Lemmatizer

Hyunjoong Kim

soy.lovit@gmail.com

github.com/{lovit, korean_lemmatizer}

한국어의 품사 체계

- 한국어의 단어는 5언 9품사로 구성되어 있습니다.
  - 조사는 체언 뒤에서 체언과 결합하여 하나의 어절을 이룹니다.
    - 명사 + 는
  - 조사는 문법 기능을 하며, 그 외의 단어들은 의미를 지닙니다.

<table>
<tr><th colspan="5">한국어 품사</th></tr>
<tr><td rowspan="4">불변어</td><td>체언</td><td>명사</td><td>대명사</td><td>수사</td></tr>
<tr><td>수식언</td><td colspan="2">관형사</td><td>부사</td></tr>
<tr><td>관계언</td><td colspan="3">조사</td></tr>
<tr><td>독립언</td><td colspan="3">감탄사</td></tr>
<tr><td>가변어</td><td>용언</td><td colspan="2">동사</td><td>형용사</td></tr>
</table>

단어와 형태소

- 형태소는 단어를 구성하는 최소 단위 입니다.
  - 명사, 부사, 조사 등은 그 자체가 형태소이기도 합니다.
  - 동사, 형용사는 “어간 + 어미” 로 구성되어 있습니다.
    - 이/Adjective + 쓰다/Eomi
  - 어간과 어미는 반드시 결합되어 단어를 이룹니다.

형태소 품사 체계

- 형태소의 품사 체계는 목적에 따라 다르게 이용할 수 있습니다.
- 세종 말뭉치는 조사나 어미의 종류도 다양하게 구분합니다.
  - 목적격 / 주격 / 서술격 조사 등
  - 연결 / 선어말 / 종결 어미 등
  - 했다 = 하/동사어간 + 았/선어말어미 + 다/어말어미

형태소 품사 체계

- 형태소의 품사 체계는 목적에 따라 다르게 이용할 수 있습니다.
- 어절의 구조를 L + [R] 로 가정하면 형태소의 종류를 줄여야 합니다.
  - 했다 = 하/동사어간 + 았/선어말어미 + 다/어말어미
    = 하/동사어간 + 았다/어미
  - 시작했다 = 시작/명사 + 하/동사형선어말어미 + 았/선어말어미 + 다/어말어미
    = 시작하/동사어간 + 았다/어미

용언의 활용

- 가변어는 어간에 결합되는 어미가 달라지면서 그 표현형이 바뀝니다.
  - 원형 (canonical form, lemma) 은 어간에 '-다/Eomi'가 결합된 형태입니다.
    - 하 + 다 / 가 + 다
  - 표현형 (surficial form) 은 그 외의 어미가 결합된 형태입니다.
    - 하 + ㄴ다 → 한다

한국어 품사				
불변어	체언	명사	대명사	수사
	수식언	관형사	부사	
	관계언	조사		
	독립언	감탄사		
가변어	용언	동사	형용사	

용언의 규칙 활용

- 용언의 활용은 두 종류로 나뉩니다.
  - 규칙 활용 / 불규칙 활용

- `규칙 활용` 은 자음/모음열 기준에서 변화가 없는 활용입니다.
  - 두 단어가 그대로 결합되는 경우와
    - 말하 + 다 → 말하다
  - 어간 마지막 글자의 종성이 없을 경우, 받침이 추가되는 경우 입니다.
    - 말하 + ㄴ다 → 말한다

용언의 불규칙 활용

- 불규칙 활용은 자음/모음열에 변화가 생깁니다.
  - 하 + 았다 → 했다
    - ㅎㅏㅇㅏㅆㄷㅏ → ㅎㅐㅆㄷㅏ

- 불규칙 활용도 몇 가지 경우로 분류됩니다.

용언의 불규칙 활용: ㄷ 불규칙

- 어간 마지막 종성이 `ㄷ`이고 어미 첫글자가 `ㅇ` 이면 `ㄷ → ㄹ`
  - 깨닫 + 아 → 깨달아
  - (질문을) 묻 + 었다 → 물었다
  - (물건을) 묻 + 었다 → 묻었다 (예외)

용언의 불규칙 활용: ㅂ 불규칙

- 어간 마지막 종성이 `ㅂ`이고 어미 첫글자가 `ㅇ` 이면 `ㅂ → ㅗ/ㅜ`
  - 더럽 + 어 → 더러워
  - 곱 + 아 → 고와
  - 아름답 + 아 → 아름다워

용언의 불규칙 활용

- 그 외에도 다양한 불규칙 활용의 규칙이 있습니다.
  - 정리된 규칙은 블로그 [1,2] 와 웹 페이지를 참고하세요 [3]

[1] https://lovit.github.io/nlp/2018/06/07/lemmatizer/
[2] https://lovit.github.io/nlp/2018/06/11/conjugator/
[3] https://namu.wiki/w/한국어/불규칙%20활용/

활용과 원형 복원

- 용언의 원형을 표현형을 변환하는 과정을 활용 (conjugation) 이라 하며, 그 반대 과정을 원형 복원 (lemmatization) 이라 합니다.
  - conjugate('하', 'ㄴ다') → '한다'
  - lemmatization('한다') → '하 + ㄴ다'

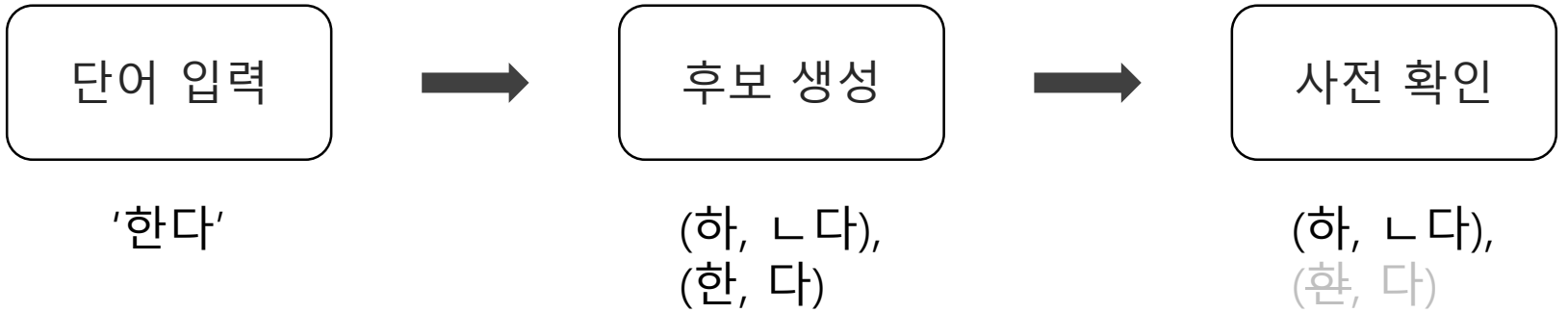
규칙 기반 원형 복원 (lemmatization)

- 원형 복원의 과정은 규칙 기반으로 구현할 수 있습니다.

```
l_last = decompose(l[-1])
r_first = decompose(r[0]) if r else ('', '', '')
if l_last[2] == 'ㄹ' and r_first[0] == 'ㅇ':
    l_stem = l[:-1] + compose(l_last[0], l_last[1], 'ㄷ')
    candidates.add((l_stem, r))
```

규칙 기반 원형 복원 (lemmatization)

- 규칙 기반으로 형태소 후보를 만든 뒤, 형태소 사전을 검색합니다.



규칙 기반 원형 복원 (lemmatization)

- 규칙 기반 방법은 새로운 어미를 탐색할 때 효과적입니다.
  - 말투에 의하여 새로운 어미는 만들어지지만, 새로운 어간은 거의 만들어지지 않습니다.
  - 새로운 어미도 활용 규칙을 따릅니다.
  - 했습니다 = (하, 았습니다) // 했어염 = (하, 았어염)

음절 단위의 사전 기반 원형 복원

- 그러나 매 단어마다 규칙 기반으로 후보를 탐색할 수 없습니다.
  - String decomposition 은 큰 계산 비용이 듭니다.
  - 게다가 규칙이기 때문에 경우의 수가 한정적입니다.

음절 단위의 사전 기반 원형 복원

- 심광섭 (2013) 에서는 음절 단위로 형태가 변하는 지점을 사전으로 저장하여 이용합니다.
  - 란 → {(랗, ㄴ), (라, ㄴ)}
    - 노란 = 노랗 + ㄴ
    - 놀란 = 놀라 + ㄴ
  - 했 → {(하, 았)}
    - 시작했다 → 시작하 + 았다

심광섭. (2013). 음절 단위의 한국어 품사 태깅에서 원형 복원. *정보과학회논문지: 소프트웨어 및 응용*, *40*(3), 182-189.

음절 단위의 사전 기반 원형 복원

- 표현형의 모습이 변하는 지점은 어간과 어미가 만나는 부분입니다.
  - 란 → {(랗, ㄴ), (라, ㄴ)}
    - 노란 = 노랗 + ㄴ
    - 놀란 = 놀라 + ㄴ
  - 했 → {(하, 았)}
    - 시작했다 → 시작하 + 았다

심광섭. (2013). 음절 단위의 한국어 품사 태깅에서 원형 복원. *정보과학회논문지: 소프트웨어 및 응용*, *40*(3), 182-189.

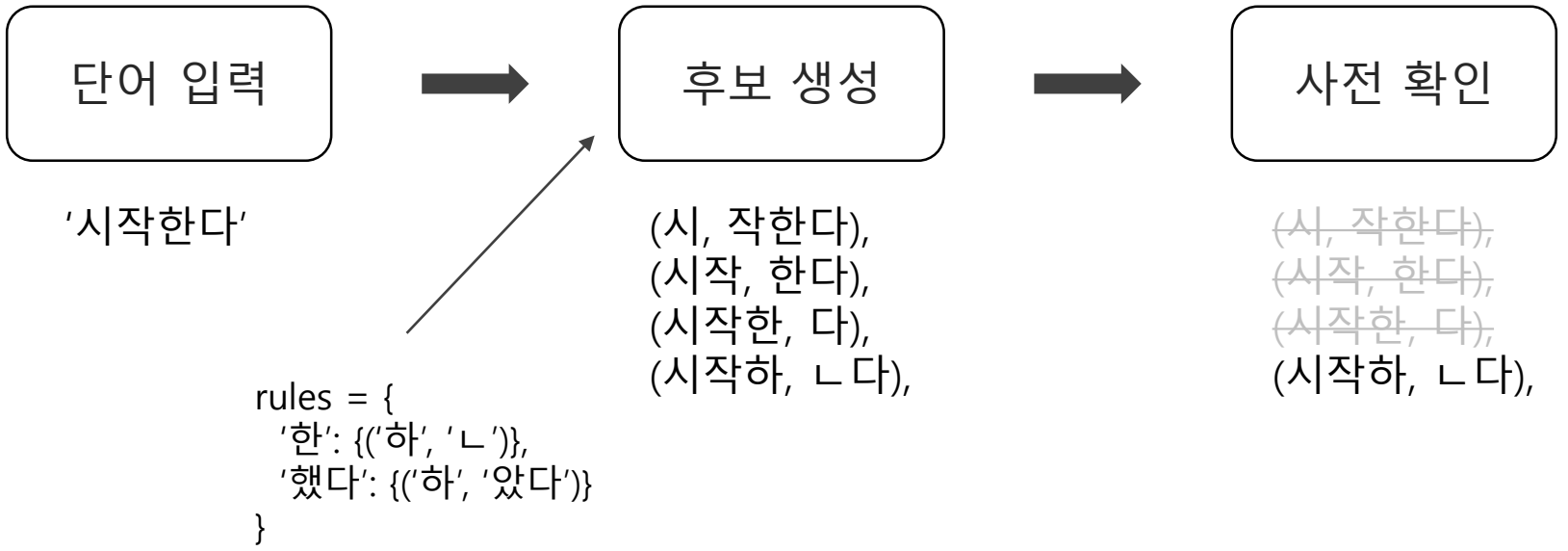
음절 단위의 사전 기반 원형 복원

- 형태소 분석 말뭉치를 이용하여 규칙을 저장합니다.
- 형태 변화는 1 음절보다 긴 부분에서 일어날 수 있습니다.

형태 변화 음절 길이	형태 변화 규칙	단어 예시
1 음절	했 = 하 + 았	시작했으니까 = 시작하 + 았으니까
1 음절	랬 = 랗 + 았	파랬던 = 파랗 + 았던
2 음절	추운 = 춥 + 은	추운데 = 춥 + 은데
2 음절	했다 = 하 + 았다	시작했다 = 시작하 + 았다
3 음절	가우니 = 갑 + 니	차가우니까 = 차갑 + 니까

음절 단위의 사전 기반 원형 복원

- 단어의 모든 부분에 대하여 형태소 후보를 생성한 뒤, 사전을 확인합니다



- 학습된 용언 분석기 (lemmatizer) 는 단어, 규칙 사전 관리가 필수입니다.
- 학습 말뭉치에 존재하지 않은 표현형도 인식 가능합니다.