Setup

- **https://github.com/yseokchoi/SejongTree2Dependency**
- https://github.com/yseokchoi/KoreanDependencyParserusingStackPointer

➢ pytorch(0.4.1)

- pip3 install http://download.pytorch.org/whl/cpu/torch-0.4.1-cp36-cp36m-win_amd64.whl (python 3.6-windows)
- pip3 install http://download.pytorch.org/whl/cpu/torch-0.4.1-cp37-cp37m-win_amd64.whl (python 3.7-windows)
- pip3 install torch (OSX)

➢ numpy, gensim

- pip3 install numpy
- pip3 install gensim

Dependency Parsing을 위한 데이터 구축 실습

2018. 08.

최용석
충남대학교
정보검색 및 지식공학 연구실

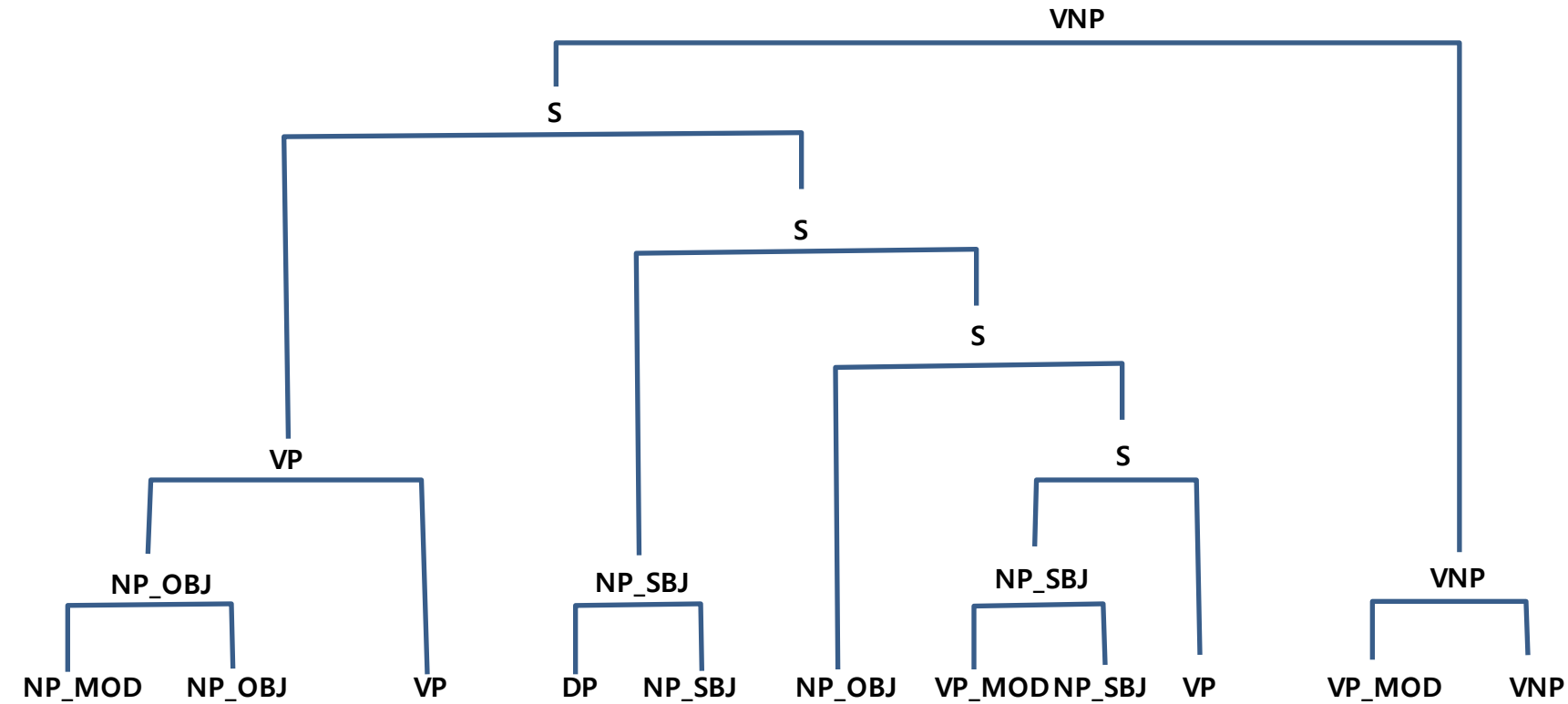
CONTENTS

- 세종 구문 코퍼스 변환: SejongTree2Dependency

- Transition-based Dependency Parser를 위한 데이터 만들기
  - ARC STANDARD + forward
  - ARC EAGER + backward

세종 구문 코퍼스 변환

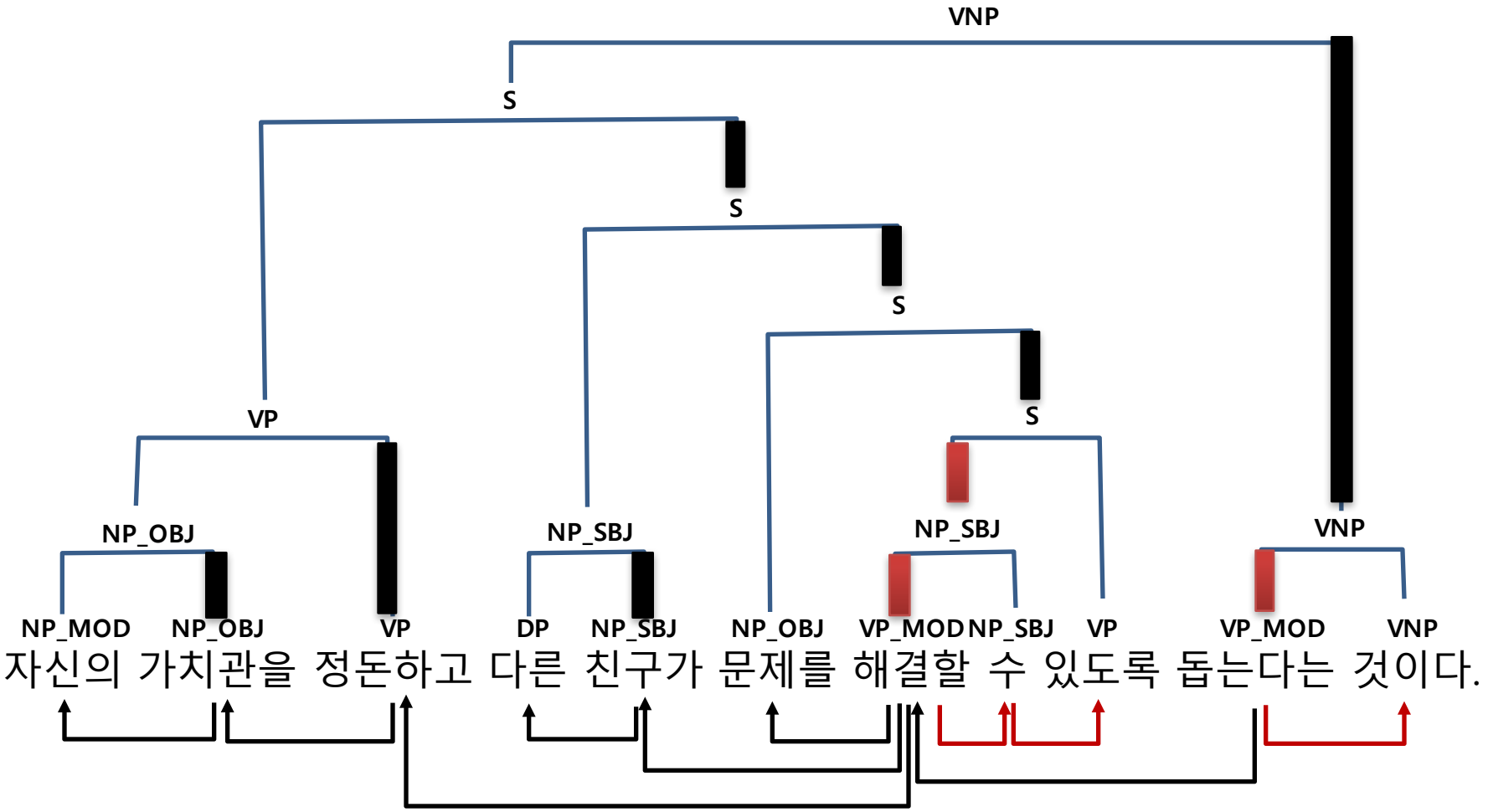
Sejong Phrase Structure



자신의 가치관을 정돈하고 다른 친구가 문제를 해결할 수 있도록 돕는다는 것이다.

세종 구문 코퍼스 변환

Sejong Phrase Structure & Dependency Structure



세종 구문 코퍼스 변환

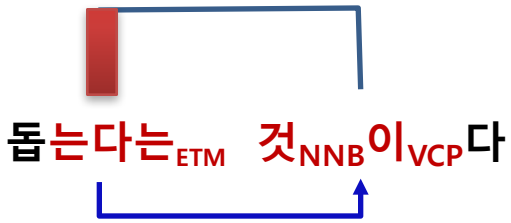
Head Initial 규칙

	Head Initial 규칙	예제
(1)	X→Y('ETM'으로끝남) Z('NNB+VCP'로시작)	서로 돕는다는 것이다
(2)	X→Y('EC로'끝남) Z('VX'로시작)	편지를 읽어 보았다
(3)	X→Y('ETM'으로끝남) W('수'포함) Z('있'으로시작)	과거를 생각할 수 있을까
(4)	X→Y('ETM'으로끝남) W('것'포함) Z('같'로시작)	웃음을 띠는 것 같았어요

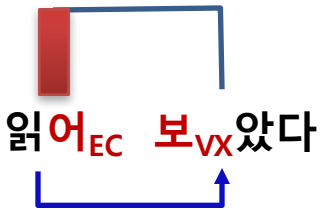
세종 구문 코퍼스 변환

Head Initial 규칙

	Head Initial 규칙	예제
(1)	X→Y('ETM'으로끝남) Z('NNB+VCP'로시작)	서로 돕<u>는다는 것</u>이다
(2)	X→Y('EC로'끝남) Z('VX'로시작)	편지를 읽<u>어 보</u>았다
(3)	X→Y('ETM'으로끝남) W('수'포함) Z('있'으로시작)	과거를 생각<u>할 수 있</u>을까
(4)	X→Y('ETM'으로끝남) W('것'포함) Z('같'로시작)	웃음을 띠<u>는 것 같</u>았어요



(1)

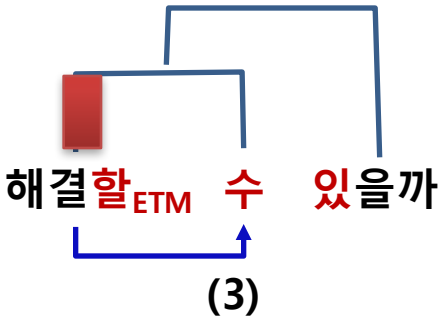


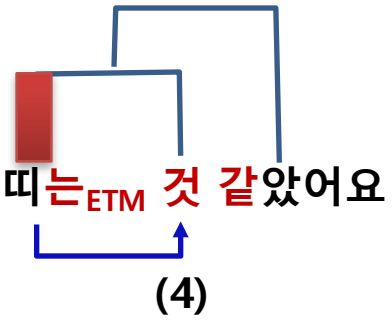
(2)

세종 구문 코퍼스 변환

Head Initial 규칙

	Head-Initial 예외 규칙	예제
(1)	X→Y('ETM'으로끝남) Z('NNB+VCP'로시작)	서로 돕는다는 것이다
(2)	X→Y('EC로'끝남) Z('VX'로시작)	편지를 읽어 보았다
(3)	**X→Y('ETM'으로끝남) W('수'포함) Z('있'으로시작)**	**문제를 해결할 수 있을까**
(4)	**X→Y('ETM'으로끝남) W('것'포함) Z('같'로시작)**	**웃음을 띠는 것 같았어요**





세종 구문 코퍼스 변환

CONLL-U Format

- CONLL-U Format(http://universaldependencies.org/format.html)

	Field	DESC
1	**ID**	Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes.
2	**FORM**	Word form or punctuation symbol.
3	**LEMMA**	Lemma or stem of word form.
4	**UPOSTAG**	Universal part-of-speech tag.
5	**XPOSTAG**	Language-specific part-of-speech tag; underscore if not available.
6	FEATS	List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available.
7	**HEAD**	Head of the current word, which is either a value of ID or zero (0).
8	**DEPREL**	Universal dependency relation to the HEAD (root if HEAD = 0) or a defined language-specific subtype of one.
9	DEPS	Enhanced dependency graph in the form of a list of head-deprel pairs.
10	**MISC**	Any other annotation.

세종 구문 코퍼스 변환

CONLL-U Format 예제

#SENTID:1
#FILE:practice.txt
#ORGSENT: 자신의 가치관을 정돈하고 다른 친구가 문제를 해결할 수 있도록 돕는다는 것이다.

#	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL*	DEPS	MISC
1	자신의	자신 의	NOUN	NNG+JKG	-	2	nmod	-	-
2	가치관을	가치관 을	NOUN	NNG+JKO	-	3	obj	-	-
3	정돈하고	정돈 하 고	VERB	NNG+XSV+EC	-	7	advcl	-	-
4	다른	다른	ADJ	MM	-	5	amod	-	-
5	친구가	친구 가	NOUN	NNG+JKS	-	7	nsubj	-	-
6	문제를	문제 를	NOUN	NNG+JKO	-	7	obj	-	-
7	해결할	해결 하 ㄹ	VERB	NNG+XSV+ETM	-	10	advcl	-	-
8	수	수	NOUN	NNB	-	7	aux	-	-
9	있도록	있 도록	VERB	VV+EC	-	8	aux	-	-
10	돕는다는	돕 는다는	VERB	VV+ETM	-	0	root	-	-
11	것이다.	것 이 다 .	ADJ	NNB+VCP+EF+SF	-	10	aux	-	-

* 변경될 수 있음

세종 구문 코퍼스 변환

CONLL-U Format 예제

#SENTID:6
#FILE:practice.txt
#ORGSENT: 웅가로는 **"실내** 장식품을 디자인할 때 옷을 만들 때와는 다른 해방감을 **느낀다"고** 말한다.

#	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL*	DEPS	MISC
1	웅가로는	웅가로 는	PROPN	NNP+JX	_	15	nsubj	_	_
2	"	"	PUNCT	SS	_	12	punct	_	SpaceAfter=No
3	실내	실내	NOUN	NNG	_	4	nmod	_	_
4	장식품을	장식품 을	NOUN	NNG+JKO	_	5	obj	_	_
5	디자인할	디자인 하 ㄹ	VERB	NNG+XSV+ETM	_	6	acl	_	_
6	때	때	NOUN	NNG	_	12	obl	_	_
7	옷을	옷을	NOUN	NNG+JKO	_	8	obj	_	_
8	만들	만들 ㄹ	VERB	VV+ETM	_	9	acl	_	_
9	때와는	때 와 는	NOUN	NNG+JKB+JX	_	10	nsubj	_	_
10	다른	다르 ㄴ	ADJ	VA+ETM	_	11	acl	_	_
11	해방감을	해방감 을	NOUN	NNG+JKO	_	12	obj	_	_
12	느낀다	느끼 ㄴ다	VERB	VV+EC	_	**15**	ccomp	_	SpaceAfter=No
13	"	"	PUNCT	SS	_	**12**	punct	_	SpaceAfter=No
14	고	고	ADP	JKQ	_	**12**	case	_	_
15	말한다.	말 하 ㄴ다 .	VERB	NNG+XSV+EF+SF	_	0	root	_	_

* 변경될 수 있음

세종 구문 코퍼스 변환

Sejong POS Tag to Universal POS Tag

대분류	세종 품사 태그	UPOSTAG
체언	NNG NNB	NOUN
	NNP	PROPN
	NR	NUM
	NP	PRON
용언	VV	VERB
	VA	ADJ
	VX	AUX
	VCP VCN	ADJ
접두사	XPN	PART
접미사	XSN	PART
	XSV	VERB
	XSA	ADJ

대분류	세종 품사 태그	UPOSTAG
관형사	MM	DET ADJ NUM
부사	MAG	ADV
	MAJ	CCONJ SCONJ
감탄사	IC	INJT
조사	J*	ADP
	JC	CCONJ
어미	E*	PART
어근	XR	NOUN

SejongTree2Dependency.pdf

세종 구문 코퍼스로부터 의존 구문 구조로 변환해보기

CNU 충남대학교 CHUNGNAM NATIONAL UNIVERSITY

- **Parameters**
  - **root_dir**: 세종 코퍼스 폴더 위치
  - **file_name(optional)**: 세종 코퍼스 파일 이름(하나의 세종 코퍼스 파일만 읽고자 할때)
  - **save_file**: 변환한 구문 구조를 저장할 파일 이름
  - **head_initial_file**: head-final 예외 규칙 파일명
  - **head_final**: 1 if ***head-final*** 0 else (default: 0)

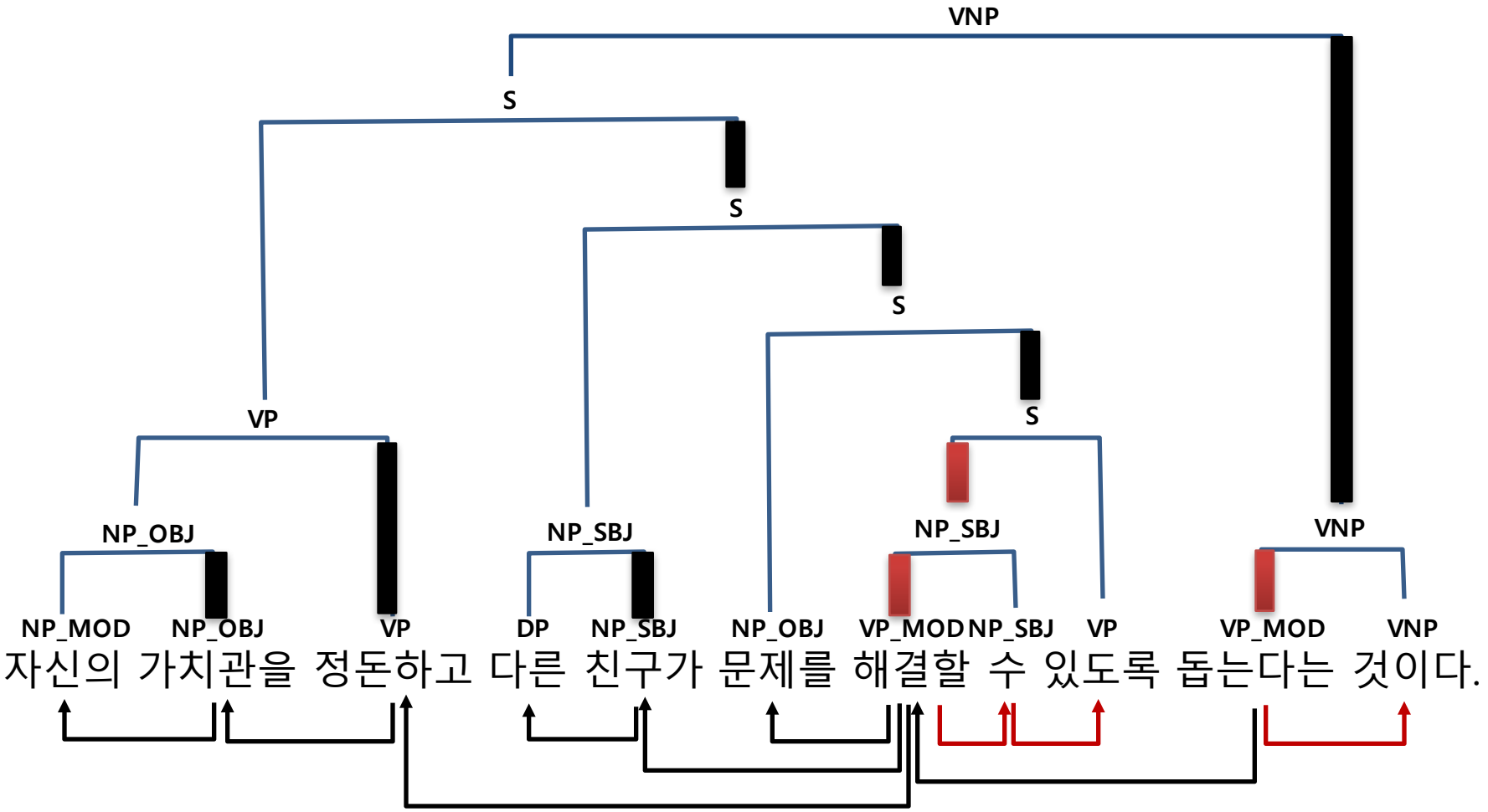
세종 구문 코퍼스로부터 의존 구문 구조로 변환해보기

- **실행 방법**

python SejongToDependency.py -root_dir *./Corpus* -save_file *sejong.conll* -head_initial_file *./Rules/linear_rules.txt*

세종 구문 코퍼스 변환

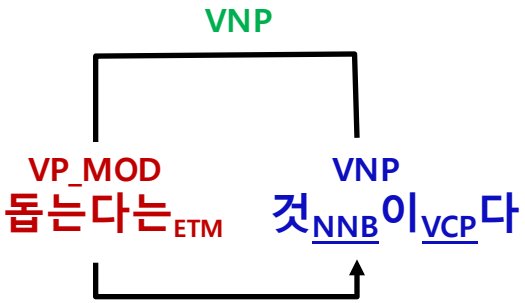
Sejong Phrase Structure & Dependency Structure



세종 구문 코퍼스 변환

Head Initial 규칙 정의

- **Head Initial 규칙 정의(두 어절)**
  - (1) parent label
  - (2) right most node label(RMN Label)
  - (3) right most node(RMN)
  - (4) left most node label(LMN Label)
  - (5) left most node(LMN)
  - (6) left context
  - (7) right context

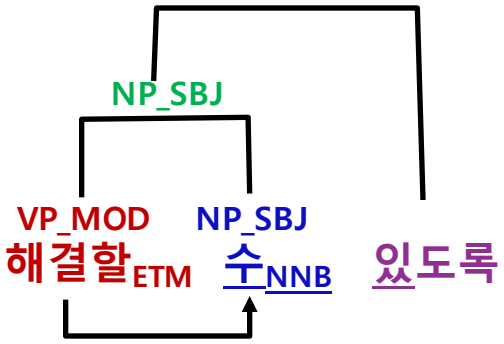


Parent Label	RMN Label	RMN	LMN Label	LMN	Left Context	Right Context
VNP*	**VP_MOD**	**-**	**VNP**	**NNB+VCP**	-	-

세종 구문 코퍼스 변환

Head Initial 규칙 정의

- **Head Initial 규칙 정의(세 어절)**
  - (1) parent label
  - (2) right most node label(RMN Label)
  - (3) right most node(RMN)
  - (4) left most node label(LMN Label)
  - (5) left most node(LMN)
  - (6) left context
  - (7) right context

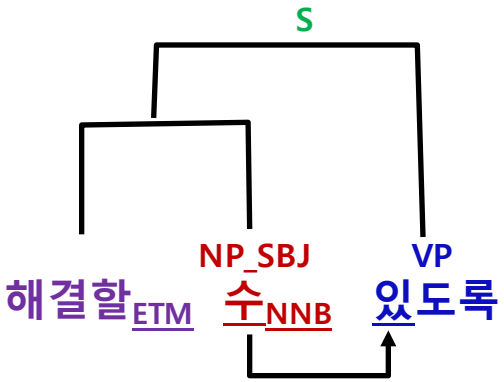


Parent Label	RMN Label	RMN	LMN Label	LMN	Left Context	Right Context
NP_SBJ	**VP_MOD**	**-**	**NP_SBJ**	**수/NNB**	-	**있**

세종 구문 코퍼스 변환

Head Initial 규칙 정의

- **Head Initial 규칙 정의(세 어절)**
  - (1) parent label
  - (2) right most node label(RMN Label)
  - (3) right most node(RMN)
  - (4) left most node label(LMN Label)
  - (5) left most node(LMN)
  - (6) left context
  - (7) right context



Parent Label	RMN Label	RMN	LMN Label	LMN	Left Context	Right Context
S*	**NP_SBJ**	**수/NNB**	**VP***	**있**	**ETM**	-

Transition-based Dependency Parser를 위한 데이터 만들기

- **Parameters**
  - **root_dir**: 세종 코퍼스 폴더 위치
  - **file_name(optional)**: 세종 코퍼스 파일 이름(하나의 세종 코퍼스 파일만 읽고자 할때)
  - **save_file**: 변환한 구문 구조를 저장할 파일 이름
  - **arc_standard:** ARC-Standard를 이용한 데이터 변환

Transition-based Dependency Parser를 위한 데이터 만들기

- **실행 방법**
  - **ARC-STANDARD**

python MakeTransitionCorpus.py -root_dir *./* -file_name *sejong.conll* -save_file sejong_*ARC_STANDARD.txt* -arc_standard

  - **ARC-EAGER**

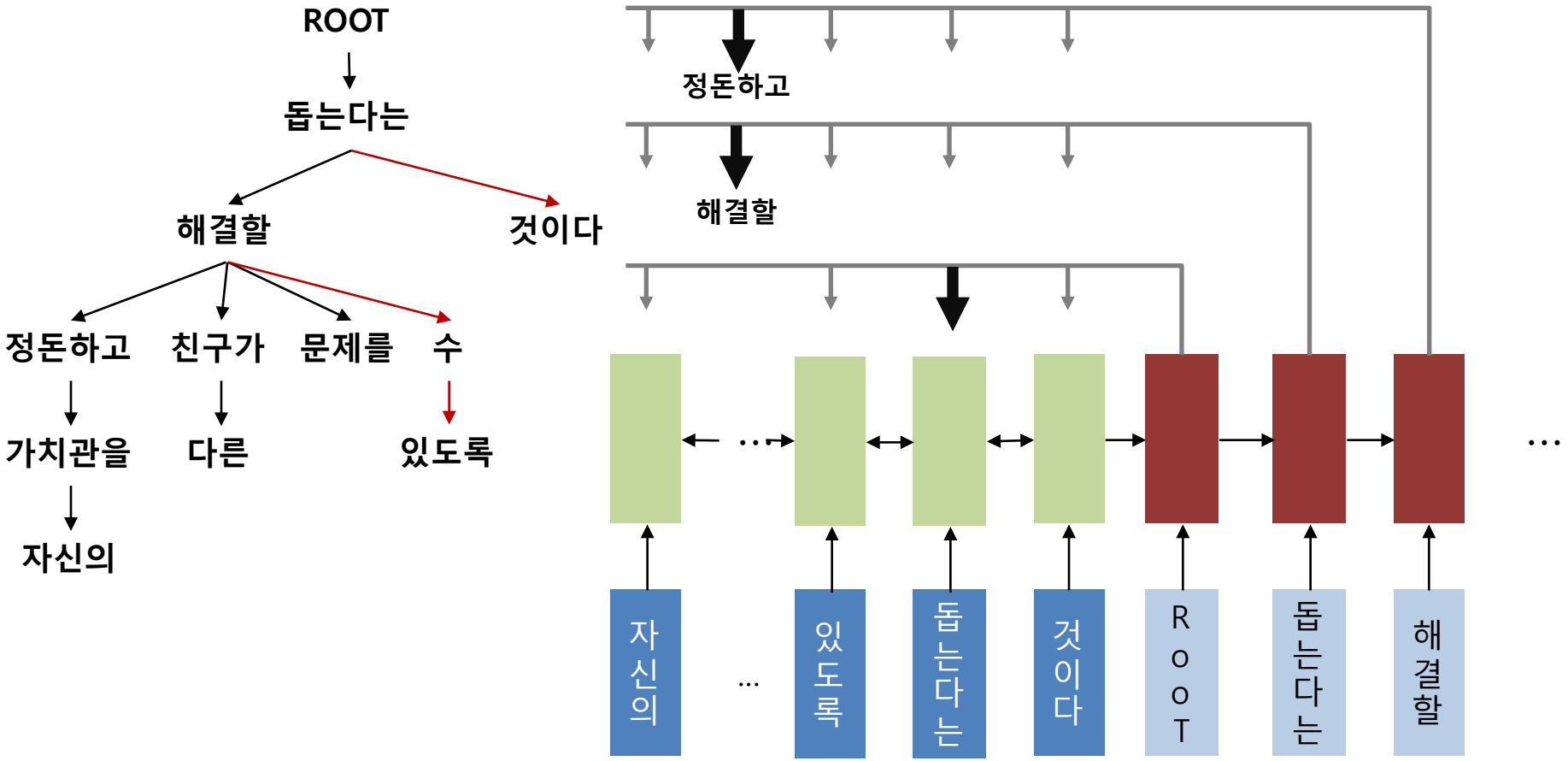
python MakeTransitionCorpus.py -root_dir *./* -file_name *sejong.conll* -save_file sejong_*ARC_EAGER.txt*

의존 구문 파서

Stack-Pointer Networks

(Xuezhe Ma, 2018)

SENT: 자신의 가치관을 정돈하고 다른 친구가 문제를 해결할 수 있도록 돕는다는 것이다.



MA, Xuezhe, et al. Stack-Pointer Networks for Dependency Parsing. arXiv preprint arXiv:1805.01087, 2018.

의존 구문 파서

실습

- **Parameters**
  - **model_path**: 저장한 모델 폴더 위치
  - **model_name**: 저장한 모델 이름
  - **output_path**: 결과 파일 저장할 폴더 위치
  - **test:** test file 위치+파일명

의존 구문 파서

실습

- **실행 방법**

python StackPointerParser_test.py --model_path *./models/* --model_name *models.pt* --output_path *./output/* --test *./test/sejong.conll*

The End
Thank you