제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

유한상태변환기만을 이용한 한국어 형태소 분석 및 품사 태깅

박원병*, 김재훈*
*한국해양대학교 컴퓨터공학과
e-mail:thewonner@empal.com, jhoon@mail.hhu.ac.kr

Korean Morphological Analyzer and POS Tagger Just Using Finite-State Transducers

Won-Byeong Park*, Jae-Hoon Kim*
*Department of Computer Engineering, Korea Maritime University

요 약

이 논문은 유한상태변환기만을 이용하여 한국어 형태소 분석 및 품사 태깅 시스템을 제안한다. 기존의 한국어 형태소 분석 시스템들은 규칙기반 형태소 분석기가 주를 이루고 한국어 품사 태깅 시스템은 은닉마르코프 모델 기반 품사 태깅이 주를 이루었다. 한국어 형태소 분석의 경우 유한상태변환기를 이용한 경우도 있었으나, 이 방법은 변환기를 작성하기 위한 규칙을 수작업으로 구축해야 하며, 그 규칙에 따라서 사전이 작성되어야 한다. 이 논문에서는 품사 태깅 말뭉치를 이용해서 유한상태변환기에서 필요한 모든 변환 규칙을 자동으로 추출한다. 이런 방법으로 네 종류의 변환기, 즉, 자소분리변환기, 단어분리변환기, 단어형성변환기, 품사결정변환기를 자동으로 구축한다. 구축된 변환기들은 결합연산(composition operation)을 이용하여 하나의 유한상태변환기를 구성하여 한국어 형태소 분석과 동시에 한국어 품사 태깅을 수행한다. 이 방법은 하나의 유한상태변환기만을 이용하기 때문에 복잡도는 선형시간(linear complexity)을 가지면, 형태소 분석기와 품사 태깅 시스템을 매우 짧은 시간 내에 개발할 수 있었다.

1. 서론

컴퓨터로 입력된 한글에 대한 의미 분석은 많은 연구들에 의하여 점진적인 발전을 보여 왔다. 하지만 의미 분석의 가장 기본이 되는 형태소 분석에 대한 연구 및 개발은 이에 미치지 못하고 있다. 비록 만족할 만한 성능을 보이기는 하지만, 기존의 형태소 분석 시스템은 미리 구축되어진 사전 및 규칙들을 가지고 입력으로 들어온 문장을 비교하여 분석하여 축약현상, 불규칙 현상의 처리 등을 분석하는 방식을 사용한다. 이 논문은 유한상태변환기만을 이용하여 한국어 형태소 분석 및 품사 태깅 시스템을 제안한다. 기존의 한국어 형태소 분석 시스템들은 규칙기반 형태소 분석기가 주를 이루고 한국어 품사 태깅 시스템은 은닉마르코프 모델 기반 품사 태깅이 주를 이루었다. 한국어 형태소 분석의 경우 유한상태변환기를 이용한 경우도 있었으나(이성진 1992), 이 방법은 변환기를 작성하기 위한 규칙을 수작업으로 구축해야 하며, 그 규칙에 따라서 사전이 작성되어야 한다. 이 논문에서는 품사 태깅 말뭉치를 이용해서 유한상태변환기에서 필요한 모든 변환 규칙을 자동으로 추출한다. 이런 방법으로 네 종류의 변환기, 즉, 자소분리변환기, 단어분리변환기, 단어형성변환기, 품사결정변환기를 자동으로 구축한다. 구축된 변환기들은 결합연산(composition operation)을 이용하여 하나의 유한상태변환기를 구성하여 한국어 형태소 분석과 동시에 한국어 품사 태깅을 수행한다. 이 방법은 하나의 유한상태변환기만을 이용하기 때문에 복잡도는 선형시간(linear complexity)를 가지면, 형태소 분석기와 품사 태깅 시스템을 매우 짧은 시간 내에 개발할 수 있었다.

이 논문의 구성은 다음과 같다. 2장에서는 기존의 형태소 분석과 품사 태깅 방법을 간단히 살펴본다. 3장에서는 한국어 형태소 분석과 품사 태깅을 위한 유한상태변환기에 대해서 기술한다. 4장에서는 사용한 말뭉치와 말뭉치로부터 형태소 정보를 축출하는 과정을 말할 것이며, 마지막으로 5장에서 결론을 맺고 향후 연구에 대해서 기술한다.

2 관련 연구 및 목표

2.1 한국어 형태소 분석기

형태소 분석기의 목표는 입력된 문장을 형태소로 분리하는 것이며, 그들 간의 결합 관계를 밝히는 것이다. 입력 문장이 정형화되어 있지 않으므로 접속 관계 처리, 불규칙 및 음운 현상 처리, 복합어 처리,

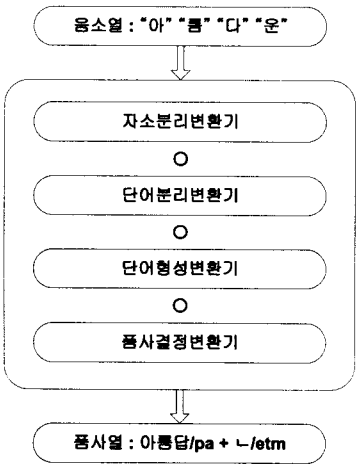
미등록어 처리 과정 등을 거친다. 이러한 처리를 위하여 기존의 형태소 분석기(강승식 2002; 임희석 외, 1995)들은 경험규칙, 사전, 규칙 명세들을 이용한다. 또한 형태소 결합 테이블, 불규칙 변형 테이블, 준말 변형 테이블 등을 사용하여 주어진 입력과 순차적으로 비교하면서 형태소를 분석한다.

2.2 한국어 품사 태깅

일반적으로 한국어 품사 태깅 시스템은 크게 형태소 분석기와 품사 태거로 구성되어 있다. 형태소 분석기는 주어진 어절에 대해서 가능한 형태소 해석 결과를 출력하는 시스템이다. 즉 형태소 분석의 중의성을 생성하는 시스템이다. 반면에 품사 태거는 형태소 분석기에서 생성된 형태소 분석의 중의성을 해소하는 시스템이다. 이와 같이 두 시스템의 역할이 분명히 구별되어 있으나, 두 시스템은 아주 밀접한 관계를 가지고 있다. 예를 들면, 형태소 분석의 중의성이 작으면 작을수록 품사 태거의 정확률은 높아진다. 또한 형태소 분석기의 성능(과분석이나 오분석 정도, 미등로어 추정 능력 등)은 품사 태거의 성능에 커다란 영향을 준다. 대부분의 한국어 품사 태깅 시스템은 은닉마르코프모델(hidden Markov model)에 기반하며 (김재훈, 1998; 임희석, 1997, 신상현 외, 1997), 대부분 96%이상의 좋은 성능을 보이고 있다.

3. 한국어 형태소 분석과 품사 태깅을 위한 유한상태변환기

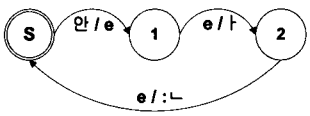
한국어 형태소 분석 및 품사 태깅을 위하여 네 종류의 유한상태변환기, 즉, 자소분리변환기, 단어분리변환기, 단어형성변환기, 품사결정변환기를 이용하며 전체 유한상태변환기의 구성은 (그림 1)과 같으며 이들을 하나의 유한상태변환기를 사용하기 위해서 Camel(Knight and Al-Onaizan, 1999)을 이용한다. 이하의 절에서 각 변환기에 대해서 구체적으로 설명할 것이다.



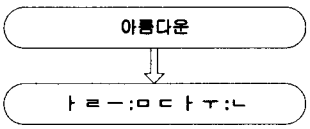
(그림 1) 제안된 유한상태변환기의 구성

3.1 자소분리변환기

분석에 소모되는 시간을 줄이기 위하여 기본적인 분석은 자소단위로 이루어진다. 이를 위하여 입력된 문자열은 자소단위로 분리되어 지는데, 이 과정에서 음가가 없는 'ㅇ'은 제거가 되고, 종성은 초성과의 구별을 위하여 앞에 ':'이 붙여진다. (그림 2)는 자소 분리를 위한 오토마타를 보여주며, (그림 3)은 자소가 분리된 결과를 보여준다.



(그림 2) 자소분리변환기



(그림 3) 자소분리의 결과

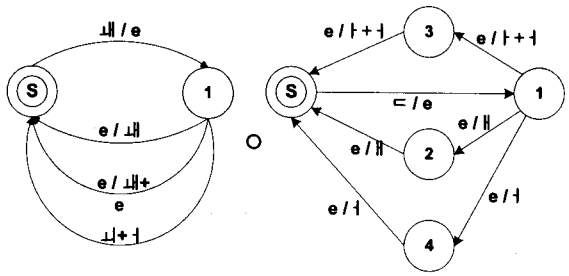
3.2 단어분리변환기

올바른 형태소 분석을 진행하기 위하여 여러 가지 불규칙 변환을 처리해야 한다. 기존의 형태소 분석 시스템은 불규칙 변환 사전 또는 규칙들을 순차적으로 비교하여 변환되는 경우를 찾아내었다. 하지만, 자소단위로 분리된 문자열들은 미리 구축되어진 불규칙 변환 오토마타를 순회하면서 결과물들을 얻어낸다. 이를 위하여 필요한 정보는 하나의 자소가 변환할 수 있는 다른 자소들의 목록과 현재의 자소가 다른 자소들과 가지는 연관관계이다. 현재의 자소가 앞의 자소와 가지는 연 관계를 이용한 bi-gram 방법을 사용하였다. 불규칙 변환을 적용시키기 위하여 두 가지의 단계를 거치게 되는데, 첫 번째 단계로 하나의 자소가 변화 가능한 자소들의 목록을 얻는 것이다. 두 번째로는 얻어진 후보 목록들 중 연관관계치가 높은 후보들의 우선순위를 높게 변화시켜주는 과정이다. 두 번째 과정을 거치게 되면서 불필요한 자소 변환 후보들을 최소화시키고, 자소간의 결합관계에 적합한 자소 변환 후보들을 얻을 수가 있다.

(그림4)의 왼쪽 변환기는 첫 번째 단계를 위한 자소 변환 오토마타를 표현한 것이다. 하나의 자소는 다른 다소로 변환 가능한 목록과 가중치를 가지는데, 첫 번째 단계에서의 가중치 차이는 그다지 크지 않게 설정하였다. 이는 자소의 변환 가능성도 중요하지만, 다른 자소와의 관계에 따른 가중치의 차이가 더 중요하다고 판단되었기 때문이다. (그림 4)의 오른쪽 변환기는 두 번째 단계를 위한 미리 구축되어진 오토마타의 도식적 구조를 보여 준다. 이는 현재 자소가 'ㅏ+ㅓ'이고, 앞의 자소가 'ㄷ'인 경우를 보여주는 것으로써 이전 자소 'ㄷ'에 의하여 최초 단계로 이동한 뒤에 현재 자소인 'ㅏ+ㅓ'에 의하여 시작 상태로 되돌아간다. 이때 현재 자소를 출력하여 주고, 각각의 간선들은 가중치를 가지고 있으므로, 이 오토마타를 거친

뒤에는 우선순위가 낮은 후보들도 높아질 수가 있다. (그림 5)는 순회 결과를 보여 준다.

(그림 4)에서 한 가지 유심히 살펴봐야 할 것은 두 개의 오토마타의 사이에 있는 'o'라는 기호이다. 이는 두 유한상태변환기의 Composition을 의미한다.



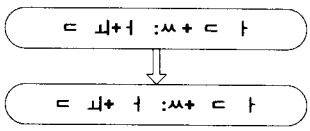
(그림 4) 단어분리변환기

"ㄷ" "ㅐ" ":ㅆ" "ㄷ" "ㅏ" "."
"ㄷ" "ㅚ+ㅓ" ":ㅆ+" "ㄷ" "ㅏ+" "." *e* 0.0504072
"ㄷ" "ㅚ+ㅓ+" ":ㅆ+" "ㄷ" "ㅏ:ㅎ+" *e* "." *e* 0.0499678
"ㄷ" "ㅚ+ㅓ" ":ㅆ+" "ㄷ" "ㅏ:ㅎ+" *e* "." *e* 0.0499678
"ㄷ" "ㅚ+ㅓ" ":ㅆ+" "ㄷ" "ㅏㄱ" *e* "." *e* *e* 0.0499678
"ㄷ" "ㅚ+ㅓ+" ":ㅆ+" "ㄷ" "ㅏㄱ" *e* "." *e* 0.0499678
"ㄷ" "ㅚ+ㅓ+" ":ㅆ+" "ㄷ" "ㅏ+" "." *e* 0.0494072
"ㄷ" "ㅚ+ㅓ" ":ㅆ+" "ㄷ" "ㅏ+" ".+" *e* 0.0229329
"ㄷ" "ㅚ+ㅓ+" ":ㅆ+" "ㄷ" "ㅏ+" ".+" *e* 0.0229329

(그림 5) 단어분리의 결과

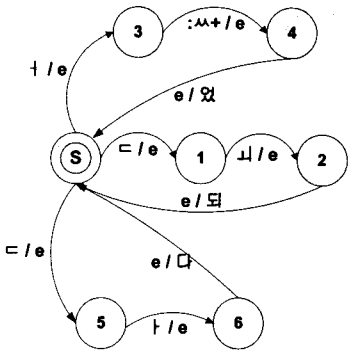
3.3 단어형성변환기

불규칙 변환 등을 효율적으로 적용하기 위한 자소 분리를 통한 처리는 3.4절에서 이루어질 최적 품사 결정의 입력으로 사용되어지는 음절 단위로 묶어져야 한다. 이를 위하여 두 가지의 세부적인 단계를 거치는데, 그 첫 번째로써 분리되어져야 하는 변환 결과물을 분리해주는 단계이다. (그림 6)은 입력된 "됐다"라는 문장에 대하여 재결합을 하기 전에 "ㅐ"가 변환되어져서 나온 "ㅛ+ㅓ"라는 자소 묶음에 대하여 +를 기준으로 "ㅛ+"와 "ㅓ"로 분리시키는 묶음분리변환기를 거친 결과를 보여 준다. 묶음분리변환기는 단어결합변환기의 보조적인 변환기이다.

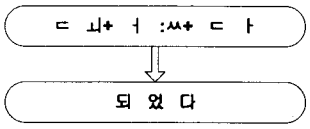


(그림 6) 자소묶음분리의 결과

분리 과정을 거친 뒤에는 미리 구축되어진 단어 사전의 오토마타를 거치면서 사전에 등록되어 있는 변환 가능한 단어들로 결합 된다. 이 과정에서 비록 우선순위는 높았으나, 결합될 수 있는 단어가 없는 후보의 경우는 분석 대상에서 제외되어 진다. (그림 7)은 단어 사전의 오토마타를 보여주는 것으로써 최종 상태에서 다시 S상태로 전이하기 전에 변환된 결과를 출력하여 주는 전이가 있다는 것을 알 수가 있다. (그림 8)은 이 오토마타를 거친 결과를 간략히 보여 준다.



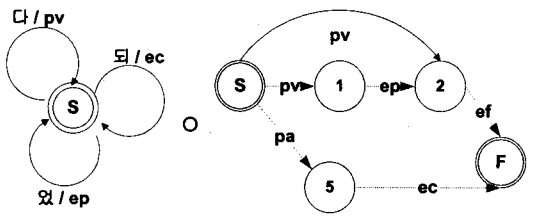
(그림 7) 단어결합변환기



(그림 8) 단어결합의 결과

3.4 최적 품사 결정

자소단위로 자르고, 불규칙 변환을 적용시킨 뒤에 다시 형태소로 재결합된 자소들은 마지막으로 품사를 결정하기 위한 단계를 거치게 된다. 비록 정확한 분석으로 형태소 단위로 나뉘어 졌다고 하더라도 의미해석상 모호한 "나는"과 같은 문장에 대하여 상황에 맞는 최적의 품사를 결정해야 하는 과정이 필요함을 알 수가 있다. 품사를 결정하는 과정 역시 2가지 과정을 거치게 되는데 첫 번째가 transition이고, 두 번째가 align이다. 첫 번째 과정에서는 입력으로 주어진 단어가 그에 맞는 품사 정보들로 변경되어지고, 두 번째 과정에서 품사관계 오토마타를 거침으로 인하여 문장을 구성함에 있어서 적합한 품사 관계 결과를 얻게 된다. 이 과정에서 가중치를 부여한 간선을 사용하기 때문에 모든 과정을 거쳤음에도 불구하고, 나오게 되는 여러 가지 결과에 대하여 적합한 우선순위를 구할 수가 있게 된다. (그림 9)의 왼쪽 변환기는 단어에 해당되는 품사들로 변환시켜주는 transition정보를 가진 오토마타를 보여 주고, 오른쪽 변환기는 품사간의 관계를 가지는 관계 오토마타를 보여준다.



(그림 9) 품사결정변환기

4. 형태소 정보의 축출

본 논문은 KAIST 말뭉치(김재훈 외, 1997)를 사용하여 확률 및 정보를 축출한다. 이 말뭉치는 다양한 장르[1]의 문장이 포함되어 있으며 총 어절수는 175,524 어절이다. 이 말뭉치를 이용해서 이 논문에서 필요한 유한상태변환기 생성을 위한 정보를 추출하였으며, 아래와 같은 정보들이 있다.

(1) 음절을 자소열 변환
갈 ㄱㅏㄹ

(2) 자소간의 변환 확률
ㅐ ㅏ+ㅣ 0.211525
ㅐ ㅐ 0.684332

(3) 자소간의 전이 확률
ㅏ ㅆ 0.176471

(4) 자소열을 이이용한 단어 생성
ㅂㅏㄱㅅㅜ 박수

(5) 단어의 품사 확률
새 mm 0.840909
새 nb 0.072727
새 nc 0.068182
새 pv 0.018182

(6) 품사와 품사 사이의 전이 확률
co ec 0.165387
co ef 0.592808

(1), (4), (5), (6)의 정보는 품사태깅 결과 정보에서 간단히 추출해 낼 수가 있다. 하지만, (2)와 (3)의 정보를 추출하기 위해서는 원문과 태깅된 결과를 비교하여 자소간의 변환과 확률을 추출하는데, 우선 원문과 태깅된 문장 간의 차이점을 다음과 같이 마킹하고,

ㅎㅐㅆㄷㅏ. ㅎㅏ+ㅣㅆㄷㅏ.

ㅎ&ㅆㄷㅏ. ㅎ&&&ㅆㄷㅏ.

이 마킹된 정보를 기반으로 하여 자소가 변환되는 목록 및 확률을 구축하는 과정을 거치게 된다.

5. 결론 및 향후 과제

이 논문은 유한상태변환기만을 이용하여 한국어 형태소 분석 및 품사 태깅 시스템을 제안한다. 기존의 한국어 형태소 분석 시스템들은 규칙기반 형태소 분석기가 주를 이루고 한국어 품사 태깅 시스템은 은닉마르코프 모델 기반 품사 태깅이 주를 이루었다. 한국어 형태소 분석의 경우 유한상태변환기를 이용한 경우도 있었으나(이성진 1992), 이 방법은 변환기를 작성하기 위한 규칙을 수작업으로 구축해야 하며, 그 규칙에 따라서 사전이 작성되어야한다. 이 논문에서는 품사 태깅 말뭉치를 이용해서 유한상태변환기에서 필요한 모든 변환 규칙을 자동으로 추출한다. 이런 방법으로 네 종류의 변환기, 즉, 자소분리변환기, 단어분리변환기, 단어형성변환기, 품사결정변환기를 자동으로 구축한다. 구축된 변환기들은 결합연산(composition operation)을 이용하여 하나의 유한상태변환기를 구성하여 한국어 형태소 분석과 동시에 한국어 품사 태깅을 수행한다. 이 방법은 하나의 유한상태변환기만을 이용하기 때문에 복잡도는 선형시간(linear complexity)를 가지면, 형태소 분석기와 품사 태깅 시스템을 매우 짧은 시간 내에 개발할 수 있었다.

현재는 주어진 품사 태깅 말뭉치에서 자동으로 정보를 축출하는 데에서 발생하는 품사 정보의 불일치가 발생한다. 이를 보완하기 위하여 품사결정 변환기의 중간 부분에 품사를 정형화 되게 변환하는 변환기를 추가할 수 있도록 해야 할 것이다. 또한, 이 논문이 실제 시스템에 적용되었을 경우 얻어지는 부수적인 효과들에 대해서도 연구를 할 것이다.

1) 신문기사(40,428 어절), 수필(41,666 어절), 교과서(50,208 어절), 기술문서(2,729 어절), 소설(40,498 어절)로 구성되었다.

참고문헌

[1] Knight, K. and Al-Onaizan, Y. (1999) A Primer on Finite-State Software for Natural Language Processing, http://www.isi.edu/publications/ licensed-sw/carmel/

[2] Manning, C. and Schütze, H. (1999) Foundation of statistical Natural Language Processing. The MIT press.

[3] 강승식, (2002) 한국어 형태소 분석과 정보검색, 홍릉과학출판사.

[4] 김재훈, "가중치 망 모델을 이용한 한국어 품사 태깅", 한국정보과학회논문지, 제25권, 제6호, pp. 951-959, 1998년.

[5] 김재훈, 김길창 (1995) 한국어 품사 부착 말뭉치의 작성요령 : KAIST말뭉치, KAIST, 전산학과, 기술문서(CS-TR-95-99).

[6] 신상현, 이근배, 이종혁 (1997) "통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템", 한국정보과학회논문지, 제24권, 제2호, pp. 160-169.

[7] 이성진, (1992) Two-level 한국어 형태소 해석, 한국과학기술원 석사학위 논문.

[8] 임희석, 김진동, 임해창 (1997) "어절 태그 변형 규칙을 이용한 한국어 품사 태거", 한국정보과학회논문지, 제24권, 제6호, pp. 673-684.

[9] 임희석, 윤보현, 임해창 (1995), "배제 정보를 이용한 효율적인 한국어 형태소 분석기", 정보과학회논문지, vol. 22, no. 6, pp. 957-964.