한국어 Lexicon에 의존한 문자 인식의 후처리

임 한 규

한국아이비엠(주) 소프트웨어연구소

A Postprocessing of Character Recognition Based on Korean Lexicon

Hankyu Lim

Korea Software Development Institute
IBM Korea

요 약

본 논문에서는 문자 인식이 끝난 한국어 원문에 대해 한국어 Lexicon에 기반을 둔 후처리의 구현을 보여주는 것을 목적으로 한다. 빈번하게 오인식되는 음절에 대해 이의 옳은 음절을 대응시킨 테이블을 만들어 놓고, 오인식이라고 정의된 음절이 출현했을 때는 이를 원래의 옳은 음절로 대체시킨 어절과 오인식된 음절이 포함된 어절에 대해 한국어 형태소 분석을 행함으로써, 올바른 형태소가 분석될 경우, 이를 옳은 음절로 간주한다. 실험결과 약 90%에서 95%에 달하는 인식율이 이 후처리 방법에 의해서 95%에서 99%로 높아졌다.

Ⅰ. 서론

문자 인식 시스템에 관한 활발한 연구 및 개발의 덕택으로 요즈음 상용화된 OCR 제품도 국내에서 출현하였고 많은 연구 단체나 대학교의 연구실 및 유수 회사 등도 상당한 기술을 축적하고 있다. 그러나 이들에 의해 연구되거나 개발되어 발표되거나 전시된 Prototype의 면면을 깊이 들여다보면 상당한 제약을 갖고 있는 실정이다. 폰트에 민감한가 하면, 입력되는 문서의 상태에 좌우되기도 하고, 인식방법의 차이에 따라 인식 속도

나 인식 성공률이 크게 달라지기도 한다. 이들 Prototype에 의해 인식된 결과는 여러 가지 이유로 인해 원문과는 다르게 인식되어, 사람에 의해서건, 맞춤법 검사 및 교정의 기능을 갖는 Wordprocessor에 의해서건 교정하는 절차가 필요하게 된다. 특히 전표류가 아닌 일반 교과서와 같은 다량의 원문을 입력하게 되면, OCR엔진의 인식속도보다는 인식된 원문의 확인, 수정 작업의 시간이 전체의 효율을 좌우하기도 한다. 이와 같이 오인식된 원문의 검사 및 교정을 위해서 문맥적인 지식을 활용하여 후처리 단계에서 이를 극복해 보려는 것이 이 논문의 목적이다.

후처리의 알고리즘의 유형으로는 상향식, 하향식 및 복합적 방식이 있다[2]. 상향식 방법은 확률적 방식에 근거한 것으로서, Viterbi 알고리즘, 수정된 Viterbi알고리즘이 있으며, 구조적 표현에 기초한 하향식 방식에는 Dictionary Lookup 알고리즘과 String Matching 알고리즘 및 Binary N-gram 알고리즘이 있다. 복합적 방식으로는 Dictionary Viterbi 방식과 Predictor-Corrector 알고리즘이 있다. 이러한 후처리 방식 중의 하나로서, 본 논문에서는 한국어 Lexicon에 기반을 둔 한국어 형태소 분석기에 의해, 오인식된 음절의 교정을 행하는 방법을 제시하려고 한다.

II. 문자 인식 엔진 및 오인식 사례

본 논문에서 언급되는 문자인식 엔진의 모델은 한국아이비엠(주)와 경북대학교간에 진행되고 있는 공동연구의 결과로서 만들어진 Prototype이다[7]. 인식 엔진은 스캐너를 통해 들어 온 문서의 이미지를 텍스트 부분만 발췌하여 각 문자 단위로 나눈 뒤, 각 문자는 다시 한글의 여섯 가지 문자 형태로 분류되어 자소 단위로 인식되고 이들이 결합되어 하나의 음절이 만들어진다. 인식된 음절은 여러 가지 요인으로 인해 원문과는 다른 형태로 인식되는데, 오인식된 음절의 예를 보면 아래 표 1과 같다.

원래 음절	오인식 결과
립	럽
녁	넉
렸	럿
행	뱅
돕	돔
업	엄
젔	겼
넝	넝
답	담
럼	렁
문	뭄
권	컨
몸	뭉
십	심

표 1. 오인식 음절의 예

오인식된 음절은 동일한 인식 엔진이라 할지라도, 특정 폰트에 의해 다소 좌우되기도 한다는 것은 부인할 수 없는 사실이다. 특히 종성의 경우, ㅂ이 ㅁ으로 잘못 인식된다거나, ㅁ이 ㅇ으로 잘못 인식되는 것과 같은 다소 구조적인 문제도 발생한다. 여기서는 입력원문의 종이의 질이나 인쇄의 질 및 폰트 등에 상관없이, 인식 과정에서 오류를 유발시키는 어떠한 원인 등을 규명하여 처리 방법을 제시하려는 것이 아니라, 오직 인식된 결과만을 가지고 이를 문맥적 지식에 의해 해결해 나가려고 한다.

Ⅲ. 한국어 형태소 분석기

여기서 언급되는 한국어 형태소 분석기는 한국아이비엠(주)과 서울대학교와의 공동 연구 결과로서 만들어진 프로그램으로서, 한국어 형태소 분석 알고리즘과 한국어 Lexicon은 이미 '가나다 Wordprocessor'(상표명)에 장착되어 맞춤법 검사 기능을 발휘하고 있다[1]. 한국어의 형태소 분석은 단어를 이루고 있는 형태소를 분리하고, 형태론적인 변형이 일어난 형태소의 원형을 복원하고, 사전과 단어 사이의 통합 관계에 의해 옳은 분석 후보를 선택하는 과정으로 이루어진다. 이의 구체적인 기능으로서는 형태소 분리 기능, 형태소 간의 결합관계 검사기능, 준말이나 복합어 처리 기능 등이 있다. 한국어의 형태소 분석은 크게 두 단계로 나누어 지는데, 하위 단계인 음절 분석에서는 음절 정보를 이용하여, 단어의 분석 순서 예측 및 문법 형태소의 분석을 행한다. 상위 단계인 복수어 단위 분석에서는 복수어 단위 정보에 의해 분석 결과의 우선 순위를 결정하고 모호성 문제를 해결한다. 가령 '가는'을 형태소 분석을 행하면 다음의 네 가지로 분석된다.

(NOUN '가') + (JOSA '는')
(VERB '가늘') + (EOMI '은')
(VERB '가') + (EOMI '는')
(VERB '갈') + (EOMI '는')

본 논문에서는 단지 Lexicon에 기반을 둔 어휘 정보만으로 그 범위를 한정하게 되므로, 다소 복합적인 구문 정보나 의미 정보를 다루지는 않는다. 따라서 위와 같이 모호성을 갖는 형태로 분석이 되더라도, 본 논문에서 별 의미를 부여하지는 않는다. 여기서 사용하고 있는 Lexicon은 어휘 Lexicon과 문법 Lexicon으로 나누어 지는데, 어휘 Lexicon은 어휘 형태소에 대한 품사 정보, 불규칙 정보 및 타 형태소와의 결합성을 나타내는 자질들로 이루어져 있으며, 약 10만 개가 넘는다. 문법 Lexicon은 조사와 선어말 어미, 어말 어미, 접사에 대한 정보를 수록하고 있는 데, 복합 조사를 포함하여 조사는 약 500개로 구성되어 있고, 두 결합 이상을 포함한 선어말 어미는 40개로 구성되어 있으며, 조사와 결합한 것까지 고려한 어미의 수는 약 760개에 달한다. 어휘 Lexicon은 Trie구조로 되어

있으며, 문법 Lexicon은 크기가 작고 빈번한 Access로 인해 주기억 장치에 적재되어 있는데 이진 탐색을 사용하고 있다. 한국어 형태소 분석 프로그램의 분석 성공률은 약 99.36%이며 분석시 Lexicon의 참조 횟수는 2.13회이다. 처리 속도는 IBM PC 386에서 약 29 msec이다. 아래의 표 2와 표 3은 어휘 Lexicon과 문법 Lexicon의 구성 예를 보여준다.

```
-------------------------
ㄱㄴ순N
ㄱ 자자N
ㄱ 자집N
가N:VI RG:VX
가가N
가가례N
......
......
-------------------------
```

표 2. 어휘 Lexicon

```
--------------------------
char *Josa[] = {
            "가",
            "같이",
            ....
            "하곤"
};
char *Eomi[] = {
            "거나",
            "거늘",
            ....
            "질"
};
--------------------------
```

표 3. 문법 Lexicon

Ⅳ. 실험결과

일단 인식이 끝난 한국어에서 위의 표 1에 정의된 오인식 음절이 나타나면 이 음절을 포함한 어절의 형태소 분석을 행한다. 또한 오인식 되었다고 정의된 음절을 원래의 음절이라고 정의된 음절로 대체한 뒤 이 어절의 형태소 분석을 행한다. 전자에서 형태소 분

석이 실패하고 후자에서 성공한 경우, 후자의 음절을 전자의 음절로 대체하여 확정한다. 전자에서 성공하고 후자에서 실패한 경우는 전자의 음절을 그대로 확정한다. 그러나 전, 후자 모두에서 성공하거나 실패한 경우는 처리시 다소 고려를 해봐야 한다. 우선 양자 모두 실패한 경우는 여러가지 요인이 있을 수 있는데, 그 단어가 고유명사이거나, 형태소 분석 프로그램에서 일부 처리되지 않는 복합명사이거나, 사투리 등의 경우와 같이 어휘 Lexicon에 적재되어 있지 않은 경우 등이 있으며, 형태소 분석 프로그램에서 처리될 수 없는 0.64%에 기인할 수도 있다. 이 경우의 해결책으로는 누락되어 있는 표제어를 Lexicon에 적재하는 등의 방법을 생각할 수 있다. 양자 모두가 형태소 분석에 성공한 경우는, 음절의 위치별 빈도수를 고려하여 빈도수가 높은 음절을 맞는 음절로 간주할 수 있으며[3], 더 깊은 연구를 위해서는 아직은 초보단계에 불과한 구문 정보나 의미정보의 처리까지 고려해야 한다. 아래의 표 4는 틀린 어절과 맞는 어절의 형태소 분석 결과를 보여주고 있다.

틀린 어절 - 형태소분석 결과	맞는 어절 - 형태소분석 결과
버럽니다 - 분석에 실패	버립니다 - 버리(V)+습니다(F)
저넉 - 분석에 실패	저녁 - 저녁(N)
버럿다 - 분석에 실패	버렸다 - 버리(V)+었(PF)+다(F)
여뱅을 - 분석에 실패	여행을 - 여행(N)+을(P)
둠고 - 분석에 실패	돕고 - 돕(V) +고(F)
거럿다 - 거르(V)+었(PF)+다(F)	거렸다 - 분석에 실패
실엄가의 - 분석에 실패	실업가의 - 실업가(N)+의(P)
꺼겼군요 - 분석에 실패	꺼졌군요 - 분석에 실패
안넝 - 분석에 실패	안녕 - 안녕(EXCL or N)
대담이 - 대담(N)+이(P)	대답이 - 대답(N)+이(P)
것처렁 - 분석에 실패	것처럼 - 것(N)+처럼(P)
때뭄에 - 분석에 실패	때문에 - 때문(N)+에(P)

컨위는 - 분석에 실패	권위는 - 권위(N)+는(P)
몽을 - 몽(N)+을(P)	몸을 - 몸(N)+을(P)
해주심사 - 분석에 실패	해주십사 - 분석에 실패
하풍을 - 하풍(N)+을(P)	하품을 - 하품(N)+을(P)

* * 기호 설명

V	: 동사
N	: 명사
P	: 조사
EXCL	: 감탄사
F	: 어말어미
PF	: 선어말어미

표 4. 틀리고 맞는 어절들의 형태소 분석 예

위와 같은 방식으로 처리함으로써 오인식된 음절의 80%를 상회하는 음절이 올바른 음절로 인식될 수 있었다. 현재 문자 인식 엔진의 인식 성공율은 문서의 종류에 따라 다소 차이를 보이고 있으나 인식율이 90%인 경우, 후처리를 거치면 약 95 %에서 98 %까지 인식율이 높아지며, 인식율이 95%인 경우, 후처리를 거치면 약 97 %에서 99 %까지 인식율이 높아진다. 여기서 발생하는 문제의 하나는,올바르게 인식되었는데, 표 1에 등재된 대로 오인식 음절로 간주되어, 형태소 분석을 행하여 이에 성공한 경우, 맞는 음절이 틀리는 음절로 해석될 수 있는 가능성이 있다는 것이다.

Ⅴ. 앞으로의 연구 방향

지금까지 살펴본 오인식의 예는 일부 제한된 폰트에 대해서만 실험한 결과로서, 다중 폰트에 대해서 실험을 확대하면 오인식된 문자의 수도 늘어날 것이며, 일부의 경우는 모순되는 경우가 예측되기도 한다. 이럴 경우, 테이블을 무한정 늘려서 해결될 문제는 아니므로 이의 처리 방법을 연구해야할 것이며, 모호성이 발생할 경우, 이를 구문 분석이나 의미 분석에 의해 해결할 수 있는 방법이 연구되어야 할 것이다. 또한 인식 단계에서 오인식 가능성을 탐지할 수 있으면 옳게 인식된 음절을 틀린 음절로 해석하는 등의 오류

를 줄일 수 있게 될 것이다. 또한 종성이 다르게 인식되는 문제는, 해당되는 종성들(ㅂ과 ㅁ, ㅁ과 ㅇ)이 들어가는 음절 모두를 대상으로 할 지는 좀 더 많은 실험을 통해서 밝혀야 할 과제이다. OCR의 적용 대상으로 현재도 쓰이고 있고, 앞으로도 확대될 좁고 특정 분야를 대상으로 한 Application을 고려할 때(예, 금융기관의 가입 및 청약서, 고객 대장, 일반 회사의 입사원서, 인사기록표), 이들 양식의 각각의 항들은 상당한 제약 조건을 가지고 있다. 이름, 주소, 우편번호, 생년월일, 주민등록번호 및 전화번호 등은 각 자리마다 고유한 특성 들을 가지고 있고 상호 연관을 가지기도 한다. 이들의 인식 또는 후처리에 제약을 가할 수 있는 조건의 기술이나, 주소 Lexicon, 성 및 이름의 빈도 순등을 활용하면 OCR제품의 상용화에 큰 기여를 하리라고 보여진다[4].

참 고 문 헌

[1] 강 승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 박사학위 논문, 1993

[2] 민 병우, 이 성환, 김 홍기, "문자 인식을 위한 후처리 기법의 사례 연구", 제1회 문자인식 워크샵 발표논문집, PP. 91-103, 1993

[3] 권 혁철, "글자 인식을 위한 한국어 정보의 이용", 단기강좌:문자인식기술, PP. 213-244, 1992

[4] 이 성환, 김 은순, "한글 주소의 오인식 수정을 위한 효율적인 후처리 알고리즘", 제4회 한글 및 한국어정보처리 학술대회, pp. 555-566, 1992

[5] Jonathan J. Hull, "Handwriting Recognition Research for Postal Automation", 제1회 문자인식 워크샵 초청 강연 및 튜토리얼 자료집, PP. 1-23, 1993

[6] Hiromichi Fujisawa, "Character Recognition Technologies and their Applications in Japan - The current status and the future", 제1회 문자인식 워크샵 초청 강연 및 튜토리얼 자료집, PP. 25-46, 1993

[7] Hangjoon Kim, Hankyu Lim, "The Study for Automatic Recognition of Hangul Document", PP. 5-73, 1992

[8] Nobuyasu ITOH, Hiroshi MARUYAMA, "A Method of Detecting and Correcting Errors in the Results of Japanese OCR", Human Interface 38-5, 1991

[9] H. Takahashi, N. Itoh, T. Amano, A. Yamashita, "A Spelling Correction Method and its Application to an OCR System", Pattern Recognition, Vol. 23, No. 3/4, PP. 363-377, 1990

[10] R. G. Casey, "Text OCT by Solving a Cryptogram", IEEE, PP. 349-351, 1986