이 人間의 知力을 빌렸거나, 단지 컴퓨터가 人間의 作業에 部分的으로 介入한다고 하는 方向의 研究가 旺盛하게 進行되고 있다.

Man-Machine시스템으로서 半自動 또는 컴퓨터補助索引은 索引作業 그 自體를 人間의 知的 活動에 있다고 認定하고 거기에 컴퓨터를 어떻게 利用하여 가는가를 研究하는 것으로 받아들이고 있다.

Barnier에 의하면 索引의 初期 進步의 하나는 ……컴퓨터가 用語의 表示 또는 標準索引語로 翻譯하는 등의 辭典編纂者의 役割을 함으로써 索引者를 도와주는 컴퓨터補助索引일 것이다. 소위 自然語索引作業은 辭典編纂의 過程과 索引의 過程을 잘 分離하여 컴퓨터에서 著者와 索引作成者가 使用한 用語를 찾게하며 또한 標準索引語로 翻譯한다. 컴퓨터에 登録된 語彙에서 찾을 수 없는 새로운 用語는 索引作成者에게 부탁하지 않고 同義語나 이미 만들어진 上位語의 下位概念이라고 判定하여 시스템에 記入할 수 있는 辭典編纂者에게 부탁한다. 이러한 種類의 컴퓨터補助索引은 훌륭한 것이다.

또한 Fangmeyer는 다음과 같이 컴퓨터補助索引의 長點을 列擧하였다.

①抄錄이나 文献中에 나타난 專門的 槪念의 相對的 重要度를 區別할 수 있다.
②모든 文献에 接近할 수 있다.
③그 文献 뿐만 아니라 參考圖書나 專門家 또는 다른 情報源에까지 適切한 索引을 하기 위해 援助를 구할 수 있다.
④文献에 明示되어 있지는 않지만 含蓄된 槪念을 公式化하고 索引하기 위해 帰納的 推理를 할 수 있다(割當索引).
⑤探索要求의 分析, 探索戰略의 公式化 그리고 探索스크리닝에 參加하기보다는 시스템의 利用者 要求에 익숙하고 親熟할 수가 있다.

컴퓨터 補助索引의 典型은 DDC (Defence Documentation Center)의 MAI (Machine－Aided Indexing)시스템에서 볼 수 있다. 이 시스템은 1967年 以來 開發을 繼續해 왔으며 年間 約 1千萬語에 달하는 文献의 索引에 使用되고 있다. 여기서는 英語単語 하나를 엔트리로 하며 "Recognition Dictionary"와 시스템이 許容되는 構文型式의 辭典인 "Format Dictionary"가 使用된다. 前者는 不用語리스트를 겸해 不用語가 되지 않는 用語는 이 辭典에 의해 그 品詞種類를 認識한다. 이 認識의 레벨은 形容詞와 独立해서 索引語가 되는 名詞 그리고 다른 非不用語와 連結해서 索引語가 될 수 있는 名詞 등으로 登録된 用語는 단지 하나의 對應하는 코드를 割當한다. 品詞種類를 表示하는 이와 같은 코드의 예는 形容詞+弱性名詞의 "Interactive Retrieval"의 型式辭典에 照會되어서 그 構文型式이 正當한지 아닌지가 検證된다. 이와같이 하여 出力되는 것이 候補索引語이며 索引者는 이 候補索引語의 리스트에서 妥當한 索引語를 決定한다.

이와 같이 MAI 시스템은 部分的 構文分析을 採用하고 2語 以上의 候補索引語를 提示하며 人間의 索引作業에 補助役割을 하고 있는 것으로 評價된다. 단 用語 各各에 단 하나의 品詞種類를 附與한다는 方法은 柔軟性이 欠如되어 候補索引語의 抽出漏落을 남긴다. MAI시스템에 適合한 예는 오하이오州의 Dayton大學에서의 材料科學分野 資料의 索引시스템에서도 볼 수 있다.

컴퓨터補助索引이 언제부터 생겼는지에 대한 文献的 確認은 困難하지만 여기서는 Doyle의 意味地圖 (Semantic Road Maps)에서 起源을 찾아보았다. Doyle은 Swanson 등이 어떻게 하면 探索者에게 心理的 連想 네트워크를 想起시키느냐라는 問題에 關聯하여 디소러스 槪念을 提示한 것이 그 意味地圖의 出發點이었다고 다음과 같이 말했다.

Swanson 등은 同義語나 關聯語의 디소러스를 이 問題의 解決策으로 提示하였다. 連想地圖는 어느 意味에서 이 解法의 擴張이다. 이것은 巨大한 自動的으로 導出된 디소러스이다. 이와 같은 地圖를 提示함으로써 探索者는 自己가 생각하고 있는 것보다 훨씬 좋은 連想네트워크를 얻을 수 있다.

Doyle은 人間에 의한 索引과 探索이 機械보다 優秀한 結果를 낳는다는 立場에 서서 文章에 쓰여진 連想結果도 人間의 連想行爲의 所産으로서 이것의 逆關係도 成立된다고 생각하였다. 그러므로 그의 意味地圖는 各索引語間의 關聯의 強度나 槪念을 圖型的으로 表示하여 索引作業을 補助하고자 하는 것이다.

또 Artandi가 1961年에 可能性을 提示한 圖書의 巻末索引의 自動作成도 큰 進展이 없이 半自動, 컴퓨터補助索引으로 移行되었다.

1966年 IBM의 Carney가 提示한 시스템에서는 컴퓨터가 重要語候補의 選出, 索引語와 그 用語를 包含한 文章의 印刷, 本文内所在의 明示, 用語의 正順化, 検索用파일의 保持 등에 利用되었다. 이 컴퓨터利用의 結果 經濟的 費用도 人間에게만 依存한 것보다는 低廉하다고 하였다. 費用의 觀點에서는 American Documentation의 索引에 컴퓨터補助索引을 適用하였다. Hines와 Harris도 똑같이 經濟的으로 바람직하다고 하고 있다.

또 Borko는 SAINT (Semi－Automatic Indexing of Natural Text)시스템을 發表하여 圖書巻末索引에 相互介在의 槪念을 導入하였다. 相互介在란 巻末索引의 作成을 連續한 處理過程이라 생각하고 人間과 컴퓨터가 서로 知的 補助와 事務的 處理를 하면서 各各의 中間結果를 받아들인다는 것이다.

最近의 開發로서는 ISI (Institute of Scientific Information)에서 發行하는 Current Contents Weekly Subject Index作成을 위해 컴퓨

터補助索引이 使用되었다.

近年에 특히 注目할 點은 컴퓨터技術一般에 있어서 온라인化한 것과 그에 隨伴한 會話方式의 普及이 컴퓨터補助索引의 實用化를 더욱 促進하고 있는 것이다. 이러한 種類의 會話型 시스템으로서는 IBM의 Bennett이 開發한 NSF(Negotiated Search Facility)와 實用시스템으로서 SSIE (Smithsonian Science Information Exchange, Inc.)의 索引方式이 있다.

3.2 內部索引法

自動索引의 困難性을 理解하기 위해서는 앞에서 말한 圖書卷末索引의 特色을 自動索引의 立場에서 다시 보아둘 必要가 있다.

一般的으로 索引은 어느 文献藏書를 對象으로 하여 거기에 包含된 個個의 文献에 대하여 행하여지는 것이다. 이 때문에 該當文献의 外部(文献藏書)를 意識한 索引이라는 意味에서 外部索引으로서 圖書卷末索引과 區別할 수 있다. 이에 對應하여 後者는 內部索引이라 呼稱할 수가 있다. 왜냐 하면 卷末索引은 그 該當文献인 圖書, 즉 著者의 어느 時點에서의 한 作品을 거의 閉鎖的으로 唯一한 索引對象으로 하여 생각되기 때문이다. 물론 読者의 探索要求에서 생기는 語法的, 意味論的 差異의 調整이 必要하다는 것은 두말 할 나위도 없으나 이것도 역시 外部索引과 比較해서 생각해 보면 明白히 容易하게 對處할 수 있다. 外部索引의 경우 文献藏書에는 여러 種類의 文献形態가 混在할지도 모르며 文献에 따라서 主題가 相異하기 때문에 用語도 여러가지일 것이다. 文献藏書가 充分히 限定된 分野에 集中하고 있는 경우 專門的·技術的 用語는 보다 넓은 分野에 있어서 보다 훨씬 明確한 意味를 갖고있다. 커버하지 않으면 안될 用語의 數도 상당히 줄어들고 語彙의 修正도 훨씬 줄어들 것이다.

內部索引은 이와같은 條件의 極限的 예라고 할 수 있다. 內部索引이 갖고 있는 自然스러운 有利性에 대해 Maloney가 指摘한 바가 있다.

①閉鎖的 性格 : 시스템의 完成後 긴 時間에 걸쳐서 새로운 用語가 入力될 問題가 없다. 어느 圖書의 索引은 그 圖書가 改訂될 때까지 変更될 必要가 없다.

②限定的 性格 : 索引 그 自體는 量的이나 索引深度上에 있어서 過大하지는 않다. 索引部分이 20 페이지 程度의 圖書이면 어떠한 圖書라도 全體를 빠짐없이 索引할 수 있다.

③個性的 性格 : 文章表現上 觀點이나 立場에서 오는 妥協의 必要가 없다. 外部索引이 直面하는 것과 같은 典據리스트나 디소러스 등에서처럼 愼重한 配慮가 內部索引에서는 實際로 별 必要가 없다.

④特定的 性格 : 索引의 各項目의 觀點, 方法, 特定性은 그 文章의 것이며 보다 넓거나 좁게 또는 高度의 背景을 가지고 資料를 찾고자 하는 探索者도 이것을 알고 있으므로 索引項目의 選定이나 記入形式에 대해 影響을 주는 일이 없다.

⑤自己檢證的 性格 : 索引에 의해 要求한 情報에 대한 到達可能性의 與否에 대한 檢證이 端的으로 可能하다.

⑥同義語問題에서의 解放的 性格 : 索引中의 語句는 本文中의 語句와 同一하다.

이와같이 內部索引은 単純한 機械的 處理技術이 比較的 容易하게 應用될 수 있는 有利性을 具備하고 있다.

上記의 有利性은 同時에 索引의 自動化의 困難性을 逆説的으로 말하고 있다. 그래서 Maloney의 指摘을 바꿔 말하면 自動索引의 困難性은 다음과 같이 整理될 수 있다.

①開放的 性格 : 外部索引(以下, 索引이라 함)은 항상 새로운 槪念의 混入과 그것에 隨伴하는 새로운 語句의 發生에 留意하지 않으면 안된다. 이와같은 索引對象의 流動性에서 辭典이나 디소러스 個個의 語句間의 關係 등은 항상 開放的이어야 한다.

②擴散的 性格 : 實際로 索引現場의 要請으로서는 情報의 量的 增大와는 意味가 強한 것은 當然하며 複雜한 自動分析(統計·構文分析 등)을 入力하는 모든 文献에 대해 행하고자 할 때는 無理가 생긴다.

③汎用的 性格 : 文献別로 著者 固有의 觀點이나 立場에 따라 用語用法上의 相異가 생긴다. 따라서 이것들을 包括的으로 處理하고

자 하는 索引시스템은 汎用性을 가져야 하며 그러기 위해서 여러가지 語彙統制上의 煩雜함이 加해진다. 또한 藏書의 專門性이 弱하면 弱할수록 語句의 意味論的 相異가 發生되고 또 이들을 包含하는 디소러스도 肥大해진다.

④一般的 性格 : 索引項目의 選定 등에서 어느 特定의 文献에만 合致하는 方式을 取할 수는 없다. 探索者도 索引과 實際 文献上의 用語와의 差異를 考慮하지 않으면 안된다.

⑤非検證的 性格 : 통상 索引 그 自體와 實際의 文献과는 時間的·空間的으로 各各 独立하여 存在하고 있다. 그래서 索引利用者, 즉 探索者는 索引에 의해서 提示된 情報의 正確性 특히 滿足度를 즉시 検證할 수 없다.

⑥同義語·同音異義語의 問題 : 이것은 汎用的 性格에 包含해서 생각할 수 있는 일이나 어느 特定의 文献에서 使用된 語句는 索引에 使用된 語句와 반드시 一致하지 않는다.

또한 보다 本質的인 問題點을 指摘하면 窮極的으로는 우리들 自身의 커뮤니케이션活動에 대해 아직도 전혀 無知하다는 것이다. 統計的인 解析에서도 원래 解明되지 못한 概念의 聯合은 統計的으로 어느 程度까지 近似할 수 있느냐는 實驗段階에서는 一步도 앞서 있지 못하다. 또 変換文法에서 말하는 것 같은 自然語가 完全히 數學的으로 記述可能하다는 證明은 現在로서는 전혀 없다고 해도 좋다. 또 自然語의 構文構造가 모두 解明되어 모든 構文構造와 組成을 列擧할 수 있다는 保障도 전혀 없다.

自動索引이 지금까지 對象으로 해온 「쓰여진 純粹한 自然語」는 역시 너무 複雜하였다.

3.3 実用化를 위한 逆說的 接近

여기까지는 言語學者 등에 의해서 自然語 그 自體에 대한 解明이 없이는 如何한 自動索引의 研究도 實用化와는 距離가 멀다는 것을 認識하게 해주었다. 그러나 自然語 그 自體가 전에 말한 바와 같은 複雜함을 갖지 않는다고 하면 實用化는 可能할지도 모른다. 즉 內部索引에서 볼 수 있었던 有利性은 무엇인가의 特殊한 環境이 具備되어 있을 경우에는 自動索引이 實用的 시스템으로서 成立可能할 것이다. 전에 指摘한 바와 같은 困難性을 克服해 가는 것이 正統한 研究努力이라고 한다면 그와 같은 困難性이 처음부터 적은 環境下에서 實用시스템을 만들어가는 것은 완전히 逆說이라 할 수 있다. 그러나 이와 같은 逆說的 接近은 만일 그것이 可能하다면 自動索引시스템의 有效性을 살리게 되고 또한 基礎研究의 努力에 刺戟劑가 될 것이다.

그러므로 여기에서는 지금까지의 自動索引·컴퓨터補助索引이 對象으로 해온 出版物形態는 아니지만 高度로 專門化되어 있으며 用語의 意味論的 問題도 극히 적으며 또한 使用되는 文體도 대단히 限定的인 特徵을 찾을 수 있다. 病理學報告書의 入力에 上記의 特殊環境을 구해 보았다. 거기에는 一般的으로 自然語라고 생각할 수 없을 程度의 特殊性을 갖고있지만 情報의 發生者이며 利用者인 病理學者間에는 自然的으로 使用되고 있는 言語環境을 찾아볼 수 있다. 病院이나 醫學研究機關에서는 近年에 컴퓨터가 活潑히 利用되어 그에 따라 病理學報告書도 效率的으로 機械検索하고자 하는 要求가 많아져서 各樣의 研究가 進行되고 있다. 이 중에서 報告書內容의 入力을 自動的으로 해보자는 움직임이 현저하다. 이 作業은 情報科學的 立場에서 索引作業으로 看做할 수 있다.

이 病理學報告書의 自動入力은 지금까지 病院內에서 独自的으로 研究開發되어 왔으며 또 自動索引의 研究者들로부터 注目되는 일도 없었다. 그러나 上記와 같은 逆說的 接近이 可能한 特殊한 環境으로서 自動索引研究의 對象으로 삼을 價値가 있다.

4. 病理学報告書의 自動入力処理

4.1 病理学報告書와 그 自動入力

病院業務의 自動化를 생각할 때 管理會計 등의 Business Applications과 特殊性이 강한 Medical Application으로 나누는 것이 매우 便利하다. 病理學報告書의 處理는 주로 後者의 一部로서 病院에 있어서의 自然語處理라는 位置設定이 可能하다.

어떠한 分野에서도 그때까지 日常的으로 행해온 業務에 컴퓨터를 導入하는 경우 하나의 커다란 心理的 抵抗은 人間이 普通 使用하고 있는 自然語를 디지털·코드로 커뮤니케이션媒體를 바꾸는 것이다. 따라서 自然語데이터가 커뮤니케이션手段의 主가 되는 業務에서는 이것이 큰 障害가 된다.

醫學分野의 專門家들이 患者의 治療를 위해 또 同僚나 學生들에게 醫學的 槪念을 伝達하기 위해서 또는 研究成果를 記述하기 위해서 使用하는 醫學데이터는 많은 경우가 非數値的이며 거의 예외없이 自然語의 単語와 數字와의 合成으로 作成되고 또한 通常 名詞句나 그 変型으로 表現된다. 病院内에서는 이와같은 限定性이 있다고는 하지만 自然語로 쓰여진 文書類가 많이 存在하고 파일化되어 있어서 從來의 方法으로는 検索要求에 응할 수 없는 것이 많다. 이와같은 自然語情報 다시 말해 醫學記述데이터는 주로 實驗室데이터로서 發生하고 報告作業이라는 一種의 索引作業을 經由하여 蓄積된다.

病院이나 醫學研究機關에서 自然語處理의 研究가 행하여지게 된 것은 近間 約 10年程度의 아주 最近의 일이다. 그리고 이 期間의 文献을 調査해 보면 이러한 種類의 研究는 거의 病理學報告書를 對象으로 하고 있음을 알 수 있다.

病理學者가 通常 取扱하고 있는 記述的 内容의 報告書는 주로 外科病理學報告書와 剖検報告書이다. 外科病理學報告書는 外科的으로 얻어진 巨視的, 微視的 特徵의 記述 및 病理學者의 所見事項에 대한 解釋으로 되어있다. 剖検報告書도 同一하나 이것은 複數의 器管系의 死後検査에 관한 詳細한 記述이라는 點이 다르다. 데이터發生量의 基準으로서는 베드(Bed)數 500의 病院에서 年間 1萬~1萬5千件의 外科病理學報告書와 300~500件의 剖検報告書가 發生하고 있다.

機械를 使用하지 않았던 從來의 報告方式은 病理學者가 組織検査時에 所見事項을 口述하는 것이다. 이 口述内容은 報告書로 作成되어 患者의 病歷記録에 附加되거나 다른 파일에 蓄積된다. 이들 報告書는 患者의 治療에 直接, 間接으로 有用하다. 또 醫學研究教育에 高價의 情報를 提供한다

病理學報告書에 記入되는 데이터는 크게 나누어 人口統計學的 데이터와 醫學診斷記述의 2種類가 있다. 이 중에서 人口統計學的 데이터는 身長, 體重, 年齡 등의 數量的인 것이나 性別 등의 擇一的 性格의 데이터가 大部分이며 比較的 単純한 機械處理가 可能하다. 이에 반해서 診斷記述은 自然語로서 作成되어 코멘트形式으로 補完的으로 作成된 部分도 있지만 거의 高度로 專門化된 病理學내지 醫學用語로서 名詞句나 그 変形으로 口述記入된다.

伝統的으로 이들 診斷記述은 ICD(International Classification of Diseases) 등의 標準코드表를 使用하여 코드化되어 検索要求에 對應해 왔다. 그러나 發生데이터 量의 增大와 코드化를 위해서 高度의 專門的 知識을 가진 專任者가 必要하다는 등의 問題로 自動코딩의 可能性이 研究되어 왔다.

커뮤니케이션手段이 言語學的 困難이 比較的 적은 限定的 自然語라는 點에 着眼하여 上記의 問題에 對處하고자 試圖를 꾀한 것은 Smith 와 Melton이 처음이었다. 그들의 指摘에 의하면 形態學的 診斷과 検索要求가 적고 単純한 것이었기 때문에 自動化가 容易했다. 自動코드化 시스팀의 實現과 各病院의 파일整備를 통해 疾病의 地域的 分布 등을 調査하기 위한 코드化된 保建데이터의 國家的 登録시스팀까지 폭넓은 提案이 되었던 것이 注目할 만하다.

4.2 現在시스팀事例

4.2.1 NCI病理学情報시스팀

1966年에 Pratt와 Thomas가 發表한 NIH(National Institutes of Health)의 National Cancer Institute Information Processing System for Pathology Data(NCI病理學情報시스팀)은 病理學報告書 파일의 生成, 維持, 探索과 파일 중의 레코드의 識別과 検索 그리고 統計諸表의 印刷 등의 機能을 갖고 있다. 取扱되는 레코드는 剖検, 外科病理學, 細胞病理學의 各報告書로서 4萬件 以上의 레코드가 収録되어 있고 診斷記述은 約 16萬件에 달했다. 診斷記述에는 S-

NOP (Systematiged Nomenclature of Pathology)가 使用되었다. 이는 病理診斷의 最少単位로서 使用되는데 SNOP는 美國病理學會가 病理學資料의 組織化를 目的으로 하여 作成한 것으로서 病理學的 所見을 記述하는 것인데 여기에는 影響을 받은 部位名(局所解剖學), 疾病에 의한 組織의 形態的 変化(病理形態學), 病原體나 薬品(病因學) 및 生理的·化學的 異常 또는 変化(機能)의 4가지 파시트로 分析記述할 수 있도록 되어 있다. 예를 들면 "Metastatic chorioca rcinoma in lungs and liver"라는 所見은

T 2800 M 8809 E 0000 F 0000 LUNGS,CHORIOCARCINOMA

T 5600 M 8809 E 0000 F 0000 LIVER,CHORIOCARCINOMA

와 같은 局所解剖學(T)의 파시트가 2種(lungs, liver) 있기 때문에 2개의 엔트리로 記述된다.

Graepel 등이 指摘한 바와 같이 이와같은 코딩構造는 SNOP를 使用하는 者가 그들 自身의 検索目的에 맞추어 設定하기 위한 道具에 그친다는 意味에서 ICD와 같은 用語表와는 다르다. 또 用語間의 關係를 表現하고 있지않기 때문에 探索構造의 設定을 위한 配慮를 附加하는 일이 SNOP를 病理學情報시스팀에서 디소러스의 位置設定에 必要不可欠한 것이 된다.

이와같이 주로 名詞句로 表示되는 病理學的 所見인 SNOP코드에 自動코드化를 全面的인 目標로 삼고있는 點이 注目할 點이다.

4.2.2 몬트리올綜合病院

카나다의 몬트리올綜合病院에서는 隣接한 마키르大學附屬病院의 施設을 包含한다는 前提下에 컴퓨터에 의존한 病理學報告書의 處理를 위해 CISP (Computerized Information System for Pathology)를 設計하였다. CISP의 對象은 NCI 病理學報告書시스팀과 同一한데 이중에서 가장 많은 外科病理學報告書는 年間 約 12萬件에 달한다.

이 外科病理學報告書는 識別, 臨床, 概略記述, 病理診斷의 4가지 部門으로 構成되어 臨床 및 病理診斷의 両部門에 대해서는 시스팀에 入力된 文章이 分析되어 內部辭典을 使用하여 統制語彙에 코드化된다. 이 辭典은 単純하지만 階層關係를 表示할 수 있도록 構成되어 있는 點과 同義語나 主題카테고리 등을 統制할 수 있도록 配慮되어 있는 點 등의 디소러스 機能을 具備하고 있고 作成은 ICDA (ICD Adaptations)와 SNOP를 合成한 것이 注目된다.

SNOP는 이 分野에서 가장 잘 普及된 用語集이기는 하지만 各様의 診斷記述을 再表現해야 하고 記入項目이 지나치게 詳細化되어 있으며 臨床과 病理의 相關에 必要한 臨床醫學用語가 欠如되어 있는 등의 不合理한 點이 있으므로 이들 欠點의 一部를 ICDA가 臨床診斷面의 코딩의 容易化를 試圖하고 있다.

4.2.3 UCLA医学附属病院

病院에 있어서 自然語處理시스팀으로서 最大規模이고 가장 넓은 對象分野를 가진 것이 UCLA 醫學部附屬病院의 Natural Language Retrieval System이다. 이 시스팀도 病理學報告書處理를 出發點으로 하고 있고 現在 解剖, 骨髓, 神經放射線, 核醫學을 包含한 5分野의 報告書類를 對象으로 하고 있다.

앞에서 Smith와 Melton이 指摘한 바를 證明한 것처럼 다음의 問題들이 이 시스팀의 開發理由이다.

①適合한 文章의 検索을 위해서는 特定한 用語 또는 그 集合의 存在與否만으로도 充分히 文章을 分別할 수 있으며 綿密한 構文論的·意味論的 分析을 행할 必要가 없다.

②文章에는 거의 모두 그 自體의 意味가 明瞭한 디스크립터가 있다.

③原語를 그대로 컴퓨터에 蓄積하여 検索時에 全文을 探索하는 것은 時間的·經濟的으로 不利하다.

④이에 반하여 컴퓨터에 의한 自動코딩은 自然語 그대로 入力할 수 있으며 自然語로 変換하여 出力도 할 수 있다.

이 시스팀의 特徵은 디소러스가 辭典과는 独立해 具備되어 있는 것이다. Lamson－IBM 디소러스라고 通稱되어 있는 이 디소러스는 2進整數로서 表現된 키워드內部의 코드間의 關係를

統制하고 辭典은 그들 內部코드와 自然語의 인터페이스로서 使用되고 있다. 이들 關係概念에는 ①類義클래스를 設置하여 用語를 그룹化하는 類義性 ②類義語 클래스를 連結함으로써 親세트와 子세트 간의 從屬性 ③論理關係와 呼稱되는 相互結合·네트워크 등이 包含된다. 디스크립터의 更新은 科學用語의 流動性을 考慮하여 容易하게 할 수 있도록 設計되어 있다. UCLA 病理學 디소러스와 UCLA 核醫學 디소러스 2種類가 具備되어 있다.

4.3 病理学報告書시스팀의 問題点과 将来

지금까지 살펴본 것처럼 病理學報告書는 一般文獻보다 言語學的 困難性이 적으므로 情報處理技術上의 問題點도 적다고 생각하기 쉬우나 索引精度의 要求에 대해서는 一般의 醫學文獻情報보다 훨씬 높고 診斷情報의 索引性能에 대한 信賴性은 索引上의 노이즈의 容認度로서 表示했을 때 1%를 充分히 下廻할 必要가 있다는 意見이 나왔다.

診斷의 確信度나 否定的 診斷에 대해서도 検索精度의 要求에 따라서 取扱을 確實히 해야 한다. 一般的으로 키워드의 存在 有無만으로 単純히 肯定的 病理診斷을 索引하고 蓄積하는데 그치지만 보다 高度의 시스팀에서는 이러한 問題를 配慮할 必要가 있다. SNOP의 경우는 코드를 別途로 割當하는 方法을 취하고 있으나 여유코드가 現在도 不足한 SNOP에서는 더욱 그 柔軟性을 낮게하고 있다고 할 수 있다. 自動入力에서는 否定的 診斷을 蓄積하느냐 마느냐 以前에 否定的 診斷이냐 아니냐의 認識이 問題가 된다. 否定語와 키워드의 共出現만으로 捕捉하는 알고리듬에서는 重要한 키워드의 索引漏落을 惹起하는 경우도 있고 또한 이것이 入力을 위한 무엇인가의 口述規制(즉, 擬似自然語入力)가 좋은 結果를 가져온다.

病理學關聯의 用語集으로는 SNOP, Lamson-IBM. 디소러스, ICDA 디소러스의 自己開發까지를 包含하여 設計上의 混亂을 招來할 뿐만 아니라 將來的인 地域醫療 시스팀이나 連邦規模의 病理學情報交換의 統合을 생각하면 그 適合性의 觀點에서 커다란 障害가 된다. 그러나 一般的 傾向으로서 語彙標準化를 위한 努力은 徵候名의 分野에서 独逸의 DOFONOS (German Syndrome Identification and Information System)의 活動이 認定될 程度로 美國에서는 活潑하지 않다.

自動코딩時에는 決定論的인 用語의 割當이 適合하기 때문에 統計的 手法이 適用된 예가 적으나 登錄해야 할 키워드나 스톱워드의 選定에서는 統計的 手法이 利用可能하다. 단 키워드의 選定은 病理學專門家의 判斷없이는 實用이 어려우므로 스톱워드選定에 관해서 노이즈의 低減化, 索引漏落의 回避 등을 위한 有效한 判定資料로서 예컨대 Harter의 β尺度가 利用될 수 있다. 初期의 自動索引研究에서 旺盛히 使用된 相關分析은 完全히 病理學研究領域 그 自體에 屬하는 것이며 情報시스팀이 直接 關與할 性質의 것은 아니다.

經濟的 觀點에서는 本來 情報検索시스팀에 있어서 그 大部分의 費用을 點有하는 入力을 위한 人件費를 節減하는 意味에서 問題가 없으나 이와 같은 效果는 큰 시스팀에서 처음으로 얻어지는 現狀이다. 이 點에서는 病理學報告書의 處理要求가 大規模의 病院이나 研究施設로 集中하고 있으므로 現在의 方向이 반드시 좋지 않다고만 말할 수는 없다.

또한 自動入力에 있어서는 SMART시스팀과 같은 複雜한 句構造解析의 準備는 오히려 시스팀의 經濟的 實用效率을 低下시킬 염려가 있으므로 構文分析的으로는 分詞句, 前置詞句의 認定이나 句構造의 正當性 検證에 그치는 것이 實際的이라 할 수 있고 또 診斷情報의 言語的 性質에서도 이 程度의 解析에 適合한 것이다.

今後의 報告書處理는 病院自動化의 一般的 趨勢에 따라서 온라인化되리라고 생각된다. 따라서 自動入力도 會話型으로 행하여진다. 컴퓨터補助索引形態를 指向할 것이다. 이경우 辭典이나 디소러스를 온라인으로 照會하고 人間이 畵面表示를 통해 모니터하는 것이 現實化될 것이다.

5. 結 論

지금까지 自動索引研究의 足跡을 찾아서 그것

을 整理하고 病理學報告書의 自動入力을 觀察하였다. 이와 같은 病院에서의 制限된 情報活動이 文献의 自動索引에 대하여 어떠한 具體的 貢献을 하였는가를 考察하여 그것을 本稿의 結論으로 하였다.

그것은 ①自動索引法의 基礎的 研究成果에 대한 貢献 ②自動索引의 實用化 促進 ③圖書館學·情報學의 한 分野로서 自動索引研究의 他領域에 대한 直接的 貢献의 3가지로 集約된다고 생각한다.

①의 例로서는 否定의 取扱이 있다. 構文構造가 너무 複雜하고 用語의 意味論的 曖昧性이 높은 경우 否定語의 取扱에 各様의 構造解析과 例外의 設定이 必要하며 또 그 決果의 信頼性도 대단히 낮은데 대하여 逆説的 接近에 의해서 限定된 分野에서는 結果의 評價도 容易하고 限定된 句構造패턴으로 否定을 取扱할 수 있게 되었다. 이와같이 結果가 評價하기 쉬운 것은 諸手法의 有效性을 検證하기 위한 材料提出에 대단히 適合하다.

本稿에서 가장 力點을 둔 것은 自動索引의 實用化 問題이다. 實現 不可能하다고 研究 그 自體가 停滯되었던 自動索引은 이와같은 分野에서 着實하게 育成되어 가고있다. 그 높은 實用性이 認定되어야만 基礎研究도 活氣가 생길 것이며 또한 社會的으로 自動索引을 收容할 素地를 만들고 確立해 가는 것이 必要하다. 그러한 意味에서도 病理學報告書나 다른 病院內部의 資料處理에만 局限하지 말고 積極的으로 實用이 可能한 環境을 開拓해야만 한다. 同時에 文献資料의 處理라고 하는 自動索引의 從來的인 思考方式에서 進一步하여 圖書館이라고 하는 情報시스템의 類型에 拘束되지 말고 다른 情報시스템, 다른 領域으로 開拓해 나감으로써 自動索引을 널리 社會的으로 알리고 그 有效性과 汎用性을 表現할 必要도 있다.

그리고 우리들 目前에는 情報의 絶對量의 過剩增加가 深刻한 問題이므로 實用的인 自動索引시스템의 確立은 現時點에서 絶對的으로 必要하다.

마지막으로 本稿에서는 研究對象을 美國에만 限定하여 漢文文化圈에서의 自動索引을 생각해보면 漢字의 機械處理 등이 近年에 刮目할 만큼 發達했지만 入力上의 絶對的 煩雜性과 高價의 費用 등을 否定할 수 없으며 自動索引의 研究事例도 극히 적다는 점 때문에 美國과의 同一한 論議는 피해야 하나 그 重要性은 認知해야 한다.