한국어 형태소 분석에서 확장된 최장 일치법을 이용한 의사 투-레벨 모델

한용기*, 이근용**, 이기오*, 이용석**
*서해대학 전자계산과
**전북대학교 컴퓨터과학과 언어정보공학실
{yghan, giolee}@sohae.ac.kr
keylee@cs.chonbuk.ac.kr, yslee@moak.chonbuk.ac.kr

(pseudo two-level model using extended longest match method in korean morphological analysis)

Y. G. Han*, K. Y. Lee**, G. O. Lee, Y. S. Lee**
*Dept. of Computer Science, Sohae College
**Dept. of Computer Science, Chonbuk National University

요 약

한국어 형태소 분석 방법 중 좌우 최장일치법은 분석 모델은 단순하지만 분석 후보의 과생성과 backtracking 발생 문제 등으로 인하여 연구가 미진하였다. 또한 Two-level 모델은 최장일치법에서 나타나는 문제점, 많은 two-level 규칙의 필요성, 그리고 중간 결과의 이용 문제로 인하여 한국어에 거의 적용되지 못하고 있다. 본 논문에서는 형태소 분석의 일반적인 모델로 알려진 Two-level 모델의 단점인 backtracking 문제와 분석 후보의 과생성 문제 그리고 중간 결과의 미사용 문제를 좌우 최장일치법을 이용하여 처리하는 방법론을 제안하고 좌우 최장일치법이 한국어 형태소 분석 방법에 효율적으로 적용될 수 있음을 제시한다.

1. 서론

형태소 분석은 여러 형태소들의 조합이 표층형태로 나타나는 어절들로부터 의미를 갖는 최소 단위인 형태소들을 분석하는 것이다. 형태소 분석 과정은 단위 형태소들의 분리, 원형 복원, 적합한 단위 형태소들의 조합 추출로 이루어진다.

이러한 형태소 분석 과정에 따라 다양한 분석 방법들이 연구 제시되었다[강승식 94a, 강승식 94b].

한국어 형태소 분석 방법 중에서 최장일치법은 좌우로 분석을 하는지, 우좌로 분석을 하는지에 따라 분석 결과가 다르게 나타나 한국어에 적합하지 않은 방법으로 인식되고 있으며 또한 이 방법을 적용했을 경우, 분석 후보의 과 생성과 이로 인한 사전의 탐색 회수가 증가하고 backtracking이 발생하여 한국어 형태소 분석 방법으로 부적합한 것으로 인식되고 있다. 따라서 이 방법론에 관한 연구는 초기에 제안된 후, 거의 이루어지지 않았다.

형태소 분석 방법론 중에서 Two-level 모델[kimmo 84]은 언어 독립적인 방법론으로 여러 언어에 적용한 결과 모델이 간단하고 효율적인 방법으로 인식되어 이 모델을 한국어 형태소 분석 방법에 적용한 연구들도 제시되었다.

그러나 많은 연구에서 Two-level 모델은 한국어 형태소 분석 과정에서 많은 형태소 분석 후보들을 생성하고 중간 결과를 이용하지 못하며 빈번한 backtracking 발생으로 한국어 형태소 분석에는 효율적이지 못한 것으로 알려졌다[김덕봉 96].

Two-level 모델은 단순한 모델이기 때문에 이 모델이 가지는 단점을 제거할 수 있다면 한국어 형태소 분석에서 단순한 모델을 정형화할 수 있다. 본 논문에서는 Two-level 모델이 가지는 단점을 해결하고 한국어 형태소 분석에서도 효율적인 확장된 최장일치법을 이용한 의사 Two-level 모델을 제안한다. 한국어 형태소 분석에 비효율적인 방법으로 알려진 최장일치법이 two-level 모델이 가지는 문제를 해결하고 한국어에도 매우 효율적인 방법임을 제시한다.

2. 최장일치법과 Two-level 모델

최장일치법은 단어를 이루는 형태소의 길이에 따라 단어를 가능한 모든 형태소로 분할 한 다음 단어를 이루는 부분 문자열 집합을 구할 때, 단어를 이루고 있는 형태소들의 집합 중에서 가장 긴 형태소를 먼저 선택하여 검사하는 방법이다. 최장일치법은 형태소 분석 방향에 따라 좌우 최장일치법과 우좌 최장일치법으로 나누어질 수 있다. 좌우 최장일치법은 알고리즘이 단순하고 이해하기 쉬운 방법이고 우좌 최장일치법은 조사, 어미를 사전에 저장하고 주기억장치에 저장하여 우에서 좌로 입력 어절을 검색하면서 음절 평가 함수를 도입하여 구하는 방법이다. 그러나 좌우 최장일치법을 적용할 경우, 가능한 형태소 분석 후보가 과다하게 생성될 수 있어 사전의 탐색회수가 많아지고 빈번한 backtracking이 발생한다는 단점을 가지고 있는 것으로 인식되고 있다. 이런 문제점들 때문에 이 방법은 한국어 형태소 분석에서 비효율적인 방법으로 인식되어 형태소 분석 방법론으로 제시되었다는 정도로 인식되었고 이로 인하여 최장일치법에 대한 연구는 이루어지지 않았다.

언어 독립적인 형태소 분석 방법으로 Kimmo Koskenniemi가 제안한 two-level 모델은 입력 문자열의 표층형과 어휘형을 일치시키기 위해서 변형 규칙들의 합성 규칙인 Two-level 규칙을 사용하는 방법이다. 이 모델에서 필요한 구성요소들은 단어의 원형, 어미, 접사 등이 저장되어 있는 트라이 구조의 사전과 통합 규칙인 Two-level 규칙 그리고 유한 자동 기계(FST: Finite State Transducer)로 이루어진다.

Two-level 모델에 의한 형태소 분석 방법은 모든 two-level 규칙을 적용하면서 가능한 모든 후보를 생성하는 동시에 트라이 구조로 된 사전과 일치하는지를 검사하여 표층형과 일치하는 어휘형을 형태소 분석 결과로 출력한다.

Two-level 모델은 굴절이 심하게 일어나는 언어에는 각 형태소를 분리하고 그 원형을 복원하는데 효율적인 방법이나 어근과 접사의 연속적인 결합이 많은 교착어에서는 많은 수의 규칙을 기술해야 하므로 구현이 어려운 것으로 알려졌다.

Two-level 모델을 한국어 형태소 분석 방법에 적용한 연구들이 제시되었는데[강승식 94b] 분석 후보의 과 생성 문제와 backtracking 문제 그리고 중간의 결과를 이용하지 못하는 문제들이 제시되어 한국어 형태소 분석 방법으로는 적합하지 못하다는 평가를 받아왔다.

형태소 분석 방법의 일반적인 분석 방법론으로 알려진 two-level 모델이 가지는 문제점이나 좌우 최장일치법이 가지는 문제점은 많은 부분이 일치하고 있다. 즉, 분석 후보의 과 생성 문제나 backtracking 그리고 많은 사전 탐색 횟수 등이 유사하다. 좌우 최장일치법을 이용하여 two-level 모델이 가지는 단점을 해결할 수 있으면 좌우 최장일치법으로도 한국어 형태소를 효율적으로 분석할 수 있다. 한국어 형태소 분석 시, 좌우 최장일치법을 이용한 의사 two-level 모델은 자소 단위의 trie 사전과 형태소 원형 복원 규칙과 사전의 탐색 과정을 통합한 약 20 여개의 한국어 two-level 규칙 그리고 FST로써 형태소 분석 후보의 과 생성 문제와 이로 인한 backtracking 문제 그리고 중간 결과를 이용하지 못한 문제를 해결한다.

3. 확장된 최장일치법을 이용한 의사 two-level 모델

확장된 최장일치법은 기존의 좌우 최장일치법이 가지는 문제점을 해결하기 위하여 원형 복원 알고리즘과 자소 단위 Trie 사전의 탐색 알고리즘을 통합한 알고리즘을 사용하고 한국어를 효율적으로 분석하기 위하여 형태소 분석의 결과가 구문분석 단계의 입력에 적합한 결과를 생성하도록 좌우 최장일치법을 개선하고 확장시킨 방법이다.

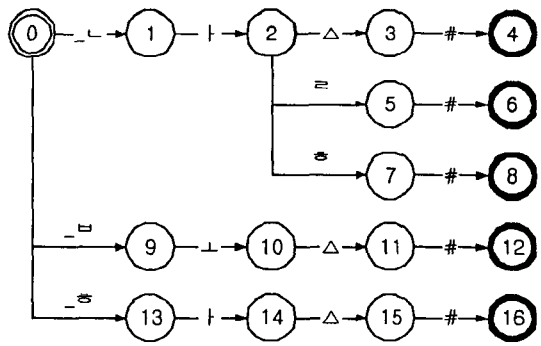
이 방법은 한국어에서 빈번하게 나타나는 여러 개의 용언들이 보조적 연결 어미를 매개로 해서 하나의 용언으로 사용될 수 있을 경우, 이들 용언을 각각의 자립된 용언들의 결합이 아니라 하나의 문법적 범주를 가지는 단일 용언으로 분석한다[이기오 96].

```
Integrate_RootFormRecovery_with Search()
{
  Make_KEY_and_Second_KEY()
  if (HAS_NODE = Exist_Node_in_Trie(KEY[0])
  {
    CURRENT_INDEX_TRIE = HAS_NODE;
    if (Exist_Node_in_Trie(KEY[1])
    {
      CURRENT_INDEX_TRIE = HAS_NODE;
      Recovery_Vowel_Root_Form();
      if (HAS_NODE=Exist_Node_in_Trie(KEY[2])
      {
        CURRENT_INDEX_TRIE = HAS_NODE;
        Recovery_Consonant_Root_Form();
      }
      else Recovery_Consonant_Root_Form();
    }
    else Recovery_Vowel_Root_Form();
  }
}
```

<표 1> 원형복원 모듈과 사전 탐색 모듈의 통합 알고리즘

한국어 형태소 분석에서 Two-level 모델을 적용할 경우, 형태소 분석 후보의 과 생성과 표층형을 어휘형에 일치시키는 과정에서 많은 backtracking이 발생하고 중간 결과를 이용하지 못하는 단점은 한국어 형태소 분석의 후보의 prefix는 같고 suffix가 다르다는 특성과 TRIE가 prefix를 공유한다는 특성을 이용하여 backtracking 문제를 해결하고 형태소 원형 복원 모듈과 사전 탐색 모듈을 통합한 알고리즘을 이용하여 과 생성 문제를 해결하며 중간 결과의 이용 문제를 해결할 수 있다. 원형 복원 알고리즘과 사전 탐색 알고리즘의 통합 알고리즘의 자세한 사항은 [김철수 98]에 기술되어 있고 여기에서는 다음과 같은 개괄적인 알고리즘만 <표 1>에 기술한다.

확장된 최장일치법을 이용한 의사 two-level 모델을 적용하여 한국어 어절 "나는 하늘을 날아 보려 하였다"에 대한 분석 과정을 보임으로써 제안한 방법을 설명한다. 예문 "나는 하늘을 날아 보려 하였다"의 자소 단위 Trie 사전 구조와 유한 자동기계(FST)는 [그림 1]과 같다. [그림 1]에서 '_'이 있는 자음은 초성을 나타내고 있고 '_'이 없는 자음은 종성을 나타내고 있다. 따라서 시작 상태인 0번 상태에서 초성 '_ㄴ'을 만나면 1번 상태로 전이가 일어나고 2번 상태에서 종성이 없거나(△) 종성 'ㄹ' 또는 'ㅎ'을 만나면 3번, 5번 상태나 7번 상태로 전이가 일어남을 알 수 있다.



<그림 1> 예문을 위한 자소 단위 Trie 사전과 FST 표현

한국어 형태소 분석에서 two-level 모델을 적용하기 위해서는 많은 two-level 규칙이 요구되는 것으로 알려졌다. [이성진 92]에서는 50여 개의 규칙과 그에 상응하는 오토마타를 사용하고 있다. 본 논문에서는 한국어 two-level 규칙을 약 20여 개를 사용하고 같은 수의 오토마타를 사용하고 있다. 여기에서는 전체 규칙을 다 기술하기는 어려우므로

한국어 two-level 규칙 중에서 예문을 처리하기 위한 규칙만 기술한다. 예문을 분석하기 위한 규칙들은 다음과 같다.

'ㄹ'탈락

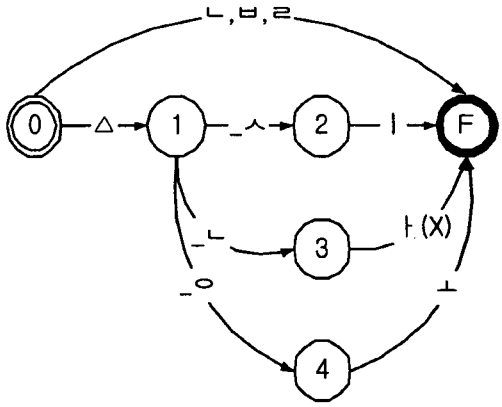
```
int IsFRem(int pos, uchar* codeS)
{
  int len = strlen(codeS);

  if(!IsJongSung(pos)) return 0;
  if(codeS[pos] == FILLCODE)
  {
    if (pos > len-3) return 0;
    if (codeS[pos+1] == ChoT &&
        codeS[pos+2] == JungL)
      return REM_F;
    if (codeS[pos+1] == ChoS &&
        codeS[pos+2] != JungK)
      return REM_F;
    if (codeS[pos+1] == ChoD &&
        codeS[pos+2] == JungH)
      return REM_F;
    return 0;
  }
  if (codeS[pos] != JongS &&
      codeS[pos] != JongQ &&
      codeS[pos] != JongF)
    return 0;
  return REM_F;
}
```

<표 2> 'ㄹ'탈락 알고리즘

기술된 'ㄹ' 탈락 알고리즘에서 'pos'는 현재 위치를 나타내고 'codeS'는 키보드 상의 S 위치의 한글 자판 'ㄴ'을 나타낸다. 'ChoT', 'ChoS', 'ChoD'는 각각 키보드 상의 T, S, D 위치의 한글 초성 '_ㅅ', '_ㄴ', '_ㅇ'을 나타내고 'JungL', 'JungK', 'JungH'는 각각 키보드 상의 T, S, D 위치의 한글 중성을 나타낸다. 또한 'JongQ', 'JongS', 'JongF'는 종성을 나타낸다. 이 규칙을 유한 자동기계로 표현하면 다음과 같다. 아래 그림에서 'ㅏ(x)'는 중성이 'ㅏ'가 아닌 경우를 나타낸다.

ㄹ탈락



<그림 2> 'ㄹ'탈락의 유한 자동기계 (FST) 표현

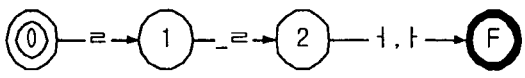
'르'불규칙

```
int IsIrrFM(int pos, uchar* codeS)
{
  if (!IsJongSung(pos)) return 0;
  if (codeS[pos] != JongF) return 0;
  if (codeS[pos+1] != ChoF) return 0;
  if (codeS[pos+2] != JungJ &&
      codeS[pos+2] != JungK) return 0;
  return IRR_FM;
}
```

<표 3> '르' 불규칙 알고리즘

'르' 불규칙에 대한 유한 자동기계의 표현은 다음과 같다.

르 불규칙



<그림 3> '르'불규칙의 유한 자동기계 (FST) 표현

ㅎ'불규칙

```
int IsIrrG(int pos, uchar* codeS)
{
  if (codeS[pos] == JongS ||
      codeS[pos] == JongF ||
      codeS[pos] == JongA ||
      codeS[pos] == JongQ)
    return IRR_G;
  return 0;
}
```

<표 4> 'ㅎ' 불규칙 알고리즘

ㅎ 불규칙

<그림 4> 'ㅎ'불규칙의 유한 자동기계(FST) 표현

예문에서 '-아/어 보다'와 '-려 들다/하다' 같은 보조적 연결어미를 매개로 여러 개의 용언들이 묶여 하나의 용언으로 표현될 수 있는 복합 용언은 어미 사전에 특정한 flag를 표시하여 이들이 다음 단어로 전이가 일어날 수 있음을 나타내어 '날아 보려 하였다'를 하나의 용언으로 분석한다. 여기에서는 복합동사의 처리에 대한 자세한 사항은 제외하기로 한다.

형태소 분석에서 처리하기 어려운 문제는 미등록어 문제이다. 예를 들어 "나는 영희와 같이 하늘을 날아 보려 하였다"의 예문에 대한 분석의 경우, 트라이 사전을 탐색할 경우 '영희'는 미등록어이므로 탐색에 실패하게 된다. 본 논문에서 제안한 형태소 분석 시스템에서는 '영희<미등록어>'를 분석하기 위하여 사전 탐색이 실패할 경우 미등록어로 간주하고 좌우 방법으로는 정확하게 미등록어의 위치를 식별하기 어렵기 때문에 우좌 최장일치법을 적용하여 조사를 기준으로 앞부분의 음절을 미등록어를 분석해내고 있다. '알콜신드롬은 많은 사람들을 죽음으로 이끌었다'에서와 같이 '미등록어 + 명사 + 조사'로 구성된 미등록어를 '미등록어 + 조사'로 분석하는 경우에 대해서는 현재 정확하게 미등록어를 식별하기 위한 연구가 진행중이다.

4. 실험 및 평가

본 논문에서 제안한 형태소 분석 시스템의 성능을 평가하기 위하여 우리는 먼저 학습 코퍼스의 형태소 분석 결과인 result set을 만들고 MATEC에서 제공한 33855 어절을 분석하여 그 결과를 비교하였다. 본 논문의 test 환경은 450Mhz의 PentiumIII에서 실험하였고 처리시간은 2분 45초, 평균 분석 후보의 개수는 1.7정도 나왔는데 그 이유는 대 분류 중심의 결과를 생성하면서 자질 형태로 세 분류 품사 정보를 표현하기 때문이다.

Test 어절의 실험 결과, 약 92%의 정확도를 얻었다. 형태소 분석 실패의 원인은 다음과 같은 몇 가지 이유로 분석되었다. 첫째, 사전에 잘못된 등록 문제로 인한 실패. 예를 들어, "치르다"는 99년판 동아 새국어 사전에는 '르' 불규칙이 있는 것으로 표시되어 있는데 실제로는 '르' 불규칙이 나타나지 않았다. 둘째, 복합명사 부분에서의 실패. "질서정연한"의 경우, '질서 + 정연 + 하 + ㄴ'가 result set에 들어있었는데 '질서 + 정연하 + ㄴ'로 분석되어야 맞는 분석 결과로 되어 result set의 잘못된 기술과 관련이 있었다. 셋째, 미등록어 추정 실패. '김대중'이라는 어절에서는 '김(nc)' + '대(nc)' + 중(xn:명사화 접미사)'로 분석되어 복합명사 추정에 의해 '김대 + 중'으로 잘못 분석되고 '김대중'이라는 미등록어를 추정하지 못했다. 그 이유는 복합명사 추정 모듈에서 느슨한 제약조건 때문이었다. 현재 복합명사 추정 모듈에서 강한 제약조건을 가하기 위하여 연구하고 있고 미등록어 문제도 연구 중에 있다.

5. 결론

본 논문에서는 한국어 형태소 분석에서 확장된 최장일치법을 이용한 의사 two-level 모델에 대하여 제안하였다. 기존의 좌우 최장일치법이나 two-level 모델은 공통적으로 형태소 분석 후보의 과 생성 문제와 이로 인한 사전 탐색에서의 backtracking 발생 문제 그리고 사전 탐색 시간의 증가 문제를 가지고 있다. two-level 모델은 중간

결과를 이용하지 못하는 문제를 가지고 있다. 이런 문제는 이들 방법이 한국어 형태소 분석 방법으로 광범위하게 적용되지 못하고 연구되지 못하는 원인을 제공한다. 본 논문은 이런 단점을 해결하기 위하여 Trie의 특성과 형태소 분석의 원형복원 모듈과 사전의 탐색 모듈의 통합을 통하여 해결하였다. 또한 좌우 최장일치법을 이용하여 two-level 모델에서 발생하는 문제를 해결하고 이 방법으로도 한국어 형태소 분석을 효율적으로 할 수 있다는 것을 제시하였다. 본 논문에서 제안된 시스템을 이용한 코퍼스 분석 결과는 최상의 분석 결과를 제시하지는 않지만 좌우 최장일치법을 이용한 분석 방법이 한국어 형태소 분석 방법에서 하나의 주된 방법이 될 수 있음을 보여주고 있다. 현재 미등록어 문제에서 '미등록어 + 미등록어 + 조사'를 '단일 미등록어 + 조사'로 분석하는 부분에 대한 연구와 복합동사 처리를 전체적으로 어떻게 생성해주어야 하는지에 대한 연구가 진행되고 있어 이런 부분이 해결된다면 더 나은 한국어 형태소 분석 방법으로 인식될 수 있을 것이다.

[참고 문헌]

[이성진 92] 이성진, 김덕봉, 서정연, 최기선, 김길창 "Two-level 모델을 이용한 한국어 용언의 형태소 해석," 한국 정보과학회 발표 논문집, 19권 2호, pp. 993~996, 1992.

[김철수 98] 김철수, "한국어 형태소 분석 환경을 효율적으로 지원하는 전자사전 구조," 전북대학교 컴퓨터과학과 박사학위 논문, 1989.

[이기오 94] 이기오, 김기철, 이용석, "형태소 분석 주도의 한국어 복합동사 처리," 제6회 한글 및 한국어 정보처리 학술대회, pp. 119~127, 1994.

[kimmo 84] Koskenniemi Kimmo, "A general computational model for word-form recognition and production," *COLING 84*, pp. 178~181, 1984.

[강승식 94a] 강승식, "다층형태론과 한국어 형태소 분석 모델," 제 6회 한글 및 한국어 정보처리 학술대회, pp. 140~145. 1994.

[강승식 94b] 강승식, "한국어 형태론의 특성과 형태소 분석 기법," 한국 정보과학회지, 12권 8호, pp. 47~59, 1994.

[김덕봉 96] 김덕봉, "예측 중심의 형태소 분석 : 한국어 어절 인식을 위한 계산 모델," 한국 과학기술원 전산학과 박사학위 논문, 1996.