구묶음을 반영한 한국어 의존 구조 말뭉치 생성

남궁영*, 김창현, 천민아, 박호민, 윤호, 최민석, 김재균, 김재훈
한국해양대학교 컴퓨터공학과
young_ng@kmou.ac.kr*

2019. 10. 12.

국립 한국해양대학교
KOREA MARITIME AND OCEAN UNIVERSITY

목 차

Ⅰ. 의존구문분석과 구묶음

ⅰ. 의존구문분석

ⅱ. 의존구문분석의 문제점

ⅲ. 구묶음

ⅳ. 구묶음을 반영한 의존구문분석

I 의존구문분석과 구묶음

❖ 의존구문분석 (dependency parsing)

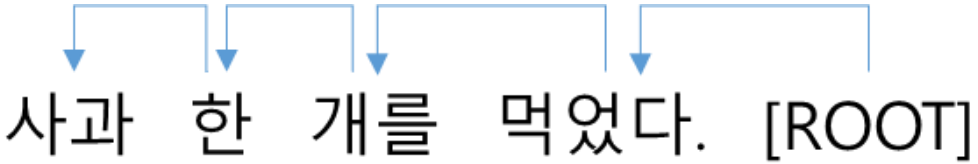
- 구문분석
  - 문장 구성성분들의 관계를 파악하는 과정
  - 문장의 구조 결정 → 의미적 중의성 해소

- 의존구문분석
  - 문장 성분 간의 지배소-의존소 관계를 파악함으로써 문장의 구조를 분석
  - 구성 요소의 위치 이동 및 생략에 유연하게 대처 가능

I 의존구문분석과 구묶음

❖ 의존구문분석의 문제점

- 지배소 결정 문제 (구문적 중심어 ≠ 의미적 중심어)



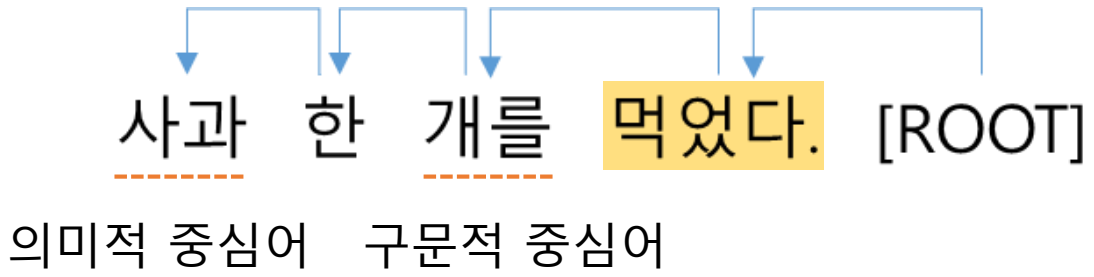
- 의존 관계를 결정해야 할 노드 수가 많음

잘 할 수 있다. [ROOT]

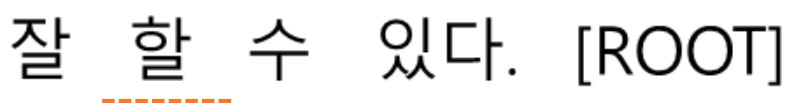
I 의존구문분석과 구묶음

❖ 의존구문분석의 문제점

- 지배소 결정 문제 (구문적 중심어 ≠ 의미적 중심어)



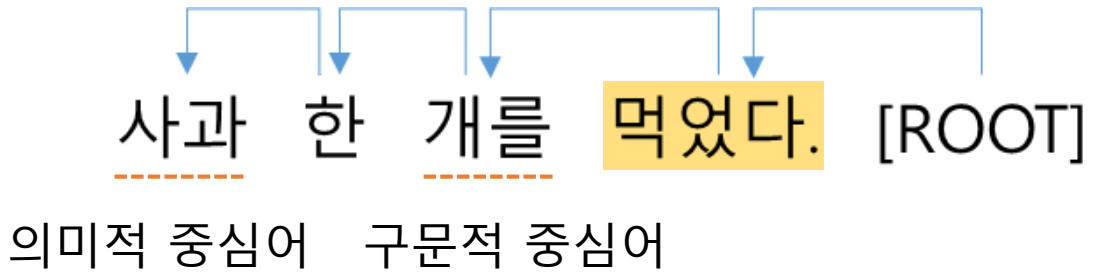
- 의존 관계를 결정해야 할 노드 수가 많음



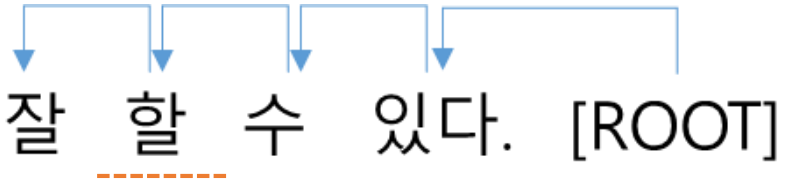
I 의존구문분석과 구묶음

❖ 의존구문분석의 문제점

- 지배소 결정 문제 (구문적 중심어 ≠ 의미적 중심어)



- 의존 관계를 결정해야 할 노드 수가 많음



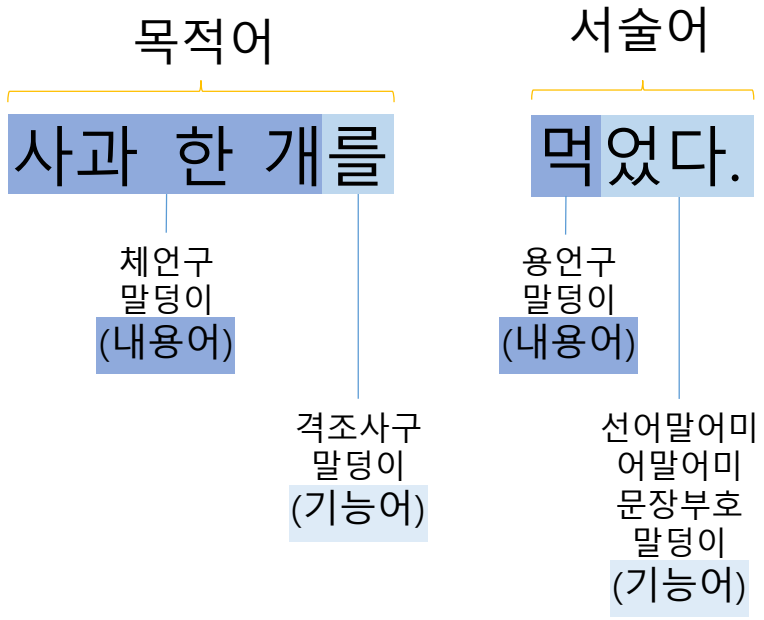
I 의존구문분석과 구묶음

❖ 구묶음 (chunking)[1,2]

- 정의
  - 형태소들을 하나의 의미 있는 구성 성분인 말덩이(chunk)로 묶는 작업
  - 부분구문분석 → 구문분석

- 말덩이 (chunk)
  - 인간이 한번에 받아 들이는 언어의 단위
  - 문법적, 의미적으로 하나의 기능을 수행
  - 연속성, 비중첩성, 비재귀성

- 문장 성분
  - 주어, 서술어, 목적어, 보어, 관형어, 부사어, 독립어
  - **내용어** **기능어**

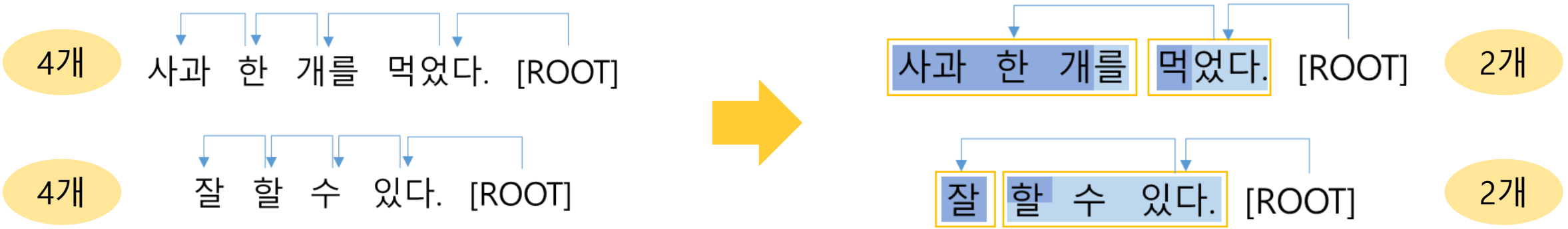


[1] S. Abney, "Parsing by chunks", Principle-based parsing, eds. Berwick, R. Abney, S. and Tenny, C., Kluwer Academic Publishers. 1991.
[2] 박의규, 나동열, "한국어 구문분석을 위한 구묶음 기반 의존명사 처리", 인지과학, vol.17, no.2, pp. 118-138, 2006.

I 의존구문분석과 구묶음

❖ 구묶음을 반영한 한국어 의존구문분석

- 문장 성분 단위가 하나의 노드 → 구문 분석의 속도 및 정확도 향상

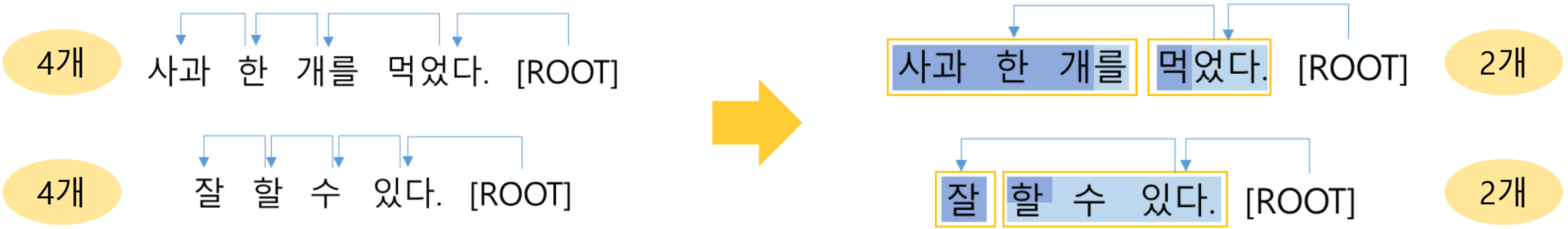


- 구문적 중심어 ≃ 의미적 중심어
  - 지배소 후위 원칙(head-final) 유지 & 의미 분석(semantic analysis) 수행 가능

I 의존구문분석과 구묶음

❖ 구묶음을 반영한 한국어 의존구문분석

- 문장 성분 단위가 하나의 노드 → 구문 분석의 속도 및 정확도 향상



- 구문적 중심어 ≃ 의미적 중심어
  - 지배소 후위 원칙(head-final) 유지 & 의미 분석(semantic analysis) 수행 가능

∴ **구묶음을 반영한 의존구문 말뭉치** → 효과적인 구문분석 가능

구묶음을 반영한 한국어 의존구문 말뭉치

i. 말뭉치 생성 방안

ii. 기존의 의존구문 말뭉치

iii. 구묶음을 반영한 의존구문 말뭉치

Ⅱ 구묶음을 반영한 의존구문 말뭉치 구축

❖ 말뭉치 생성 방안

- 말뭉치 구축: 시간, 비용, 인적 자원 ↑

- 기존 의존구문 말뭉치[1] → 구묶음을 반영한 의존구문 말뭉치
  - CoNLL 형식

∴ **구묶음을 반영한 의존구문 말뭉치**로 변환하는 알고리즘 기술

1) 최용성, 이공주, **"한국어 구절 구문 코퍼스의 의존 구문 구조 트리로의 변환에서 중심어 전파 규칙"**, 제30회 한글 및 한국어 정보처리 학술대회 논문집, pp. 514 519, 2018.

Ⅱ 구묶음을 반영한 의존구문 말뭉치 구축

❖ 기존의 의존구문 말뭉치

- 토큰 단위 (**형태소** 기반)

#ORGSENT:프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	프랑스의	프랑스 의	PROPN	NNP+JKG	_	4	nmod	_	_
2	세계적인	세계 적 이 ㄴ	ADJ	NNG+XSN+VCP+ETM	_	4	acl	_	_
3	의상	의상	NOUN	NNG	_	4	nmod	_	_
4	디자이너	디자이너	NOUN	NNG	_	6	nmod	_	_
5	엠마누엘	엠마누엘	PROPN	NNP	_	6	nmod	_	_
6	웅가로가	웅가로 가	PROPN	NNP+JKS	_	11	nsubj	_	_
7	실내	실내	NOUN	NNG	_	8	nmod	_	_
8	장식용	장식 용	NOUN	NNG+XSN	_	9	nmod	_	_
9	직물	직물	NOUN	NNG	_	10	nmod	_	_
10	디자이너로	디자이너 로	NOUN	NNG+JKB	_	11	obl	_	_
11	나섰다.	나서 었 다 .	VERB	VV+EP+EF+SF	_	0	root	_	_

Ⅱ 구묶음을 반영한 의존구문 말뭉치 구축

❖ 구묶음을 반영한 의존구문 말뭉치

- 문장 성분 단위 (**말덩이** 기반)
- 내용어(contents) + 기능어(function)

{ 구묶음을 반영한 의존 구문 말뭉치 }

```
# text = 프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.
```

ID	FORM(conts)	FORM(func)	LEMMA	UPOSTAG	XPOSTAG	CHUNKTAG	HEADS	DEPREL
1	프랑스	의	프랑스 의	PROPN	NNP+JKG	NX+JMX	3	nmod
2	세계_적_이	ㄴ	세계 적 이 ㄴ	ADJ	NNG+XSN+VCP+ETM	CX+ETX	3	acl
3	의상_디자이너_엠마누엘_웅가로	가	의상 디자이너 엠마누엘 웅가로 가	PROPN	NNG+NNG+NNP+NNP+JKS	NX+JKX	5	nsubj
4	실내_장식_용_직물_디자이너	로	실내 장식 용 직물 디자이너 로	NOUN	NNG+NNG+XSN+NNG+NNG+JKB	NX+JKX	5	obl
5	나서	었_다_.	나서 었 다 .	VERB	VV+EP+EF+SF	PX+EPX+EFX+SYX	0	root

{ 기존 의존 구문 말뭉치 }

```
#ORGSENT:프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.
```

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	프랑스의	프랑스 의	PROPN	NNP+JKG	_	4	nmod	_	_
2	세계적인	세계 적 이 ㄴ	ADJ	NNG+XSN+VCP+ETM	_	4	acl	_	_
3	의상	의상	NOUN	NNG	_	4	nmod	_	_
4	디자이너	디자이너	NOUN	NNG	_	6	nmod	_	_
5	엠마누엘	엠마누엘	PROPN	NNP	_	6	nmod	_	_
6	웅가로가	웅가로 가	PROPN	NNP+JKS	_	11	nsubj	_	_
7	실내	실내	NOUN	NNG	_	8	nmod	_	_
8	장식용	장식 용	NOUN	NNG+XSN	_	9	nmod	_	_

변환 과정 및 알고리즘

ⅰ. 변환 과정

ⅱ. 변환 알고리즘

Ⅲ 변환 과정 및 알고리즘

❖ 실제 변환에 사용되는 요소들을 중심으로 과정 설명

{ 기존 의존 구문 말뭉치 형식 }

ID	FORM	XPOSTAG	HEAD
1	프랑스의	NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

{ 구묶음을 반영한 의존 구문 말뭉치 형식 }

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	content	function			

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

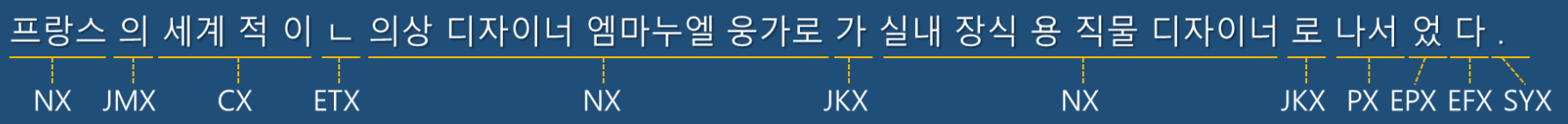
문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

- 문장 성분 단위를 얻기 위해 구묶음 수행
  - 입력: 형태소/품사
- CHUNKTAG 열 생성
- 예시
  - 프랑스의 세계적인 의상 디자이너 엠마누엘 웅가로가 실내 장식용 직물 디자이너로 나섰다.



Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

프랑스 의 세계 적 이 ㄴ 의상 디자이너 엠마누엘 웅가로 가 실내 장식 용 직물 디자이너 로 나서 었 다 .

NX JMX CX ETX NX JKX NX JKX PX EPX EFX SYX

{ 기존 의존 구문 말뭉치 형식 }

ID	FORM	XPOSTAG	HEAD
1	프랑스의	NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

{ 구묶음을 반영한 의존 구문 말뭉치 형식 }

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			

Ⅲ 변환 과정 및 알고리즘

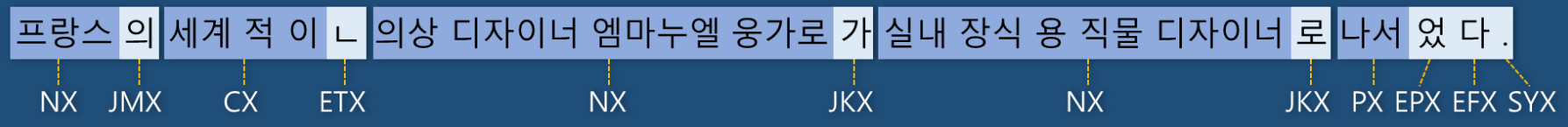
말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정



{ 기존 의존 구문 말뭉치 형식 }

ID	FORM	XPOSTAG	HEAD
1	프랑스의	NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

{ 구묶음을 반영한 의존 구문 말뭉치 형식 }

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

프랑스 의 세계 적 이 ㄴ 의상 디자이너 엠마누엘 웅가로 가 실내 장식 용 직물 디자이너 로 나서 었 다 .

NX JMX CX ETX NX JKX NX JKX PX EPX EFX SYX

{ 기존 의존 구문 말뭉치 형식 }

ID	FORM	XPOSTAG	HEAD
1	프랑스의	NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

{ 구묶음을 반영한 의존 구문 말뭉치 형식 }

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			

Ⅲ 변환 과정 및 알고리즘

- 말덩이 표지 생성
- 문장 성분 구성
- 새로운 지배소 결정에 사용할 사전 생성
- 문장 성분의 중심어 결정
- 문장 성분의 지배소 및 관계명 결정

{ ID 사전 }

new_ID (문장성분)	old_ID (토큰)
1	1
2	2
3	3, 4, 5, 6
4	7, 8, 9, 10
5	11

{ 역 ID 사전 }

old_ID (토큰)	new_ID (문장성분)
1	1
2	2
3	3
4	3
5	3
6	3
7	4
8	4
9	4
10	4
11	5

- 문장 성분의 중심어에 해당하는 토큰의 HEAD 정보를 새로운 ID로 매핑할 수 있음

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

- 문장 성분의 중심어 결정 → 문장 성분의 최종 HEAD 결정

old_ID (토큰)	new_ID (문장성분)
5	3
6	3
7	4
8	4
9	4
10	4
11	5

		XPOSTAG	HEAD
		NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			
1	프랑스	의	NNP +JKG	NX+JMX	
2	세계 적 이	ㄴ	NNG+XSN+VCP +ETM	CX+ETX	
3	의상 디자이너 엠마누엘 웅가로	가	NNG+NNG +NNP+NNP +JKS	NX+JKX	
4	실내 장식 용 직물 디자이너	로	NNG+NNG+XSN +NNG+NNG +JKB	NX+JKX	
5	나서	었 다 .	VV+EP +EF+SF	PX+EPX +EFX+SYX	

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

- 문장 성분의 중심어 결정 → 문장 성분의 최종 HEAD 결정

문장 성분을 이루고 있는 내용어 토큰 중 그 HEAD가 문장 성분 내에 없는 것

old_ID (토큰)	new_ID (문장성분)
5	3
6	3
7	4
8	4
9	4
10	4
11	5

		XPOSTAG	HEAD
		NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			
1	프랑스	의	NNP +JKG	NX+JMX	
2	세계 적 이	ㄴ	NNG+XSN+VCP +ETM	CX+ETX	
3	의상 디자이너 엠마누엘 웅가로	가	NNG+NNG +NNP+NNP +JKS	NX+JKX	
4	실내 장식 용 직물 디자이너	로	NNG+NNG+XSN +NNG+NNG +JKB	NX+JKX	
5	나서	었 다 .	VV+EP +EF+SF	PX+EPX +EFX+SYX	

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

old_ID (토큰)	new_ID (문장성분)
5	3
6	3
7	4
8	4
9	4
10	4
11	5

- 역 ID 사전을 참조하여 변환한 말뭉치의 지배소(HEAD) 결정
- 의존 관계명 등은 중심어가 갖고 있는 정보를 따름

		XPOSTAG	HEAD
		NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			
1	프랑스	의	NNP +JKG	NX+JMX	
2	세계 적 이	ㄴ	NNG+XSN+VCP +ETM	CX+ETX	
3	의상 디자이너 엠마누엘 웅가로	가	NNG+NNG +NNP+NNP +JKS	NX+JKX	
4	실내 장식 용 직물 디자이너	로	NNG+NNG+XSN +NNG+NNG +JKB	NX+JKX	
5	나서	었 다 .	VV+EP +EF+SF	PX+EPX +EFX+SYX	

Ⅲ 변환 과정 및 알고리즘

말덩이 표지 생성

문장 성분 구성

새로운 지배소 결정에 사용할 사전 생성

문장 성분의 중심어 결정

문장 성분의 지배소 및 관계명 결정

old_ID (토큰)	new_ID (문장성분)
5	3
6	3
7	4
8	4
9	4
10	4
11	5

- 역 ID 사전을 참조하여 변환한 말뭉치의 지배소(HEAD) 결정
- 의존 관계명 등은 중심어가 갖고 있는 정보를 따름

		XPOSTAG	HEAD
		NNP+JKG	4
2	세계적인	NNG+XSN+VCP+ETM	4
3	의상	NNG	4
4	디자이너	NNG	6
5	엠마누엘	NNP	6
6	웅가로가	NNP+JKS	11
7	실내	NNG	8
8	장식용	NNG+XSN	9
9	직물	NNG	10
10	디자이너로	NNG+JKB	11
11	나섰다.	VV+EP+EF+SF	0

ID	FORM		XPOS TAG	CHUNK TAG	HEAD
	cont	func			
1	프랑스	의	NNP +JKG	NX+JMX	3
2	세계 적 이	ㄴ	NNG+XSN+VCP +ETM	CX+ETX	3
3	의상 디자이너 엠마누엘 웅가로	가	NNG+NNG +NNP+NNP +JKS	NX+JKX	5
4	실내 장식 용 직물 디자이너	로	NNG+NNG+XSN +NNG+NNG +JKB	NX+JKX	5
5	나서	었 다 .	VV+EP +EF+SF	PX+EPX +EFX+SYX	0

Ⅲ 변환 과정 및 알고리즘

❖ 변환 알고리즘

```
def To_Chunk_Dependency_Corpus(dependency_corpus):
    # 문장 성분 단위로 분리
    toConst = To_Constituent(dependency_corpus)

    # look-up 사전 생성
    idDict = ID_Dictionary(toConst) # {new_ID: old_ID}
    idDict_reversed = ID_Dictionary_Reversed(idDict)
                                    # {old_ID: new_ID}

    # 문장 성분 내의 중심어 선정
    for old_id in idDict(new_ID):
            # 토큰의 head에 해당하는 id가
            # 문장 성분 내에 없으면 이 토큰을
            # 해당 문장 성분의 중심어(content)로 선정
            if not old_id.HEAD in idDict(new_ID):
                    content_list = Add_to_Content_List(old_id)

    # 선정한 중심어를 토대로 역 ID 사전을 이용하여
    # 문장 성분의 최종 지배소 및 관계명 결정
    for old_id in content_list:
            new_head = idDict_reversed(old_ID)
            new_relation = old_id.DEPREL
            new_upostag = old_id.UPOSTAG
```

비교 및 정략적 분석

ⅰ. 의존구문 말뭉치 비교

IV 비교 및 정량적 분석

❖ 의존구문 말뭉치 비교

	의존 구조 말뭉치 (원본)	의존 구조 말뭉치 (정제 후)	구묶음을 반영한 의존 구문 말뭉치
문장 수	62,345	49,292	49,292
표지 종류 수	45	45	**18**
형태소 수	1,566,560	1,054,859	**1,054,859**
말덩이 수	·	·	**804,854**
행 단위	토큰 (token)	토큰 (token)	**문장 성분 (constituent)**
노드 수 = 전체 행 수	713,238	526,378	**377,301**
의존 관계 태그 수	50	50	50

- 띄어쓰기 오류, 형태소 오류, 구묶음 오류를 포함한 13,053 문장 제외

IV 비교 및 정량적 분석

❖ 의존구문 말뭉치 비교

	의존 구조 말뭉치 (원본)	의존 구조 말뭉치 (정제 후)	구묶음을 반영한 의존 구문 말뭉치
문장 수	62,345	49,292	49,292
표지 종류 수	45	45	**18**
형태소 수	1,566,560	1,054,859	**1,054,859**
말덩이 수	·	·	**804,854**
행 단위	토큰 (token)	토큰 (token)	**문장 성분 (constituent)**
노드 수 = 전체 행 수	713,238	526,378	**377,301**
의존 관계 태그 수	50	50	50

- 띄어쓰기 오류, 형태소 오류, 구묶음 오류를 포함한 13,053 문장 제외

결론 및 향후 연구

ⅰ. 결론

ⅱ. 향후 연구

V 결론 및 향후 연구

❖ 결론

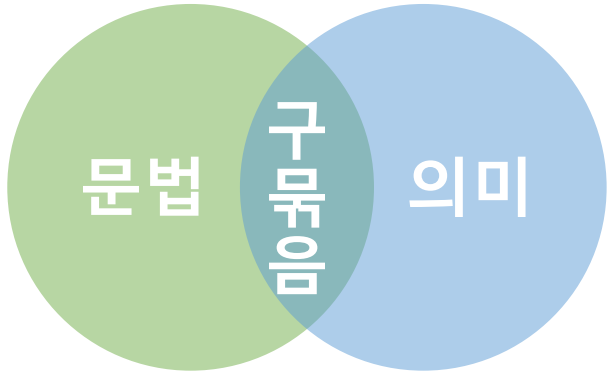
- **구문 분석 말뭉치** → **구묶음을 반영한 의존 구문 말뭉치**
- 지배소 결정의 방향성 문제 → 구문적 중심어와 의미적 중심어 일치
- 계산 복잡도 ↑ → 노드 수 감소 계산 복잡도 ↓

V 결론 및 향후 연구

❖ 향후 연구

- 본 논문을 통해 구축한 구묶음을 반영한 의존구문 말뭉치를 이용하여 한국어 의존구문분석 수행 및 비교

- 지속적으로 한국어 구묶음 분야 연구



감 사 합 니 다.

남궁 영, 김 재훈
한국해양대학교 컴퓨터공학과
young_ng@kmou.ac.kr

2019. 10. 12.