# moHANA: 해석 정보 정밀화를 통한 형태소 해석 정확률 향상 방안

2008년 2월 21일

서승현*, 강인호*, 김재동**

* 주식회사 워드워즈, ** 카네기멜론대학교/언어공학연구소

# 1. 서론

- 한국어는 형태적인(morphological) 언어적 특성이 강한 언어
- 한국어의 단어조어적인 특성을 효율적으로 기술하는 것이 형태소 분석의 정확률 및 기능을 높이는 방법
- 다차원 해석 사전을 이용하는 형태소 분석기 **moHANA(morphological Hangul Analyzer)**에서 한국어의 형태적인 특성을 반영한 복합어 및 파생어의 처리방식과 이를 이용한 띄어쓰기 오류 어절 처리 방식에 대해 소개

# 2. 복합어 분석(1)

## 2.1 복합명사 처리

◆ 복합명사들을 한국어 형태소 분석에서 정확하게 분석해내는 것이 그 성능의 하나의 척도

◆ 명사들에 "복합명사 형성 제약"정보를 두어 처리

→ ¡복합명사 제약¡의 처리 전,

(1) 고아지지 *고아_{ncn} + 지지_{ncp}고_{pv} +
아_{ef} + 지_{aux} + 지_{ef}

(2) 기독교의 기독교_{ncn}+의_{j}
*기독_{ncn} + 교의_{ncn}

→ ¡복합명사 제약¡의 처리 후,

(3) 고아지지 고_{pv} + 아_{ef} + 지_{aux} + 지_{ef}

(4) 기독교의 기독교_{ncn} + 의_{j}

# 2. 복합어 분석(2)

## 2.2 파생명사 처리

- 한국어의 단어조어방식 중에 발달한 것이 “접미사, 접두사”와 같은 파생접사를 이용한 파생법
- 생산적인 접사들의 특성을 사전과 문법에 기술하여 처리하는 것이 효율적

→ '현', '주' 에 접미사 및 명사결합 제약 정보 입력 전,

(5) 성주현 성주현_{nq_per}
*성주_{ncn} + 현_{nfix}
*성주_{nq_loc} + 현_{nfix}

(6) 이성주 이성주_{nq_per}
*이성_{ncn} + 주_{nfix}

→ '현', '주¡에 접미사 및 명사결합 제약 정보 입력 후,

(7) 나카타현 나가타_{nq_forloc} + 현_{nfix}
미네소타주 미네소타_{nq_forloc} + 주_{nfix}

# 3. 분석실패 어절 분석(1)

## 3.1 다단계 해석 문법

◆ 해석 실패할 경우, 다단계 해석 문법의 다음 단계의 문법을 활용하여 입력에서의 오류를 추정하여 해석을 시도

→ 적용전,

(8) 학생및선생 학생_{ncn} + 및선생_{unk}
안적는 안적는_{unk}
엄마와아이 엄마_{ncn} + 와아이_{unk}

→ 적용후,

(9) 학생및선생 학생_{ncn} + 및_{ad} + 선생_{ncn}
안적는 안_{ad} + 적_{pv} + 는_{ef}
엄마와아이 엄마_{ncn} + 와_{j} + 아이_{ncn}

# 3. 분석실패 어절 분석(2)

## 3.2 부분 기분석 사전

◆ 자주 쓰이는 형태적 패턴을 찾아 기분석 사전에 등록하여 형태소 분석시 분석에 이용

(10) 편지에대한 편지_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef}
청주에가는방법 청주_{ncn} + 에_{j} + 가_{pv}는_{ef} + 방법_{ncn}

◆ 다단계 해석 문법 + 부분 기분석 사전

: 형태소 간의 문법 결합 강도를 조절하고, 부분 기분석 사전을 사용

(11) 지가공시및토지등의평가에대한법률
지가공시_{ncp} + 및_{ad} + 토지_{ncn} + 등_{nfix} + 의_{j}
+ 평가_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef} + 법률_{ncn}
성매매알선등행위의처벌에대한법률
성매매_{ncn} + 알선_{ncp} + 등_{nfix} + 행위_{ncn} + 의_{j}
+ 처벌_{ncp} + 에_{j} + 대하_{pv} + ㄴ_{ef} + 법률_{ncn}

# 4. 실험 및 고찰 (1)

? 상용검색엔진 사용 **keyword 2**만어절

표 **1** 정보 정밀화를 통한 형태소 분석 성능의 비교

| | avg # cand | prec | fail rate |
|---|---|---|---|
| Base | 2.976 | 0.859 | 0.068 |
| + 접미사 세분 | 1.676 | 0.862 | 0.111 |
| + 부분 기분석 사전 | 1.671 | 0.866 | 0.104 |
| + 다단계 해석 문법 | 1.654 | 0.914 | 0.047 |

- **avg # cand :** 입력 어절 당 출력하는 해석결과수
- **prec :** 출력한 해석 결과 중 정답을 포함하는 제대로 해석된 어절의 비
- **fail rate :** 분석 실패한 어절의 비
- **Base : 1**차원적인 품사 위주의 품사 정보를 이용한 기본 모델
- ¡**+** 접미사 세분¡ **:** 의미 정보를 활용한 형태소 품사 정보의 세분화를 통해 얻어낸 모델
- ¡**+** 부분 기분석 사전¡과 ¡**+** 다단계 해석 문법¡**:** 바로 윗줄의 모델에 **+** 다단계 해석 문법 부분 기분석 사전과 다단계 해석을 통한 모델의 확장

# 4. 실험 및 고찰 (2)

◆ 접미사 세분 적용

- 유사한 정확률을 보이면서, 평균 약 3개 정도의 후보를 제공하던 형태소 분석기가 평균 약 1.7개의 결과를 출력
- 세분화를 통해 형태소 연결 가능 여부 조건이 강화되어 해석 실패율이 증가

◆ 부분 기분석 사전과 다단계 해석 문법의 적용

- 정확률을 향상시키며 해석 실패율을 57% 가까이 감소시킴.

# 5. 결론

- 한국어 어절을 정확하게 형태소 분석을 하기 위해서는 한국어의 언어적 특성을 고려
- 형태소의 정보를 세밀화하고 각각의 정보들간의 언어적인 특성을 형태소 분석에 이용하면 분석의 정확률을 높이며 분석 실패한 어절의 수를 줄임.
- 정확한 형태소의 분석은 정보검색의 정보 제공에서의 정보의 적합성을 높임.
- 각각의 정밀화된 형태소의 언어적 정보를 이용한다면, 자연언어처리에 효율적인 방법론을 제시의 가능