CRF기반 한국어 형태소 분할 및 품사 태깅에서 두 단계 복합형태소 분해 방법

나승훈[O] 김창현, 김영길
한국전자통신연구원
nash@etri.re.kr, chkim@etri.re.kr ,kimyk@etri.re.kr

Two-Stage Compound Morpheme Segmentation in CRF-based Korean Morphological Analysis

Seung-Hoon Na[O], Chang-Hyun Kim, Young-Kil Kim
Natural Language Processing Laboratory
Electronics and Telecommunication Research Institute

요 약

본 논문은 CRF기반 한국어 형태소 분석 및 품사 태깅 과정에서 발생하는 미등록 복합형태소를 분해하기 위한 단순하고 효과적인 방법을 제안한다. 제안 방법은 1) 복합형태소를 내용형태소와 복합기능형태소로 분리하는 단계, 2) 복합기능형태소를 분해하는 두 단계로 구성된다. 실험 결과, 제안 알고리즘은 Sejong데이터에 대해, 기존의 lattice HMM 대비 높은 복합형태소 분해 정확률 및 두드러진 속도 개선을 보여준다.

주제어: 형태소분석, 복합형태소 분해, 복합기능형태소, 내용형태소, 최장일치

1. 서론

최근의 한국어 형태소 분석 연구는 입력문 전체에 대해 통계기반 모델을 수행하는 방법이 주된 축을 이루고 있다 [1,7,9,10,11]. 이 중, [7]은 한국어 형태소 분석을 형태소 분할 및 품사 태깅으로 나누어 각 단계에 CRF를 적용하는 방법을 제안했으며, [11]는 이를 확장하여 띄어쓰기, 형태소 분할 및 품사 태깅 단계를 모두 통합하는 joint모델을 제안했다.

그러나, 이들 연구에서는 복합형태소 처리 과정이 별도로 필요하다. 복합형태소 (compound morpheme)란 복수개의 단위형태소 (atomic morpheme)로 구성되는 형태소를 일컫는데, 표층형이 입력문과 일치하는 특징을 가지고 있다. 복합 형태소를 단위형태소로 분할하는 간단한 방법은 학습데이터에 나타나는 기분석 패턴을 이용하는 것이다 [7,11], 그러나, 복합형태소는 용언류의 활용형이 많아 대부분 개방어이므로 미등록어가 필연적으로 발생하게 된다. 미등록 복합형태소 처리를 위한 일반적인 모델로, [7]은 lattice HMM기반 방식을 제안하였다. 그러나, 이 방법은 높은 시간 복잡도를 지니며, 결과적으로 CRF와 이질적인 생성 모델의 도입으로 형태소 전 분석 과정을 한 가지 모델로 통일시키지 못했다.

본 논문은 미등록 복합형태소 처리를 위해 보다 단순하고 효율적인 방법을 제안한다. 제안 방법의 핵심은 기능형태소만을 기분석화하는 것으로, 다음의 두 단계로 이루어진다. 1) 복합형태소를 내용형태소와 복합기능형태소로 분리하는 단계, 2) 복합기능형태소를 분해하는 단계이다. Sejong 품사 부착 코퍼스에서 실험 결과, 제안 방법은 기존 lattice HMM방식에 비해 복합형태소 분해 정확률을 향상시킬 수 있을 뿐 아니라 두드러진 속도 향상을 보여주었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 다루고, 3장에서는 제안하는 두 단계 복합형태소 분해 방법을 소개하고, 4장에서는 실험을, 5장에서는 결론 및 향후 계획에 대해서 언급한다.

2. 관련 연구

최근의 한국어 형태소 분석의 연구는 입력문 전체에 별도의 언어학적 지식 사용없이 통계기반 모델을 수행하는 방법이 주된 축을 이룬다 [1,7,9,10,11]. 그러나, 한국어는 축약/불규칙 변이로 인해, 사전형과 표층형과 일치하지 않은 문제가 있어, 중국어의 경우와 달리 단어분리 (word segmentation) 및 품사 태깅 (POS tagging) 방법을 직접 적용할 수 없다. 이른바 *표층형과 사전형 불일치 문제*를 해소하는 방식에 따라, 최근 통계 기반 한국어 형태소 분석 연구는 다음의 두 가지 방식으로 나눌 수 있다.

먼저, 형태소 단위의 접근법으로, 표층형과 사전형 불일치를 복합형태소의 도입을 통해 해결하는 방식이다 [7,11]. 서론에서 언급했듯이, 복합형태소 분해의 대표적인 방법은 학습 코퍼스에 발생하는 복합형태소와 분해 결과를 기분석 패턴을 사용하는 것이다. 미등록 복합형태소의 일반적 처리를 위해, [7]는 lattice HMM의 생성 모델의 사용을 제안했다.

두 번째 방법은 [9,10]의 음절 단위의 접근법이다. 표층형으로부터 사전형으로 원형 복원을 위해 *음절 복원 사전*을 사용하는데, 이는 일종의 *음절 레벨에서의 기분석패턴*으로, 해당 음절/태그에 대해 원래 사전형에 대응되는 음절열/태그열을 저장해놓은 것이다. 가령, 축약의

경우, *복합음절*을 도입하여, 복합음절/음절복합태그에 대해 원래의 사전형의 음절열/태그열을 얻어낼 수 있도록 음절 복원 사전을 구성한다 (가/VVEP=>가/VV+았/EP).

3. 제안 방법

3.1. 개요

제안 방법은 다음의 두 단계로 구성된다.

1) 내용형태소와 복합기능형태소로 분할 (음절테이블 활용)
2) 복합기능형태소 분할 (기분석 활용)

그림 1에 제안 방법의 도식도를 보여준다. 첫 번째 단계에서는 복합형태소를 *내용형태소* (content morpheme)와 *기능형태소* (functional morpheme)로 분할한다. 내용형태소는 단위형태소로 구성, 기능형태소는 여러 개의 단위형태소로 이루어진 *복합기능형태소* (compound functional morpheme)이다. 두 번째 단계에서는, 복합기능형태소를 기분석패턴을 이용하여 단위기능형태소로 분해하는 방식으로 [7]과 동일한 방법을 사용한다.

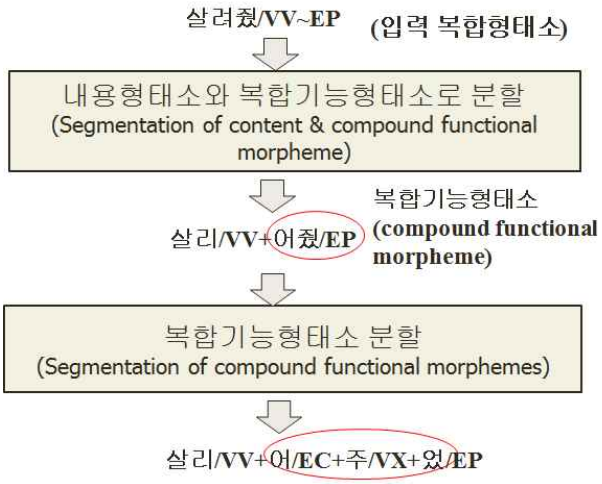

그림 4 제안 복합형태소 분해 절차

이해를 높기 위해, 아래, 그림 1의 제시된 예를 보다 상세히 기술한다. 복합형태소의 입력 예는 "살려줬/VV~EP "이며, VV~EP는 *복합태그*이다. 우선 "살려줬/VV~EP"이 복합형태소 기분석 패턴에 있는지를 조사한다. 기분석 패턴에 없을 경우 제안하는 2단계 과정을 통해 분해가 시도된다.

첫 번째 단계에서는 주어직 복합형태소가 *내용형태소*인 "살리/VV"와 *복합기능형태소*인 "어줬/EP"으로 분핸해된다. 여기서, EP는 *복합기능형태소의 복합태그*로서 3.2절에서 보다 명확히 정의된다. 다음으로, 두 번째 단계에서는, 복합기능형태소 "어줬/EP"에 대한 기분석패턴에 기반하여 분해를 시도한다. 기분석 패턴으로부터 "어줬/EP: 어주/VX+었/EP"을 얻어내, 최종적으로 "살리/VV+어주/VX+었/EP"의 분해 결과를 얻는다.

제안 분해 방법은 본래 복합형태소단위로 기분석 패턴 중, 기능형태소에 대한 기분석 리스트만 유지하면 되기 때문에 원칙적으로 기분석 패턴의 개수를 대폭 감소시킬 수 있게 된다. 무엇보다 복합기능형태소는 1개의 단일형태소로 이루어지고, 2개 이상이 많지 않다는 점, 기능형태소는 용언류보다는 closed word에 해당되어 대부분은 사전에 등재 가능하는 점으로 인하여, 복합기능형태소의 기분석 패턴의 사용은 복합형태소 기분석패턴 기반 방법보다 현실성이 높다고 할 수 있다.

두 단계 복합형태소 분해를 위한 지식/사전은 모두 학습데이터로부터 자동으로 얻어진다. 다음 절부터, 분해 지식/사전을 어떻게 학습하는지 상세히 기술하도록 한다.

3.2. 복합형태소 분해 지식 학습

제안 복합형태소 분해를 수행하기 위해서는 다음의 세 가지 정보가 필요하다.

- **음절 매핑 테이블**, ST (*syllable mapping table*): 입력 표층형 s에 대해서 ST(s)는 원래 사전형과 해당되는 빈도수의 리스트를 리턴한다. ST의 예는 다음과 같다.

- ST(했)= {하았}
  ST(려)= {리어}
  ST(줬)= {주었}
  ST(써)= {쓰어}
  ST(겨)= {기어}
  (표기의 단순화를 위해 빈도수는 생략하였다)

  본 논문에서는 축약 문제로 한정, 입력음절열을 한음절, 출력음절로 2음절로 제한하여 설명한다. 물론, 제안 방법은 불규칙 활용 등을 포괄하는 형태로 알고리즘을 손쉽게 확장될 수 있다.
- **복합기능형태소 기분석 사전**, FLEX (*preanalyzed patterns for compound functional morpheme*): 입력 형태소 m, t에 대해서 FLEX(m,t)는 단위기능형태소 분해 결과 및 빈도수의 리스트를 리턴한다.
- **내용형태소 사전**, CLEX (*lexicon of content morphemes*): 입력 형태소 m, t에 대해서 CLEX(m,t)는 내용형태소 분해 결과 및 빈도수의 리스트를 리턴한다.

위의 세 가지 정보 중 음절 매핑 테이블 (ST)와 복합형태소 기분석 사전 (FLEX)은 복합형태소의 기분석 리스트로부터, 내용형태소 사전 (CLEX)은 학습코퍼스의 단위형태소로부터 얻어진다. ST, FLEX, CLEX모두 리스트는 빈도수로 정렬(sorting)하여, 가장 빈도수가 높은 엔트리가 리스트의 첫 번째에 위치하도록 배치한다.

보다 명확한 설명을 위해, 아래, ST, FLEX, CLEX의 학습 과정을 형식적으로 기술한다.

분해 지식 학습을 위한 기본 정보 추출

복합형태소에 대한 기분석 엔트리가 $m/t: m_1/t_1, m_2/t_2, \cdots, m_k/t_k$로 주어졌다고 하자. l_i는 m_i의 음절 갯수를, $m_i[j]$는 m_i의 j번째 음절을, $m_i[j \cdots k]$

는 m_i의 j부터 k까지의 음절열을, $m_i[j\cdots]$는 m_i의 j부터 마지막음절까지의 음절열을 가리킨다고 하자. "+"는 음절열의 결합 (concatenation) 연산자라고 정의하자.

학습의 핵심은 주어진 복합형태소 음절열 m과 이를 구성하는 단위형태소들로부터, *head, tail, compressed* (축약음절), *decompressed1* (분해요소음절1), *decompressed2* (분해요소음절2) 의 정보를 추출하는 것으로, 각각에 대한 정의는 다음과 같다.

정의 1. *head*는 처음 단위형태소 m_1의 l_1-1번째까지의 음절열로 정의한다

head = $m_1[1\cdots l_1-1]$

정의 2. tail은 복합형태소의 m의 l_1+1부터 마지막까지의 음절열로 정의한다.

tail = $m[l_1+1\cdots l]$

여기서 l은 m의 길이이다.

정의 3. *compressed*는 복합형태소의 m의 l_1-1번째 음절을 축약된 음절로 정의한다.

compressed = $m[l_1-1]$

정의 4. 축약된 음절 *compressed*에 대응되는 사전형 두음절은 *decompressed1*와 *decompressed2* 다음과 같이 정의된다.

decompressed1 = $m_1[l_1]$

decompressed2 = $m_2[1]$

여기서, 두 개의 분해요소음절, *decompressed1, decompressed2*은 축약음절 *compressed*에 대응되는 정보이다.

예를 들어, "구겨진/VV~ETM: 구기/VV+어/EC+지/VX+ㄴ/ETM"의 경우, 위의 정의에 따라, *head, tail, compressed, decompressed1, decompressed2* 정보는 다음과 같이 주어진다.

head	구
tail	진
compressed	겨
decompressed1	기
decompressed2	어

정의 5. 위의 정의로부터, 내용형태소의 사전형 *cm*, 복합기능형태소 표층형 *fm*은 각각 다음과 같이 정의된다.

cm = m_1

fm = *decomposed2* + *tail*

앞서의 예에서, *cm*과 *fm*은 다음과 같이 정의된다.

cm	구기
fm	어진

위의 정보에 기반, ST, FLEX는 아래와 같이 학습된다.

음절매핑테이블 (ST) 구축: 축약음절 *compressed*에 대해 분해요소음절 *decompressed1+decompressed1에 대한 매핑을* 음절매핑테이블의 한 엔트리로 구성한다.

복합기능형태소 (FLEX) 기분석 사전 구축: 기능형태소 *fm*에 대한 단위형태소들은 $m_2/t_2,\cdots,m_k/t_k$임로, 이들은 복합기능형태소 기분석 패턴의 엔트리로 추가한다.

여기서, 복합기능형태소의 복합태그 *ft*는 t와 $t_1\cdots t_k$에 의존하는데, 다음은 몇 가지 예시들이다.

1. 마지막태그만 사용: t_k
   예) 해진/ETM (행해진/VV~ETM: 행하/VV+아/EC+지/VX+ㄴ/ETM)
2. 시작태그만 사용: F-t_1
   예) 해진/F-VV (행해진/VV: 행하/VV+아/EC+지/VX+ㄴ/ETM)
3. 시작 및 마지막 태그: t_2-t_k
   예) 해진/EC~ETM (행해진/VV~ETM: 행하/VV+아/EC+지/VX+ㄴ/ETM)
4. 시작부터 마지막 태그: = $t_2\cdots t_k$
   예) 해진/EC~VX~ETM (행해진/VV~ETM: 행하/VV+아/EC+지/VX+ㄴ/ETM)

본 논문에서는, [7]의 복합형태소의 복합태그의 정의를 참조하여 복합기능형태소의 복합 태그를 다음과 같이 정의한다. 즉, 용언류에 대해서는 시작과 마지막 태그가 주어지므로, 방법 1을 사용, 나머지형태소에 대해서는 모두 방법2를 사용한다. 즉, 방법3-4는 복합형태소가 복합태그가 해당 정보를 제공해줄 수 있을 때에만 활용될 수 있는 것으로 본 논문에서는 고려되지 않는다.

내용형태소 (CLEX) 사전 학습: 위의 두 정보 ST, FLEX와 달리, 단위형태소 사전은 품사 부착 학습코퍼스 전체로부터 얻어진다. 사전의 엔트리는 단위형태소의 사전형과 품사태그 그리고 빈도수로 구성된다.

3.3. 최장 일치 기반 복합형태소 분해 알고리즘

위의 세 가지 지식, ST, FLEX, CLEX를 바탕으로, 본 논문에서 복합 형태소 분해 알고리즘은 최장일치법에 기반한다. 분해 알고리즘의 핵심적인 가정은 *복합형태소내의 축약음절은 최대 단 한 개만이 존재한다*는 것이다. 즉, 복합형태소내의 어느 위치에서 축약이 발생했는지를 판단하는 것이 알고리즘의 메인이 된다.

분해 알고리즘은 *복합형태소 표층형의 마지막 음절부터 순차적으로 해당 음절이 축약음절인지를 조사*한다. 축약여부는 음절매핑테이블을 참조하여 이를 2음절로 분해하고 분해된 음절 앞부분은 내용형태소로, 뒷부분은 기능형태소로 포함된다고 가정하고, 이들이 실제 CLEX, FLEX사전에 있는지 확인한다. 만약, 모두 사전에 있으면 바로 해당 음절위치에서 축약이 이루어진 것이다. 그렇

지 않으면, 음절 위치를 앞으로 한칸 이동하여, 다시 축약음절인지 아닌지 조사한다.

상세한 분해 알고리즘은 다음과 같다.

```
input: m/t  ; 복합형태소

ct = getCTag(t); 내용형태소 태그
ft = getFTag(t); 기능형태소 태그

for (i=l;i ≥ 1;i--)
    syllSet = ST(m[i]); 분해요소음절 획득
    if syllSet == NULL:
       continue
    for each decomps in syllSet:
        cm = m[1 ··· i-1] + decomps[1] ; 내용형태소
        fm = decomps[2] + m[i+1 ··· l]; 기능형태소

        if CLEX(cm,ct) is not empty
          && FLEX(fm,ft) is not emptry:
            return 해당 분해 결과
return NULL
```

여기서, getCTag(t)와 getFTag(t)는 복합 태그 t로부터 각각 내용형태소의 태그 및 복합기능형태소의 태그를 얻어내는 함수이다. 위의 알고리즘에서 CLEX, FLEX에 모든 복수개의 엔트리가 있는 경우에는 기분석 결과를 제시할 때, 애매성이 발생하게 된다. 에매성이 발생하는 경우, 가장 높은 빈도수를 갖는 엔트리를 취했다.

4. 실험

실험을 위해, 세종코퍼스 약 25여만문장중 80%는 학습데이터로, 20%는 테스트데이터로 사용하였다. 형태소 분할 및 품사 태깅 모델로 [11]과 유사하게, 형태소 분할과 태깅을 동시에 수행하는 음절 기반 결합 모델을 적용하였다. 이때 사용되는 자질은 [7]의 형태소 분할을 위해 사용되는 자질과 동일하다.

표 1은 [7]의 CRF기반 형태소 분할/품사 태깅 과정에서, 제안 방법과 기존 방법의 복합형태소 분해 정확도를 비교한 결과이다. 표 1에서 보다시피 제안 방법은 기존 lattice HMM기반 방법보다 높은 분해 정확률을 보여주고 있다.

표 1. 복합형태소 분해에서 제안 두 단계 방법과 lattice HMM 방법 비교 (복합형태소만을 대상으로 평가)

방법	성능 (Accuracy)
기분석+lattice HMM [7]	92.66%
기분석+제안방법	**93.84%**

표 2는 전체 형태소 분석에서 [7]의 CRF기반 방법의 복합형태소 분해 단계가 끝난 후 최종 성능을 비교한 것이다. 표 2에서 보다시피 최종 성능에서도, 제안 방법은 기존 CRF기반 방법대비 다소 높은 성능을 보여준다.

표 3는 전체 테스트 문장에 대해 복합형태소 분해 소요시간에 대해 기존 방법과 제안 방법을 비교한 것이다. 표에서 확인되는 것처럼, 제안 방법은 기존 방법과 비교하여 훨씬 적은 시간을 소요하였다.

요약 하면, 제안 방법은 기존방법 대비, 분해 정확률을 높일 뿐 아니라, 시간 복잡도를 두드러지게 개선할 수 있음을 보여준다.

표 2. 음절 기반 형태소분할/태깅에서 분석 결과에서 제안 복합형태소 분해 방법과 기존 방법과의 비교 (전체 형태소 대상)

방법	성능 (F-measure)
CRF+기분석	97.08%
CRF+기분석+lattice HMM [7]	97.21%
CRF+기분석+제안방법	**97.23%**

표 3. 제안 방법과 기존 방법의 복합형태소 분해의 시간복잡도 비교

방법	소요 시간
lattice HMM [7]	65.76 sec
제안 방법	**4.77 sec**

5. 결론

본 논문은 미등록 복합형태소 분해를 위해 단순하고 효율적인 두 단계 방법을 제안하였다. 제안 방법은 두 단계로 구성되는데 먼저 복합형태소를 내용형태소와 복합기능형태소로 분해하고, 다음으로 복합기능형태소를 기분석패턴으로 처리한다. 결국, 제안 방법은 복합기능형태소에 대해서만 기분석 패턴을 유지하기 때문에 기존의 방식에 비해서 기분석 패턴의 수를 크게 줄일 수 있는 장점이 있다. 실험 결과, 제안 방법은 기존의 lattice HMM기반 분해보다 정확도와 시간 복잡도를 모두 개선하였다.

본 논문에서는 음절매핑테이블을 구성할 때 축약음절만을 고려하였으나, 일반적으로는 불규칙 변형 등 보다 다양한 유형을 포괄하도록 확장하는 것이 필요하다. 또한, 제안 분해 알고리즘은 최장일치 방식에 기반을 두고 있으나, 전체적으로 가장 개연성이 높은 분해 결과를 찾는 글로벌 분해 방법을 도입하여 분해 알고리즘을 보다 정교화 하는 것도 흥미로운 주제가 될 것이다.

참고문헌

[1] Lee, D., & Rim, H. (2005). Probabilistic Models for Korean Morphological Analysis. IJCNLP (pp. 197-202).

[2] Lee, G. G., Cha, J., & Lee, J. (2002). Syllable-Pattern-Based Unknown- Morpheme Segmentation and Estimation for Hybrid Part-of-Speech Tagging of Korean. Computational Linguistics, 28(1).

[3] Shim, K., & Yang, J. (2002). MACH - A Supersonic Korean Morphological Analyzer. COLING (pp. 939-945).

[4] 강승식. (1994). 다층 형태론과 한국어 형태소 분석 모델. 한글 및 한국어정보처리 학술대회 (HCLT) (pp. 140-145).

[5] 강승식, & 장병탁. (1996). 음절 특성을 이용한 범용

한국어 형태소 분석기 및 맞춤법 검사기. 정보과학회 논문지, 23(5).
[6] 강승식. (2002). “한국어 형태소 분석과 정보 검색“, 홍릉과학출판사
[7] 나승훈, 양성일, 김창현, 권오욱, 김영길. (2012). CRF에 기반한 한국어 형태소 분할 및 품사 태깅. In 한글 및 한국어 정보처리.
[8] 신준철, & 옥철영. (2012). 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기. 한국정보과학회 논문지, 39(5), 415-424.
[9] 심광섭. (2011). 형태소 분석기 사용을 배제한 음절 단위의 한국어 품사 태깅. 인지과학, 22(3), 327-345.
[10] 심광섭. (2013). 음절단위의 한국어 품사 태깅에서 원형 복원. 정보과학회논문지: 소프트웨어 및 응용, 40(3), 182-189.
[11] 이창기. (2013). Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅. In 한국컴퓨터종합학술대회 (pp. 604-606).