效率的인 데이터베이스 製作과 情報檢索을 위한 自動綴字校正

Automatic Spelling Correction for Efficient Data Base Production and Information Retrieval

김 병 혜*
(Kim, Byung-Hye)

抄 錄

本稿는 書誌 데이터베이스製作과 情報檢索觀點에서 自動綴字校正에 對하여 論한다. 여기에는 흔히 발견되는 綴字誤謬의 類型과 書誌 데이터베이스內의 綴字誤謬가 미치는 影響에 대하여 記述하고 있으며, 自動綴字校正시스템의 일반적 구성에 관해서는 文書의 정규화, 綴字檢證, 綴字校正, 使用者 인터페이스로 나누어 記述하고 있다.

ABSTRACTS

This paper discusses automatic spelling correction in a point of view bibliographic Data Base production and information retrieval. Types of commonly detected spelling errors and impact of spelling errors in bibliographic data bases are described here. Document normalization, spelling verification, spelling correction and user interface for general construction of automatic spelling correction systems are described.

* 産業研究院 附設 産業技術情報센터 電算室.

Ⅰ. 序　　論

文獻 데이터베이스가 일반화되면서 컴퓨터는 데이터베이스 製作上 문제가 되는 校正作業(proofreading)을 돕고 있다. 맞춤법이 틀린 單語를 찾아 고치고, 띄어쓰기를 해주며, 文法이 틀린 文章을 찾아 이에 대한 助言을 주는 등 그 역할은 다양하다. 이러한 기능은 워크시트(worksheet)作成者에게 주어지는 指針書(使用文字 및 用語表記法 포함)의 상당 부분을 커버하며, 키인(key-in)과정에서 발생하는 오류에 대한 處理를 하여 준다. 현재 校正作業을 돕는 시스템은 크게 다음과 같은 세 부류로 나누어진다.

① 綴 字 校 正 (spelling correction)

入力텍스트를 토큰(스페이스, 콤마, 점, 콜론 등에 의해 區分되어지는 單位를 말하는데, 앞으로 토큰과 單語는 같은 의미로 쓴다)으로 나누어 토큰 單位로 處理한다. 즉, 맞춤법이 틀린 單語를 찾아 이를 올바른 單語로 校正한다.

② 文 法 檢 査 (grammar checking)

綴字校正을 포함하는 좀 더 포괄적인 校正이다. 이는 綴字校正만으로는 할 수 없는 것들, 가령 文法的으로 맞지 않는 單語를 찾아 이를 바르게 校正하는 것 등을 포함한다.

③ 文 體 檢 査 (style checking)

이상적인 校正을 추구하는 것으로 綴字校正과 文法檢査를 포함한다. 이는 入力텍스트에서 복잡하거나 어색한 文章을 찾아 이를 평이한 文章으로 校正하는 것을 目標로 하고 있다.

그러나 이 중에서 綴字校正시스템만이 현재 기술로 볼 때 가장 실용적이며 또 널리 쓰이고 있다. 文法檢査와 文體檢査시스템은 아직까지는 실험적인 수준이다.

本稿는 우리말 데이터베이스 製作時 적은 努力으로 당장 도움이 될 것으로 기대되는 綴字校正시스템만을 관심의 대상으로 삼고자 한다.

綴字校正시스템에 관한 硏究는 1957년경부터 시작되었는데, 초기에는 특정문맥에 대하여 特定入力裝置에서 비롯되는 오류를 발견하고 고치는 쪽에 관심이 집중되었었다. 또한 초기에는 旅客機의 搭乘者 名單의 오류를 찾는다든가, 通信用 모르스 부호를 인식한다든가, 컴퓨터 프로그램의 컴파일 오류를 校正한다든가 등의 문제에 대한 硏究도 있었다. 그러나 초기의 硏究結果는 실용적인 면에서 보면 그

가치가 별로 없는 실험용에 불과했다.

1971년 Ralph Gorin이 만든 DEC-10에서 운용되는 SPELL은 최초의 應用綴字·檢査機로 기록되고 있다. 이 프로그램은 그동안 여러 번 수정이 되어 현재까지도 널리 이용되고 있으며, 우리에게 많은 교훈을 주고 있다.

온라인 데이터베이스의 급격한 증가는 綴字校正에 대한 硏究方向을 기계 가독형 텍스트(machine readable texts)의 교정쪽으로 바꾸어 놓았다. 이는 많은 데이터베이스 專門生産機關들이 그들의 生産性 向上과 他生産機關보다 데이터베이스의 質的 優位를 점하기 위한 戰略에도 기인한다. CAS(Chemical Abstracts Service)에서는 SPEEDCOP(SPElling Error Detection COrrection Project)이라는 프로젝트를 통해 이와 같은 綴字校正시스템을 開發하여 활용하고 있다. 또 電氣·電子分野의 데이터베이스를 제공하는 것으로 유명한 INSPEC(INformation Service for Physics, Electronics & Computing)에서도 데이터베이스 製作過程에서 綴字校正시스템의 도움을 받고 있다.

本稿에서는 데이터베이스 製作과 情報檢索側面에서 고찰한 철자교정 시스템을 분석·기술하도록 하였으며 우리말을 중심으로 한 철자교정 시스템은 강재우(1990) 등이 수행한 硏究들이 다소 있지만 모두 實驗的인 수준에 그치고 있어 영어권에서 개발·이용되고 있는 것들을 중심으로 분석 기술하였다. 이러한 고찰을 통해 아직 태동기에 있는 우리말 데이터베이스 製作에 일조하며 보다 向上된 情報서비스가 이루어지기를 기대한다.

Ⅱ. 書誌 데이터베이스내의 綴字誤謬의 結果

Bourne(1977)은 11개의 온라인 데이터베이스 (〈表 1〉)를 대상으로 檢索시스템이 생성한 索引語의 綴字誤謬에 대하여 다양한 分析을 하였다. 이는 "APPLE"에서 "AQUA"까지, "GRAPE"에서 "GREECE"까지, 그리고 "PLUM"에서 "PLUTO"까지의 세가지 범주를 調査對象索引語로 선정하여 샘플 데이터베이스에서 각 범주에 대한 誤謬를 조사한 것에 바탕을 두고 있다. 여기에서 綴字誤謬는 크게 철자가 틀린 誤謬와 띄어 쓰기가 잘못된 오류의 두가지를 지칭한다.

〈表 2〉는 각각의 데이터베이스에서 발견한 綴字誤謬의 頻度를 나타낸 것인데,

〈表 1〉 Bourne이 調査한 데이터베이스

ERIC	complete Research in Education and Current Index to Journals in Education files from ERIC
CAC	Chemical Abstracts Condensates(1972－)
BIOSIS	BIOSIS Previews(1972－)
NTIS	complete Government Reports Announcements(1964)
SSCI	Social Sciences Citation Index(1972－)
EI	COMPENDEX from Engineering Index(1970－)
CAIN	complete CAtaloging and INdexing from National Agricultural Library
PA	Psychological Abstracts(1967－)
ISMEC	Information Service in MEChanical Engineering from INSPEC(1973-)
ABI	Abstracted Business Information INFORM(1975－)
PATS	Chemical Market Abstracts and Equipment Market Abstracts from Predicasts

여기에서는 英語가 아닌 言語의 綴字誤謬가 고려되지 않았으며, 單語의 변형형인 略語도 誤謬에 포함되지 않았다. 따라서 실제 誤謬의 頻度數는 더 높다고 할 수 있다.

Bourne이 調査對象으로 선정한 索引語는 전체 索引語의 0.13～0.52% 정도 밖에 되지 않아 그 信賴度가 높다고 할 수 없지만 이러한 誤謬를 측정하기에는 충분하다. 〈圖 1〉에서 볼 수 있듯이 索引語의 誤謬率은 좋게는 0.5% 미만에서 나쁘게는 23%까지 데이터베이스에 따라 다양하다.

그러면 이러한 綴字誤謬가 우리에게 미치는 영향에 대하여 데이터베이스 供給者의 立場과 檢索서비스를 提供하는 기관의 입장과 이용자의 입장에서 살펴 보자.

데이터베이스 供給者는 데이터베이스내의 綴字誤謬로 인하여 데이터베이스 內容이 든 테이프의 販賣에 어려움이 따를 수 있는 등 큰 타격을 받을 수 있다. 비교적 높은 比率의 綴字誤謬를 가진 테이프를 계속적으로 공급한다면 抄錄과 索引서비스의 지대한 공헌조차도 희미해질 것이며, 테이프 購買者가 이러한 誤謬를 심각하게 받아 들인다면 契約을 파기하고 다른 供給者에게로 눈을 돌릴 것이다.

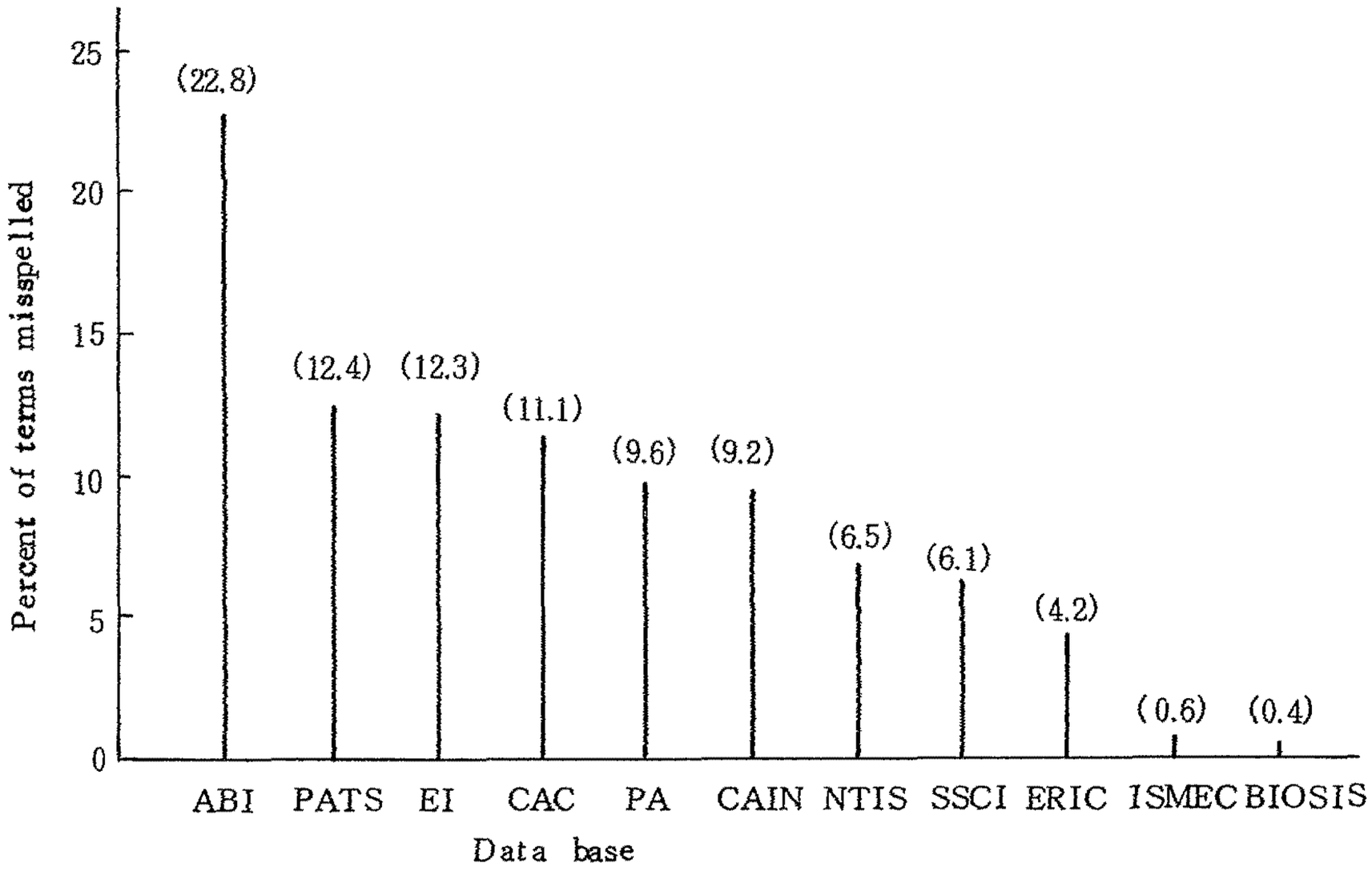
일반적으로 기계 가독형 레코드는 册子型 索引이나 抄錄을 만드는데 이용된다. 따라서 테이프 레코드내의 綴字誤謬는 印刷物에 그대로 반영될 뿐 아니라 抄錄이나

〈表 2〉 Bourne이 선정한 세가지 범주에 대한 데이터베이스別 綴字誤謬의 頻度

APPLE-ACUA	COMBINED FILE	ERIC (FILE1)	CA (FILE3)	BIOSIS (FILE5)	NTIS (FILE6)	SSCI (FILE7)	EI (FILE8)	CAIN (FILE10)	PSYCH AB (FILE11)	INSPEC ISMEC (FILE14)	ABI INFORM (FILE15)	PATS CMA-EMA (FILE16)
TOTAL NUMBER OF TERMS:	1,352	57	226	259	273	118	227	493	201	133	167	226
TOTAL NUMBER OF TERMS MISSPELLED:	268 (19.8)	2 (3.5)	45 (19.9)	2 (.8)	37 (13.6)	14 (11.9)	51 (22.5)	74 (15.0)	29 (14.4)	1 (.8)	55 (32.9)	49 (21.7)
TOTAL NUMBER OF CITATION POSTINGS:	198,299	2,137	26,197	13,812	41,244	6,261	26,381	20,834	30,067	1,938	9,314	20,114
TOTAL NUMBER OF CITATION POSTINGS TO MISSPELLED TERMS:	576 (.29)	2 (. 09)	79 (.30)	2 (.01)	45 (.11)	20 (.32)	77 (.29)	144 (.69)	43 (.14)	1 (.05)	80 (.86)	83 (.41)
GRAPE-GREECE												
TOTAL NUMBER OF TERMS:	1,822	73	241	296	518	125	373	539	149	189	83	211
TOTAL NUMBER OF TERMS MISSPELLED:	104 (5.7)	3 (4.1)	13 (5.4)	0 (0)	17 (3.3)	2 (1.6)	23 (6.2)	33 (6.1)	8 (5.4)	1 (.5)	7 (8.4)	8 (3.8)
TOTAL NUMBER OF CITATION POSTINGS:	117,030	2,002	14,137	8,858	37,034	1,820	8,625	20,387	14,507	746	4,409	4,505
TOTAL NUMBER OF CITATION POSTINGS TO MISSPELLED TERMS:	174 (.15)	16 (.80)	14 (.10)	0 (0)	36 (.10)	2 (.11)	23 (.27)	56 (.27)	10 (.07)	1 (.13)	8 (.18)	8 (.18)
PLUM-PLUTO												
TOTAL NUMBER OF TERMS:	434	12	102	153	86	34	57	129	35	37	31	45
TOTAL NUMBER OF TERMS MISSPELLED:	19 (4.4)	1 (8.3)	5 (4.9)	1 (.6)	3 (3.5)	1 (2.9)	7 (12.3)	0 (0)	0 (0)	0 (0)	2 (6.4)	3 (6.7)
TOTAL NUMBER OF CITATION POSTINGS:	11,513	78	967	2,156	2,286	302	1,264	1,427	1,074	129	560	1,270
TOTAL NUMBER OF CITATION POSTINGS TO MISSPELLED TERMS:	25 (.22)	2 (2.56)	7 (.72)	1 (.05)	3 (.13)	1 (.33)	6 (.47)	0 (0)	0 (0)	0 (0)	2 (.36)	3 (.24)
TOTAL												
TOTAL NUMBER OF TERMS:	3,608	142	569	708	877	277	657	1,161	385	359	281	482
TOTAL NUMBER OF TERMS MISSPELLED:	391 (10.8)	6 (4.2)	63 (11.1)	3 (.4)	57 (6.5)	17 (6.1)	81 (12.3)	107 (9.2)	37 (9.6)	2 (.6)	64 (22.8)	60 (12.4)
TOTAL NUMBER OF CITATION POSTINGS:	326,842	4,217	41,301	24,826	80,564	8,383	36,270	42,648	45,648	2,813	14,283	25,889
TOTAL NUMBER OF CITATION POSTINGS TO MISSPELLED TERMS:	775 (.24)	20 (. 47)	100 (.24)	3 (.01)	84 (.10)	23 (.27)	106 (.29)	200 (.47)	53 (.12)	2 (.07)	90 (.63)	94 (.36)

註 : ( ) 안은 %임.

〈圖 1〉 데이터베이스別 綴字誤謬의 比較



索引의 蓄積에 있어 가외의 접근점 (extra access point) 을 갖게 한다. 이는 파일 크기를 증가시키고 校正과 編輯을 어렵게 만드는 原因이 된다.

대부분의 온라인 檢索시스템에서는 각각의 檢索接近點에 대한 索引이 저장되어 있다. 이런 온라인 파일에 수년씩이나 綴字가 틀린 項目이 저장되어 있다면 기억장소의 낭비뿐 아니라 가외의 컴퓨터 使用時間과 같은 중요한 運營上의 부담을 초래 할 수 있다. 결과적으로 테이프 供給者가 틀린 철자를 제거한 정련된 테이프를 提供한다면 온라인 檢索서비스 費用은 훨씬 절약될 수 있을 것이다.

檢索서비스만을 하는 중간 위치에 처해 있는 기관들은 테이프 供給者로부터 제공받은 誤謬가 있는 레코드로 인해 利用者로부터 비난을 받을 수 있다. 실제로 情報檢索을 하는 대부분의 利用者들은 誤謬에 대한 책임을 테이프 供給者보다는 서비스機關에 묻는다. 利用者들은 근본적인 실수가 테이프 供給者에게 있다는 것을 알더라도 서비스機關이 그 오류를 고쳐 品質管理를 해야 한다고 생각한다. 그러나 서비스機關이 이와 같은 誤謬를 校正하는 것은 運營上 많은 어려움이 따른다. 그렇다고 이런 誤謬를 그냥 방치할 수 없는 것이 檢索서비스機關의 이미지와 信賴度가 손상되어 利用度가 떨어지는 문제가 제기될 수 있기 때문이다.

綴字誤謬는 檢索시스템 내부의 기생충으로 간주될 수 있다. 왜냐하면 綴字誤謬는 온라인 파일 크기를 증가시키고, 컴퓨터가 非效率的으로 運營되도록 하며, 원하지

않는 檢索結果를 利用者에게 안겨 주기 때문이다.

利用者는 綴字誤謬로 인한 손해를 가장 직접적으로 받는다. 綴字誤謬는 서비스 料金의 일부를 차지하므로 利用者에게 가외의 費用이 더 들게 하며, 필요없는 檢索結果를 유발시켜 利用者의 귀중한 시간을 빼앗는다.

또한 利用者는 綴字誤謬로 인해 書誌事項中 資料出處에 관한 중요한 情報를 잃어 버릴 수 있다. 틀린 文字가 오른쪽에 치우쳐 있다면 문자열이 유사한 單語에 의해 보완될 수도 있지만 앞글자가 틀린 경우에는 찾을 수 없다. 예를 들면, "情報管理研究"가 "情報管理究"로 된 綴字誤謬는 "情報管理$"에 의해 찾아질 수 있지만, "報管理研究"로 된 綴字誤謬는 찾을 수 없다.

그리고 綴字誤謬는 포괄적으로 檢索하여 재현율을 높이고자 할 때에도 利用者의 時間과 努力을 많이 빼앗는다. 문자열이 유사한 檢索項目에서 利用者는 올바른 항목과 철자가 틀린 항목을 골라내야 하며, 그것을 모든 檢索 質疑語에 포함시켜야만이 원하는 結果를 얻을 수 있다. 다시 말하면 재현율을 높이기 위해서는 원래의 檢索項目外에 가외의 항목을 追加시켜야 하는데, 이때 관련이 없는 綴字誤謬單語로 인해 관심어가 방해를 받는다면 利用者는 문자열의 範圍를 정확히 정해 이를 여러 개의 명령어로 나누어 檢索해야 한다.

Ⅲ. 綴字誤謬의 類型

일반적으로 綴字誤謬는 著者의 無知(author ignorance)나 타이핑 誤謬이거나 시스템 에러로 발생한다. 여기에서 著者의 無知에 의한 誤謬는 지속적인 특성 즉, 한번 범한 誤謬를 계속해서 범하는 특성이 있다. 이 誤謬는 보통 단어의 發音과 맞춤법이 서로 다른데서 오는 경우가 많다.

入力 誤謬는 키보드에서 키인할 때 생기는 誤謬이다. 현재 온라인 데이터베이스에서는 入力 오류가 다른 어떤 誤謬보다도 頻度數가 많으며 가장 문제가 된다. 이 오류는 키보드의 자판 배열에 관계가 있거나 키인하는 손가락의 움직임에 관련이 많다. 따라서 이 誤謬는 어느 정도 豫測이 가능하다.

시스템 에러에 의한 綴字誤謬는 文字暗號化(character encoding)및 전송(transmission) 메커니즘과 외부의 전파방해에 기인한다.

몇몇 硏究者들은 이와 같이 발생한 綴字誤謬單語들에 대하여 統計的 分析을 하여 유용한 결과들을 발표하였다. Damerau(1964)는 모든 綴字誤謬의 80 % 이상이 다음과 같은 誤謬라고 지적하였다.

① 인접한 두 글자가 뒤바뀐 誤謬(transposition):

例) "the"가 "hte"로 된 誤謬

② 가외의 한 글자가 追加된 誤謬(insertion):

例) "the"가 "thea"로 된 誤謬

③ 한 글자가 削除된 誤謬(deletion):

例) "the"가 "th"로 된 誤謬

④ 한 글자가 다른 글자로 바뀐 誤謬(substitution):

例) "the"가 "ahe"로 된 誤謬

Pollock(1983)은 CAS의 SPEEDCOP 프로젝트를 수행하면서 일곱개의 온라인 데이터베이스로부터 약 2,500 萬 單語를 抽出하고, 이 중에서 綴字誤謬인 單語만을 대상으로 다양한 分析을 하여 그 結果들을 발표한 바 있다. 이들의 分析結果에 따르면 2,500 萬 單語中에는 綴字誤謬인 單語가 5 萬 이상이나 있고, 그 綴字誤謬의 90～95 %가 단지 한 글자만 틀린 오류이며, 그 綴字誤謬單語 중에서 세번째 글자가 가장 많이 틀린 글자라고 한다.

Pollock의 調査에서는 한 글자가 削除되어 발생한 誤謬가 전체 오류의 30～40 %를 차지하여 가장 많았고, 두번째는 가외의 한 글자가 추가되어 발생한 誤謬(25～35 %), 세번째는 한 글자가 다른 글자로 바뀌어 발생한 誤謬(15～20 %),네번째는 인접한 두 글자가 뒤바뀌어 발생한 오류(10～15 %)였다. 이 이외에도 앞에서 언급한 오류가 두번 이상 반복하여 일어나는 誤謬도 4～9 % 정도 있었다.

더 재미있는 사실을 살펴보면 한 글자가 削除되어 발생한 誤謬에서는 같은 글자가 반복되어 나올 때 한 글자를 빠뜨린 경우("OMISSION"이 "OMISION"으로 된 誤謬)가 7.5 %를 차지하고, 이런 誤謬로 가장 많이 생략된 알파벳은 L,F,M, S이다. 이와는 반대로 한 글자가 추가되어 발생한 誤謬에서는 한 글자만 있어야 하는데 두번 연속 반복된 誤謬가 全體 誤謬의 45 %를 차지하여 이 오류의 절대부분을 점하고 있다. 그리고 한 글자가 다른 글자로 바뀌어 發生한 誤謬에서는 한 母音이 임의의 다른 母音으로, 한 子音이 다른 子音으로 바뀐 경우가 85 %를 차지하고, 대부분이 같은쪽 손에 의해 잘못 타이프된 오류이다. 또 인접한 두 글자가 뒤바뀌어 發生한 誤謬에서는 이와는 반대로 73 %가 母音과 子音, 또는 子音과

母音의 位置가 바뀐 誤謬이고, 84%가 오른손과 왼손, 또는 왼손과 오른손에 의해 잘못 타이프된 誤謬이다.

Ⅳ. 綴字校正시스템 構成

綴字校正시스템은 일반적으로 文書正規化 (document normalization), 綴字檢證 (spelling verification), 綴字校正 (spelling correction), 使用者 인터페이스(user interface)의 네가지 모듈로 構成된다. 여기에서 문서정규화는 綴字校正시스템의 실용성을 강화하는데 目的을 두고 있는 중요한 부분이다. 만약 文書가 正規化되지 않는다면 入力 텍스트를 구성하고 있는 單語들을 區分하는 간단한 것 조차 불가능하다.

綴字檢證과 綴字校正은 기능적으로 구별되어 있지만, 槪念的으로는 상당히 서로 관련되어 있다. 綴字檢證이 入力 토큰에 대하여 綴字誤謬인지 아닌지만을 검사한다. 그 뒤에 入力 토큰이 綴字誤謬이면 綴字校正에서 이를 오류의 種類에 따라 거기에 맞는 校正을 하고, 入力 토큰이 綴字誤謬가 아니면 綴字校正 루틴은 무시된다.

使用者 인터페이스는 綴字校正시스템을 이용하는 利用者와 직접적인 관계를 갖는 아주 중요한 부분이다. 綴字誤謬를 利用者에게 알리고 이를 편리하게, 또 效果的으로 고칠 수 있는 環境을 마련하는 것이 이 부분의 主要 機能이다.

1. 文書正規化 (Document Normalization)

綴字校正시스템은 정규화된 문서를 전제로 개발되어야 하며, 시스템의 실용적인 側面에서 매우 중요하다. 문서정규화는 다음과 같은 것들이 고려되어야 한다.

첫째는 大·小文字 (case), 폰트, 文字세트 등에 관한 文字暗號化 (character encoding)의 標準化이다. 이는 같은 種類의 單語가 시스템에 의해 다르게 처리되는 것을 막아 준다. 예를 들면, "HUMAN"과 "human"은 印刷上 차이는 있지만 綴字的으로 동일하며, 두 單語 모두 올바른 英語 單語이기 때문에 이 둘은 綴字校正시스템에 의해 같은 의미로 처리되어야 한다. 또한 24×24 폰트로 된 "人間"이라는 單語와 32×32 폰트로 된 "人間"이라는 單語도 같은 의미로 處理되어야 한다.

둘째는 文書編輯機(워드프로세서, 데이터入力·編輯시스템 등)의 制御文字들에 관한 것이다. 이들 文字들은 대개 印刷하면 나타나지 않고, 시스템에 의존적이고, 使用者의 취향에 따라 다양하게 나타난다. 따라서 이들 文字들은 일반적으로 綴字校正시스템에서 고려하지 않는다.

셋째는 숫자를 포함하는 토큰들에 관한 것이다. 숫자 오류를 校正하는 것은 주변상황 정보뿐만 아니라 文書作成者의 의도까지도 이해되어야 가능하기 때문에 일반적으로 이러한 토큰들은 綴字校正機에 의해 무시된다. 보통 이러한 토큰들은 利用者가 직접 校正하도록 남겨 두는 것이 보편화되어 있다.

넷째는 하이픈 (hyphen) 에 관한 것이다. 하이픈은 대개 複合語, 住民登錄番號에서와 같이 식별자 (delimiter) 로 쓰인다. 이 경우는 하이픈 양쪽의 單語를 별개의 토큰으로 간주하면 별 문제가 없으나 入力라인의 끝에서 單語分節을 위해 사용하는 하이픈은 다소 問題가 있다. 이때 分節된 單語는 다시 완전히 원상 복귀시키는 것이 불가능하다.

다섯째는 어포스트로피 (apostrophe) 에 관한 것이다. 어포스트로피는 일반적으로 綴字校正시스템에서 文字로 간주된다. 그들은 대부분 省略된 文字의 지시자(don't, I'd) 로 쓰이거나 소유격 (king's, John's) 을 나타내기 때문이다. 토큰의 시작과 끝에 있는 어포스트로피는 식별자로 간주되어야 하는데, 그 이유는 그들이 때때로 引用符號 ("the token 'ten'… ") 로 쓰이기 때문이다. 그러나 복수의 소유격(kids') 도 또한 單語의 끝에 어포스트로피를 갖는다.

여섯째는 接辭處理에 관한 것이다. 辭典에는 보통 단어의 원형만이 들어간다. 이는 辭典의 크기를 줄이고 辭典探索時間을 줄이기 위한 目的에 기인한다. 따라서 이와 같은 辭典을 효과적으로 이용하기 위해서는 入力 토큰에서 接辭 (s, er, ly, able 등)를 분리하고 그 單語의 원형을 찾는 것이 절대적으로 필요하다. 그러나 "ACCEPTIBLE", "ACCEPTER"와 같은 接辭의 오용에 대한 綴字誤謬를 발견하기위해서는 辭典에 접속 가능한 接辭에 관한 情報가 들어 있어야 한다.

2. 綴 字 檢 證

綴字를 檢證하는 方法은 크게 다음과 같은 세가지 方法이 알려져 있다.

(1) 辭典探索方法

이 接近方法의 주된 요소는 辭典이다. 辭典은 맞춤법에 맞는 단어들로 구성되어

있어야 한다. 이 方法은 먼저 각각의 入力 토큰을 사전에서 찾아보고, 그 토큰이 辭典에 있으면 올바른 單語로 간주하고, 사전에 없으면 綴字가 틀린 단어로 본다. 이 接近方法의 장점은 處理結果가 정확하다는 것이다. 그러나 어느 정도 쓸 수 있는 辭典(맞춤법에 맞는 單語의 수가 주어진 텍스트의 綴字檢證을 하기에 충분하지는 않지만 적당한 수로 구성된 사전)이 없으면 빠른 處理가 곤란하다. 일반적으로 올바른 綴字檢證을 하는데 필요한 사전의 단어 수는 비서들이 사용하는 單語水準인 2萬 내지 4萬 정도가 적당한 것으로 알려져 있다.

비록 광범위한 辭典이 텍스트의 많은 부분을 인지하더라도 수백만 單語를 처리하다 보면 일치되지 않는 單語들이 많이 생길 것이다. 그리고 자연 언어는 정확히 예측할 수 없기 때문에 사전이 정말로 크지 않다면 綴字가 틀리다고 出力한 單語中 진짜 誤謬인 것은 40% 이상 기대하기 어려울 것이다. 더구나 頻度가 낮은 單語를 辭典에 포함시키면 철자오류인 것을 올바른 綴字로 인지하게 할 수 있다. 예를 들면, "CLOSET"이라는 單語는 실내장식 데이터베이스에는 빈번히 출현하나 科學分野의 데이터베이스에서는 "CLOSE"의 최상급인 "CLOSEST"의 綴字誤謬일 가능성이 더 크다. 科學分野의 辭典에 "CLOSET"을 포함시킨다면 옳을 가능성 보다는 틀릴 가능성이 더 크다. 따라서 사전의 크기를 어느 정도 이상 증가시키는 것은 반대의 結果를 가져오기 쉽다.

(2) 統計的 方法

統計的 方法은 앞에서 기술한 辭典探索方法과는 달리 사전에 전혀 의존하지 않는다. 대신에 텍스트에서 綴字誤謬인 單語를 찾기 위하여 이 方法은 각 入力 토큰에 대하여 이상 계수(coefficient of peculiarity)를 계산한 다음 이상 계수가 基準値보다 큰 토큰들을 綴字誤謬對象으로 삼는다. 이와 같은 分析은 비록 정확성은 辭典探索方法보다 떨어지지만 辭典探索 루틴이 필요없기 때문에 處理速度가 빠르다는 장점이 있다.

이 方法은 다양한 텍스트 分析結果에 바탕을 두고 있다. 텍스트 分析의 가장 일반적인 形態로는 n-문자쌍(n-gram) 분석이라는 것이 있다. n-문자쌍이란 길이가 $n < L$인 문자열 $C_1C_2 \cdots C_L$에 대한 임의의 세그먼트를 말한다. 예를 들면, 문자열 "ABCD"의 2-문자쌍(digram)은 "AB", "BC", "CD"이고, 3-문자쌍(trigram)은 "ABC", "BCD"이다. n-문자쌍분석의 정확도는 入力 텍스트의 글자 구성이 사용중인 辭典의 글자구성을 제대로 반영하느냐에 비례한다. 이와 같은 n-문자쌍 분석에서는 3-문자쌍 분석이 가장 이상적인데, 그 이유는 2-문

자쌍은 너무 부정확하고, 3 이상의 n-문자쌍 分析은 統計的 計算이 너무 복잡하기 때문이다.

또한 綴字와 發音의 分析에 바탕을 둔 統計的 方法도 있다. 發音情報를 이용하면 "pneumonia"가 "numonia"로 된 것과 같은 誤謬를 찾는데 유용하다.

(3) 混 合 方 法

混合方法은 辭典探索方法과 統計的 方法의 장점만을 이용한 방법으로 현재 綴字校正시스템에서 가장 많이 쓰고 있다.

3. 綴 字 校 正

綴字校正은 크게 6가지 정도의 方法들이 고려될 수 있다. 가장 간단한 方法은 대화식 철자교정기에서 綴字誤謬가 발생할 때마다 여기에 적합한 올바른 單語를 利用者가 지적하여 시스템이 記憶하는 경우이다. 이 方法은 바람직하지 못한 면도 없지 않지만 반복되는 誤謬를 校正하는데 效果가 있다.

두번째로 생각해 볼 수 있는 것은 辭典內에 맞춤법에 맞는 單語와 일상적으로 틀리는 모든 單語를 두고 처리하는 方法이다. 이 方法은 사전내에 모든 가능한 誤謬를 넣어 둘 수만 있다면 아주 좋을 수 있지만 현실적으로 이는 불가능하다.

셋째는 Ⅲ章에서 살펴본 綴字誤謬에 대한 統計에 바탕을 두어 校正하는 方法이다. 이 方法을 개략적으로 살펴 보면 다음과 같다.

① 먼저 辭典에서 발견되지 않은 각각의 토큰에 대해 인접한 두 글자를 바꿔보거나, 임의의 한 글자를 빼보거나, 임의의 한 글자를 토큰내에 挿入하거나, 임의의 한 글자를 다른 글자로 바꿔보거나 하여 가능한 單語의 리스트를 構成한다. 이들이 入力 토큰에 대한 교정후보들이다.

② 만약 그 리스트가 정확히 한 개의 候補만을 갖는다면 그 單語가 원하는 單語인지를 利用者에게 물어 본다.

③ 만약 그 리스트가 몇개의 候補를 갖는다면 이것을 명시하여 利用者가 여기에 대한 조작(select, replace, replace and remember, accept, accept and remember, edit 등)을 하도록 한다.

넷째는 키보드의 자판 배열에 바탕을 두어 校正하는 方法이다. 이는 키인할 때 발생하는 綴字誤謬를 校正할 때 유용하다.

다섯째는 單語의 發音에 바탕을 두어 校正하는 것이다. 이것은 pf 대신에 f를

사용한다든가, qu 대신에 k를 사용하는 것과 같은 명백한 誤謬를 고치는데 도움을 준다.

여섯째는 토큰과 사전에 들어 있는 單語에 공통으로 들어 있는 문자열의 길이에 바탕을 두어 校正하는 것이다. 이것은 統計的인 分析에 바탕을 두고 있다. 확률이 높은 單語는 대개 바른 綴字로 선택될 수 있다. 그러나 이것은 매우 부정확하다.

4. 使用者 인터페이스

綴字校正시스템은 그 기능 못지 않게 사용하기 편해야 한다. 이를 위해서는 시스템에서 제시하는 것들을 利用者가 쉽게 알아 볼 수 있도록 해야 하고, 綴字誤謬인 單語를 利用者가 편리하게 고칠 수 있도록 하며, 利用者가 시스템에서 校正한 情報를 요구하면 이를 자세히 보여 줄 수 있어야 한다.

UNIX의 SPELL과 같은 초기의 시스템에서는 入力 텍스트를 한꺼번에 處理하여 綴字誤謬인 單語들을 순서적으로 보여주는 일괄 처리방식으로 운영되었다. 따라서 시스템이 제시한 綴字誤謬가 실제로 入力 텍스트에 있는가를 보기 위해서는 이를 종이에 옮겨 적은 후에 다시 文書編輯機를 이용해야만이 가능했다.

좀 더 진보된 시스템에서는 畵面 한쪽에는 誤謬리스트를 보여 주고, 다른 한쪽 畵面에서는 入力 텍스트를 校正할 수 있도록 하였다. 좀 더 최근의 시스템에서는 綴字誤謬를 畵面上에서 쉽게 알아 볼 수 있도록 그 부분을 반짝이게 한다든가, 메뉴를 제공하여 利用者가 원하는 기능을 쉽게 선택할 수 있게 한다든가, 利用者가 하나의 키 조작만으로 綴字誤謬單語를 올바르게 校正할 수 있도록 하는 등의 편리한 기능이 가능하다.

데이터베이스 製作에서 필요한 利用者 인터페이스 부분은 좀 더 특이한 면이 있을 수 있을 것이다. 이를 위해서는 실제로 데이터베이스를 제작하는 과정에서 綴字校正시스템이 이러한 使用者 인터페이스 부분을 어떻게 처리하였는가를 살펴보는 것이 중요하다. INSPEC에서의 경우를 살펴보자.

〈圖 2〉는 INSPEC에서 데이터베이스 製作過程인 校正作業(proofreading)을 위해 컴퓨터가 出力한 리스트이다. 이 리스트는 綴字校正시스템이 綴字가 틀린 부분에 대한 情報를 아래에 주고 있다. 또 정정된 리스트를 利用者가 쉽게 온라인으로 校正할 수 있도록 하기 위해 INSPEC에서는 온라인 畵面도 〈圖 2〉와 같은 形態로 構成하여 제공하고 있다.

〈圖 2〉 INSPEC의 데이터베이스 校正用 리스트

NEW RECORD

```
0     .   1644-86006-A001-R
1     1   A001-R
2     1   341-8
4     1   a4260Bq a4280Lc a4255Pq b4320Jh b4130+d
5     .   1644-86006 .
6     1   Threshold current analysis of InGaAsP-InP ‖ridge-waveguide
      2   lasers‖
7     1   Amann, A.-C.
      2   Siegmuller, B.
      3   (Siemens AG Res. Labs., Munchen, Germany)
8     1   The marked depression of the refractive index by ‖injected
      2   carriers‖ in the «InGaAsP-InP ant/material system
      3   essentially influences the performances of ‖laterally
      4   innomogeneously pumped laser diodes‖ such as the usual
      5   ridge-waveguide-structures. This is because, together with
      6   the built-in index guiding an ‖M-shaped waveguide‖ is
      7   formed that may become leaky even for the fundamental
      8   waveguide mode. The characteristics of these waveguides are
      9   calculated and compared with previous results. Hence, the
     10   waveguide structure is studied for a wide range of device
     11   parameters showing that ‖optical losses‖ can be kept
     12   within acceptable limits by applying effective ‖index
     13   steps‖ in excess of 0.02. Furthermore, the ‖mode gain‖
     14   is calculated for the corresponding ‖carrier profiles‖
     15   and the threshold current is estimated. According to the
     16   experiment, the results show that the threshold currents around
     17   20 (Ma) can be achieved with appropriate waveguide structures
     18   at wavelengths of 1.3 μm and 1.55 μm)
10    .   20
19    1   0267-3932/86/$2.00+0.00
65    1   refractive index depression; III-V semiconductors
      2   threshold current analysis; leaky waveguide;
68    .   T
71    1   indium-compounds
      2   *gallium compounds
      3   gallium arsenide
      4   SJL
      5   optical waveguides
      6   laser transitions IZZ at 1.3 and 1.55 μm;
73    .   C InGaAsP-InP laser ridge waveguide, threshold current
      2   anal.
140   .   curr 20 mA.
      2   wave 1.3 μm:
      3   wave 1.55 μm:
92    .   INPUT 23-Dec-1986 06:17:18
```

Corrections marked

Computer generated errors and warning messages

```
??? WARNING IN FIELD 8 Field does not begin with upper case
??? WARNING IN FIELD 8 Spelling error detected
*** ERROR IN FIELD 8 missing right Chemical Delimiter: line 2
??? WARNING IN FIELD 71 Spelling error detected
*** ERROR IN FIELD 71 Field 71 line 2 not on Thes. - 1 errors
```

Ⅴ. 結　　論

지금까지 綴字校正시스템에 대하여 데이터베이스 製作과 情報檢索側面에서 살펴보았다. 이를 요약하면 다음과 같다.

첫째, Ⅱ章에서는 書誌 데이터베이스내의 綴字誤謬의 頻度와 이들 오류가 우리에게 미치는 영향에 대하여 세가지 관점(데이터베이스 供給者觀點, 情報서비스機關觀點, 그리고 利用者觀點)에서 論하였다.

둘째, Ⅲ章에서는 Damerau(1964)와 Pollock(1983)이 조사한 內容을 중심으로 이러한 綴字誤謬의 類型에 대하여 알아 보았다.

셋째, Ⅳ章에서는 일반적인 綴字校正시스템의 構成에 대하여 文書正規化, 綴字檢證, 綴字校正, 使用者 인터페이스의 네 부분으로 나누어 이들 각각에 대하여 기술하였다.

本稿는 이러한 考察을 통하여 綴字校正시스템이 데이터베이스 製作에 상당한 도움을 주고, 正確한 情報를 利用者에게 전달하는데 필요하며, 시스템 維持 및 管理費用을 절감시키며, 利用者의 계속적인 信賴를 받는데 견인차 역할을 수행함을 확인할 수 있었다. 또한 우리의 20여 년에 걸친 情報서비스 경험 등을 살린다면 韓國語 綴字校正시스템을 적은 노력으로 구축하여 앞으로 있을 우리말 데이터베이스 製作에 效果的으로 적용시킬 수 있을 것이라고 확신할 수 있었다.

〈參 考 文 獻〉

1. Angell, R.C., Freund, G.E. and Willett, P., "Automatic Spelling Correction Using a Trigram Similarity Measure," *Information Processing & Management*, vol.19, no.4, 1983, pp.255 ~ 261.

2. Berghel, H.L., "A Logical Framework for the Correction of Spelling Errors in Electronic Documents," *Information Processing & Management*, vol. 23, no. 5, 1987, pp.477 ~ 494.

3. Bourne, C.P., "Frequency and Impact to Spelling Errors in Bibligraphic Data Bases," *Information Processing & Management*, vol.13, 1977, pp.1 ~ 12.

4. Cornew, R.W., "A Statistical Method of Spelling Correction," *Information and Control*, vol.12, 1968, pp.79 ~ 93.

5. Damerau, F.J., "A Technique for Computer Detection and Correction of Spelling Errors," *CACM*, vol.7, no.3, 1964, pp.171 ~ 176.

6. Heidorn, G.E., Jensen, K., Miller, L.A., Byrd, R.J., and Chodorow, M.S., "The EPISTLE Text-Critiquing System," *IBM SYST J*, vol.21, no.3, 1982, pp.305 ~ 326.

7. INSPEC, "Introducing INSPEC's Production & Editorial Systems."

8. Muth, F.E., and Tharp, A.L., "Correcting Human Error in Alphanumeric Terminal Input," *Information Processing & Management*, vol.13, 1977, pp.329~337.

9. Peterson, J.L., "Computer Programs for Detecting and Correcting Spelling Errors," *CACM*, vol.23, no.12, 1980, pp.676 ~ 687.

10. Peterson, J.L., "A Note on Undetected Typing Errors," *CACM*, vol.29, no.7, 1986, pp.633 ~ 637.

11. Pollock, J.J., "Spelling Error Detection and Correction by Computer : Some Notes and Bibliography," *Journal of Documentation*, vol.38, no.4, 1982, pp.282 ~ 291.

12. Pollock, J.J., and Zamora, A., "Collection and Characterization of Spelling Errors in Scientific and Scholarly Text," *Journal of American Society for Information Science*, vol.34, no.1, 1983, pp.51 ~ 58.

13. Pollock, J.J., and Zamora, A., "System Design for Detection and Correction of Spelling Errors in Scientific and Scholarly Text," *Journal of American Society for Information Science*, vol.35, no.2, 1984, pp.104 ~ 109.

14. Pollock, J.J., and Zamora, A., "Automatic Spelling Correction in Scientific and Scholarly Text," *CACM*, vol.27, no.4, 1984, pp.358 ~ 368.

15. Srihari, S.N., "Computer Text Recognition and Error Correction," *IEEE Computer Society Press*, 1984.

16. Yonnakoudakis, E.J., and Fawthrop, D., "The Rules of Spelling Errors," *Information Processing & Management*, vol.19, no.2, 1983, pp.87 ~ 99.

17. Yannakoudakis, E.J., and Fawthrop, D., "An Intelligent Spelling Error Corrector," *Information Processing & Management*, vol.19, no.2, 1983, pp.101~108.

18 . Zamora, A., " Automatic Detection and Correction of Spelling Errors in a Large Data Base," *Journal of American Society for Information Science*, Jan. 1980, pp.51 ~ 57.

19 .강재우, 「接續情報를 利用한 한글 綴字 및 띄어쓰기 검사기의 設計 및 具現」, 韓國科學技術院 碩士學位論文, 1990.

20 . 박종만, 「效率的인 韓國語 형태소 분석기 및 綴字檢査 校正機의 具現」, 서울大學校 碩士學位論文, 1990.