구문 분석을 위한 한국어 말덩이 정의

남궁영, 김창현†, 천민아, 박호민, 윤호, 최민석, 김재훈

한국해양대학교 컴퓨터공학과, 한국전자통신연구원†

young_ng@kmou.ac.kr

2018. 10. 13.

목 차

I 말덩이란 (1/4)

❖ 한국어 처리

… → 형태소 분석 → 구문 분석 → 의미 분석 → …

- 문장 구성 요소들 간의 이동 및 생략이 자유로움.
- 구문 분석 단계에서 처리해야 할 성분의 수가 많음.

구문 분석의 **중의성** 문제 유발

⇒ 구문적으로 **하나의 역할을 수행**하는 형태소들을 묶어 구를 형성한 뒤
이를 이용해 구문 분석을 수행할 수 있다.

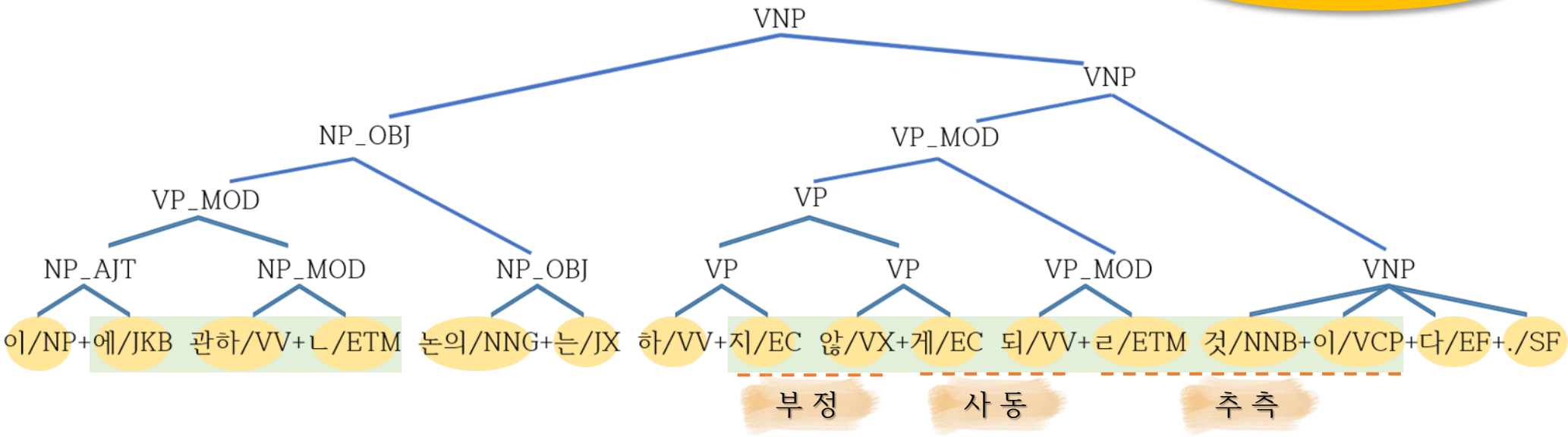
예] 할 수 있다 → 하/VV+ㄹ/ETM 수/NNB 있/VV+다/EF → 하/VV+(**PUX** ㄹ/ETM 수/NNB 있/VA)+다/EF

I 말덩이란 (2/4)

이에 관한 논의는 하지 않게 될 것이다.

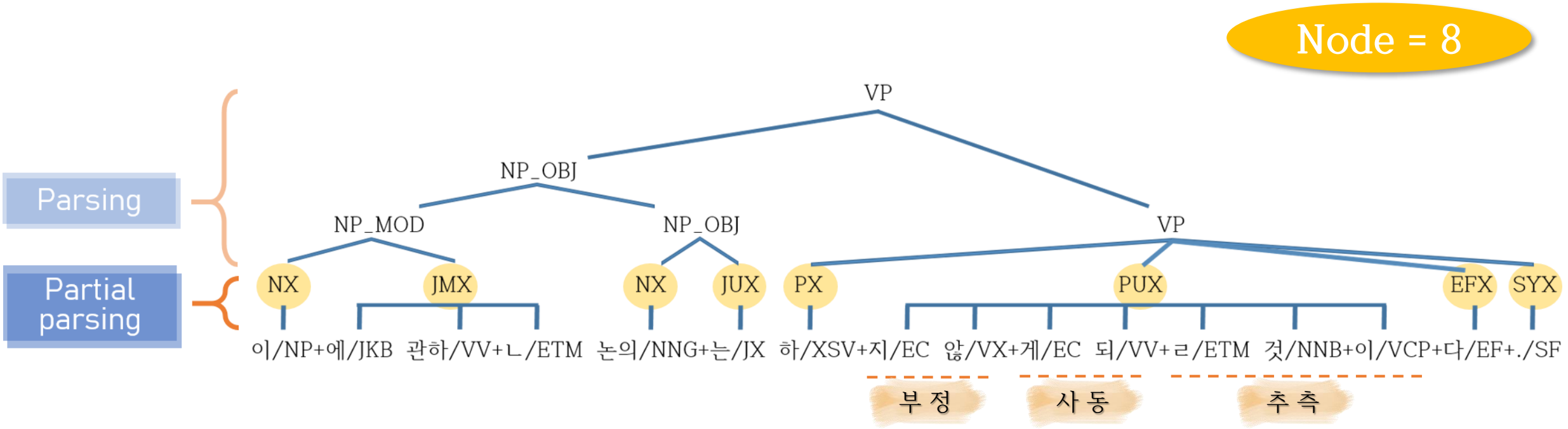
Node = 16

Parsing



I 말덩이란 (3/4)

이에 관한 논의는 하지 않게 될 것이다.



입력 성분 수를 줄여 계산의 복잡도를 감소시키고 구문 분석의 **정확도를 높일 수 있다**.

I 말덩이란 (4/4)

❖ 한국어 처리

… → 형태소 분석 → 구문 분석 → 의미 분석 → …

부분 구문 분석

- **구묶음(chunking)** : 구문 분석 단계 이전에 문장 내의 형태소들을 하나의 말덩이로 묶어 구문 분석의 부담을 경감시키는 과정
- **부분 구문 분석(partial parsing)** : 입력 문장에 대해 구묶음을 수행하는 것
- **말덩이(chunk)** : 부분 구문 분석의 기본 단위

⇒ 한국어 문장의 모든 구성 성분에 대해 구문 분석을 위한 단위로서의 **말덩이(chunk)의 기준 및 정의 제시**

❖ 품사 태그: 세종 품사 태그, 예문: 세종 말뭉치 등

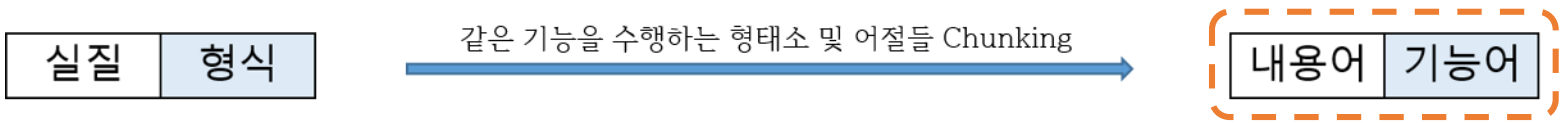
Ⅱ 말덩이의 특징 (1/3)

❖ 개 요

- The typical **chunk** consists of a single **content word** surrounded by a constellation of **function words**, matching a fixed template.*

- 인간이 한번에 받아들이는 언어의 구조(performance structures)가 있으며,
  이는 자연스럽게 한번에 발화되는 단위 또는 내용어(syntatic head, content word)를 기준으로 분절되는 말덩이(chunk) 등으로 표상된다.**

- 한국어의 형태상 갈래: 교착어(첨가어)



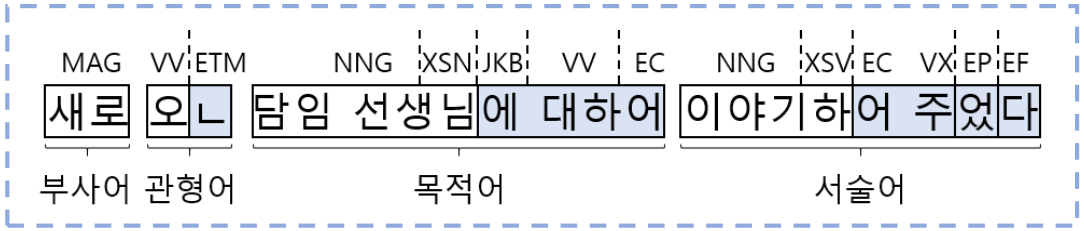
- 기존의 구문 분석 방법에서도 그대로 이용 가능
- Head-final 방식으로도 구문 분석 가능

* S. Abney, "Parsing by chunks", Principle-based parsing, eds. Berwick, R. Abney, S. and Tenny, C., Kluwer Academic Publishers. 1991.
** J. P. Gee and F. Grosjean, "Performance structures: A psycholiguistic and linguistic appraisal", Cognitive Psychology vol.15, no.4, pp. 411-458. 1983.

Ⅱ 말덩이의 특징 (2/3)

❖ 말덩이를 단위로 구묶음을 수행한 문장은 다음과 같은 **특징**을 가진다.

- 완전한 구묶음 이후의 문장은 한국어의 **7가지 구성성분으로 표현 가능**하다.
  (주어, 서술어, 목적어, 보어, 관형어, 부사어, 독립어)



- **서술어를 제외한 문장 성분**

내용어	또는	내용어	기능어

- **서술어**

내용어	기능어	...	기능어

ㄴ하나 이상의 보조용언, 연결어미, 선어말어미, 종결어미 등의 말덩이

- **내용어 말덩이**

  반드시 **의미적 중심어**(semantic head)를 가진다.

- **기능어 말덩이**

  의미적 중심어를 가지지 않고 그 자체로 **하나의 덩어리**를 형성한다.

Ⅱ 말덩이의 특징 (3/3)

- **병렬구문**은 의미적 중심어가 여러 개 존재하므로 각각의 구절을 하나의 말덩이로 간주한다.
- 동일한 구가 연속해서 등장할 경우 **최장일치**를 기본으로 한다.

- 말덩이를 이루는 구성 성분들은 **연속적**이어야 한다(no discontinuous).
- 문장 내의 하나의 형태소는 반드시 하나의 말덩이에 속하며,

  서로 다른 말덩이에 중복하여 구묶음 되지 않는다. 즉, **비중첩성**을 지닌다(no center-embedded).
- 말덩이 내의 구문 구조는 선형으로, 트리 구조를 형성하지 않고 **비재귀성**을 지닌다(no recursive).

Ⅲ 단위 및 표지

내용어 말덩이

1. 체언구 (NX)
2. 본용언구 (PX)
3. 지정사구 (CX)
4. 부사구 (AX)
5. 관형사구 (MX)
6. 독립어구 (IX)

기능어 말덩이

7. 보조용언구 (PUX)
8. 조사
    1) 격조사구
        ① 주격/목적격/보격/부사격 (JKX)
        ② 관형격 (JMX)
        ③ 호격 (JVX)
    2) 보조사구 (JUX)
    3) 접속조사구 (JCX)
9. 어미
    1) 선어말 어미 (EPX)
    2) 연결 어미 (ECX)
    3) 전성 어미 (ETX)
        ① 관형형
        ② 명사형
    4) 종결 어미 (EFX)
10. 문장 부호 (SYX)

Ⅲ 단위 및 표지

1. 체언구 말덩이 (NX)

- 주어, 목적어, 보어의 내용어가 된다.

1) 단일 명사, 대명사, 수사 (체언구 표준 말덩이)

- (NX 학교/NNG)에 갔다.

2) 명사열

- (NX 전철역/NNG 주변/NNG)에서 안내지를 보았다.

3) 명사(열) + 접사

① 명사(열) + XSN

- (NX 아이/NNG+들/XSN)과 공원에서 놀았다.

② XPN + 명사(열)

- 그는 (NX 풋/XPN+사과/NNG)를 좋아한다.

❖ **표준 말덩이** : 한 형태소가 다른 성분들과 함께 말덩이를 이루지 않고 **자기 자신이 하나의 말덩이를 형성**하는 경우 이를 표준 말덩이라 한다.

Ⅲ 단위 및 표지

2. 용언구 말덩이 (PX)

- 서술어, 관형어의 내용어가 된다.

1) 단일 동사 또는 형용사 (용언구 표준 말덩이)

- 사회가 (PX 겪/VV)는 사건들

2) (XPN) + XR + (XSV / XSA)

- 지방 자치제가 철저히 (PX 시행/XR+되/XSV)고 있다.

3) 동사열

- (PX 조사/XR 및/MAG 관찰/XR+하/XSV)였다.

Ⅲ 단위 및 표지

3. 지정사구 말덩이 (CX)

- 지정사구 말덩이 또한 체언과 용언의 성질을 모두 지니고 있다.
- 서술어에서 내용어로서 기능한다.

1) 체언 + VCP

- 바로 그 (CX 책/NNG+이/VCP)었다.

Ⅲ 단위 및 표지

4. 부사구 말덩이 (AX)

- 부사어의 내용어가 된다.

1) 단일 부사 (부사구 표준 말덩이)

- (AX 빨리/MAG) 달린다.

2) 같은 부사가 반복해서 나올 경우

- (AX 너무/MAG 너무/MAG) 좋았다.

Ⅲ 단위 및 표지

5. 관형사구 말덩이 (MX)

- 체언구를 수식하며 관형어의 내용어가 될 수 있다.

1) 단일 관형사 (관형사구 표준 말덩이)

- (MX 저/MM) 예쁜 꽃

2) 같은 관형사가 반복해서 나올 경우

- (MX 무슨/MM 무슨/MM) 학교라던데.

3) 체언 바로 앞에 여러 관형사가 올 경우

- (MX 저/MM) (MX 헌/MM) 책이 내 거야.

Ⅲ 단위 및 표지

6. 독립어구 말덩이 (IX)

- 독립어의 내용어가 된다.
- ‘체언 + JKV’로 이루어진 독립어를 제외한 경우가 이에 해당한다.

1) 단일 독립어구 (독립어구 표준 말덩이)

- (IX 아/IC), 벌써 가을이 왔나보다.

Ⅲ 단위 및 표지

7. 보조용언구 말덩이 (PUX)

- 보조 용언구: 본용언 없이 홀로 존재할 수 없으며, 그 뜻 또한 본용언을 보조하는 의미를 지닌다.

1) 보조 용언을 포함하는 경우

- 꿈이 이루어지(PUX 도록/EC 하/VX)자.

2) 의존 명사를 포함하는 경우*

- 같은 꿈을 꾸(PUX 는/ETM 셈/NNB 이/VCP)다.

3) 여러 보조용언구가 연속해서 나올 경우 **최장 일치를 기본으로** 한다.

- 학생이 배우(PUX ㄹ/ETM 수/NNB 있/VV+도록/EC 하/VX)어야 이상적이다.

* 황이규, 이현영, 이용석, “형태소 및 구문 모호성 축소를 위한 구문단위 형태소의 이용”, 한국정보과학회논문지:소프트웨어 및 응용, vol.27, no.7, pp.784-793, 2000.

Ⅲ 단위 및 표지

8. 조사 말덩이

▪ 격조사구 말덩이

- 주격/목적격/보격/부사격 (JKX)

  1) **단일 격조사 (격조사구 표준 말덩이)**

     - 나의 노력(JKX 을/JKO) 기억해주세요.

  2) **격조사 + JX**

     - 나(JKX 에게/JKB+는/JX) 믿음이 있다.

  3) **격조사 상당 어구**

     - 담임 선생님(JKX 에/JKB 대하/VV+어/EC)이야기했다.

- 관형격 (JMX)

  1) **단일 관형격조사 (관형격조사구 표준 말덩이)**

     - 나(JMX 의/JKG) 집

  2) **관형격조사 상당 어구**

     - 다른 사람(JMX 을/JKO 위하/VV+ㄴ/ETM) 배려

- 호격 (JVX)

  1) **단일 호격조사 (호격조사구 표준 말덩이)**

     - 대자대비하신 부처님(JVX 이시여/JKV)

Ⅲ 단위 및 표지

8. 조사 말덩이

- 보조사구 말덩이 (JUX)

  1) 단일 보조사 (보조사구 표준 말덩이)

  - 이것 (JUX 부터/JX) 시작하자.

  2) 보조사 상당어구

  - 이것 (JUX 만/JX+도/JX) 못하다.

- 접속조사구 말덩이 (JCX)

  1) 단일 접속조사 (접속조사구 표준 말덩이)

  - 철수(JCX 와/JC) 영희는 서로 친구이다.

  2) 접속조사 상당 어구

  - 영희(JCK 뿐/JX+만/JX 아니/VA+라/EC) 철수도…

Ⅲ 단위 및 표지

9. 어미 말덩이

- 선어말 어미 말덩이 (EPX)

  1) **단일 선어말 어미 (선어말 어미 표준 말덩이)**

  - 전력으로 달리(**EPX** 었/EP)다.

  2) **복합 선어말 어미구**

  - 진지 잡수(**EPX** 시/EP+었/EP)어요.

- 연결 어미 말덩이 (ECX)

  1) **단일 연결 어미 (연결 어미 표준 말덩이)**

  - 그녀의 눈은 크(**ECX** 고/EC) 아름답다.

  2) **복합 연결 어미구**

  - 그녀의 눈은 크(**ECX** ㄹ/ETM 뿐/NNB+만/JX 아니/VA+라/EC) 아름답다.

Ⅲ 단위 및 표지

9. 어미 말덩이

- **전성 어미 말덩이**
  - 관형형 전성 어미 말덩이 **(EMX)**

    - 내가 살(**EMX** 던/ETM) 집
  - 명사형 전성 어미 말덩이 **(ENX)**

    - 먹고살(**ENX** 기/ETN+에/JKB) 충분하다.
  - 부사형 전성 어미 말덩이 **(EAX)**

    - 지금만큼은 어른답(EAX 게/EC) 행동해야 한다.

- **종결어미 말덩이 (EFX)**

  1) **단일 종결 어미 (종결 어미 표준 말덩이)**

     - 진행하고 있(**EFX** 습니다/EF).

10. 문장 부호 말덩이 (SYX)

- 사람이(**SYX** …/SE+…/SE) 많이 왔군요(**SYX** !/SF)

Ⅳ 결론 및 제언

❖ 한국어 부분 구문 분석 말뭉치 구축을 위한 **말덩이**의 특성과 종류 제시

❖ 부분 구문 분석의 결과를 이용해 구문 분석을 수행 時

입력 노드 수↓ → 계산량↓ → 정확도 향상

∴ 구문 분석의 **중의성 문제 해소**에 기여

❖ Chunk corpus 제작 중 → 4만 문장 공개 예정

감 사 합 니 다

남궁영, 김창현†, 천민아, 박호민, 윤호, 최민석, 김재훈

한국해양대학교 컴퓨터공학과, 한국전자통신연구원†

young_ng@kmou.ac.kr

2018. 10. 13.