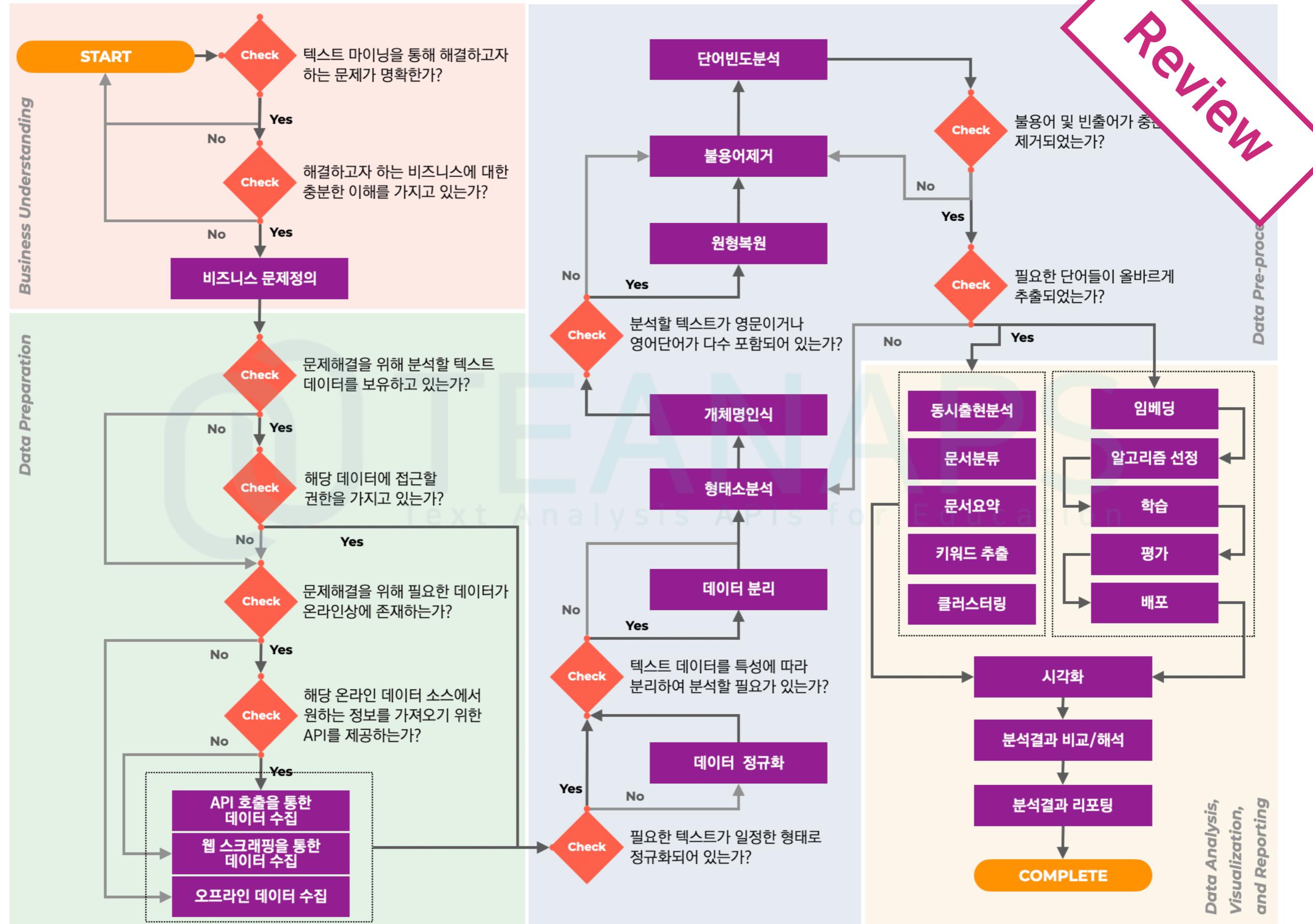


ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)



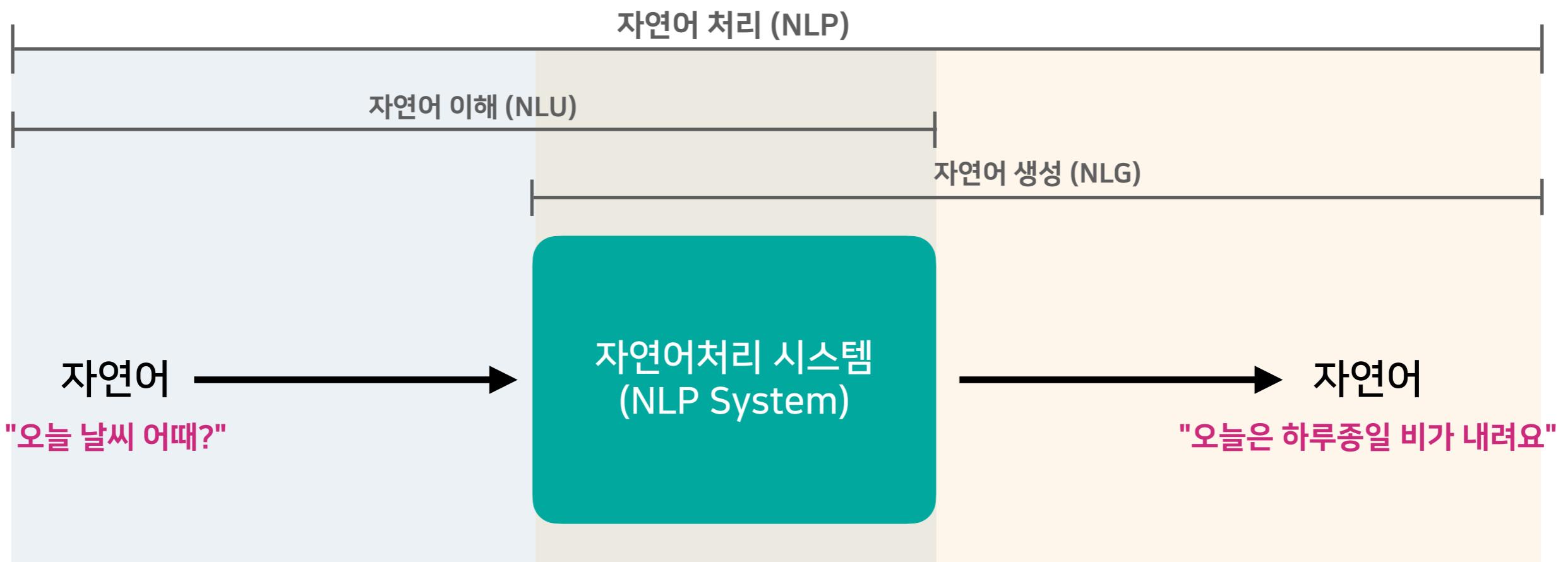
WEEK 05

Pre-processing Text Data

자연어처리 개요 (Natural Language Processing, NLP)

자연어처리 구성요소

- 컴퓨터가 자연어를 이해하거나 생성할 수 있도록 하는 일련의 과정
- **자연어** (natural language) : 사람들이 일상생활에서 자연스럽게 사용하는 언어
- **자연어 이해** (Natural Language Understanding, NLU) : 문자로된 언어를 입력으로 목적에 맞게 내부적으로 처리해나가는 과정
- **자연어 생성** (Natural Language Generation, NLG) : 주어진 정보를 바탕으로 자연어 문장을 생성하는 과정



자연어처리 유형

어휘/형태소 분석 (Lexical/Morphology Analysis)

- 문서를 단어(_{word}) 단위로 구분하고 단어의 표면적/내재적 의미를 분석하는 방법
- **형태소** (_{morpheme}) : 뜻을 가진 가장 작은 말의 단위 (동사, 명사, 조사, 문장부호 등의 품사)
- **형태소 분석** (_{morphology analysis}) : 문장을 형태소 단위로 분리하고 품사(_{part of speech})를 태깅하는 과정
- 주요 형태소 품사 : 일반명사(_{NNG}), 고유명사(_{NNP}), 형용사(_{VA}), 동사(_{VV}), 일반부사(_{MAG}), 어미(_{EC}), 외국어(_{SL}), 숫자(_{SN})

구분	내용													
원문	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홫 스 퍼 의 승 리 를 이 꿀 었다 .													
음절	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홫 스 퍼 의 승 리 를 이 꿀 었다 .													
형태소	손흥민 이 골 을 작렬 하 며 토크 넘 홫스퍼 의 승리 를 이끌 었다 .													
어절	손흥민이 골을 작렬하며 토크 넘 홫스퍼의 승리를 이끌었다 .													

자연어처리 유형

형태소 품사 태그표

구분	품사태그	설명	구분	품사태그	설명
체언	NNG	일반명사	선어말 어미	EP	선어말어미
	NNP	고유명사		EF	종결어미
	NNB	의존명사		EC	연결어미
	NR	수사		ETM	명사형 전성어미
	NP	대명사		ETD	관형형 전성어미
용언	VV	동사	접두사	XPN	체언접두사
	VA	형용사		XSN	명사파생 접미사
	VX	보조용언		XSV	동사파생 접미사
	VCP	긍정지정사		XSA	형용사 파생 접미사
	VCN	부정지정사		XR	어근
관형사	MM	관형사	부호	SF	마침표, 물음표, 느낌표
부사	MAG	일반부사		SP	쉼표, 가운뎃점, 콜론, 빗금
	MAJ	접속부사		SS	따옴표, 괄호, 줄표
감탄사	IC	감탄사		SE	줄임표
조사	JKS	주격조사		SO	물결표, 숨길표, 빠짐표
	JKC	보격조사		SW	기타 기호
	JKG	관형격조사		NF	명사추정범주
	JKO	목적격조사		NV	용언추정범주
	JKB	부사격조사		NA	분석불능범주
	JKV	호격조사	미식별	SL	외국어
	JKQ	인용격조사		SH	한자
	JX	보조사		SN	숫자
	JC	접속조사			

자연어처리 유형

구문 분석 (Syntax Analysis)

- 문장에 출현하는 단어 사이의 구조적 관계를 분석하는 방법
- 한 문장 내 단어 사이의 구조적 관계는 트리(tree) 구조로 표현할 수 있으며, 트리 구조로 표현이 어려운 경우에는 문법적으로 맞지 않은 문장이라고 볼 수 있음
→ 문법적으로 맞는 경우일지라도 의미적(semantics)으로는 맞지 않은 문장일 수도 있음

구분	내용
원문	손 흥 민 이 골 을 작 렬 하 며 토크 넘 홫 스퍼 의 승 리 를 이 끌 었 다 .

구문분석

형태소	손흥민	이	골	을	작렬	하	며	토크	넘	핫스퍼	의	승리	를	이끌	었	다	.
-----	-----	---	---	---	----	---	---	----	---	-----	---	----	---	----	---	---	---

자연어처리 유형

| 의미론적 분석 (Semantics Analysis)

- 문장이 의미적으로 올바른지 여부를 분석하는 방법
- 문법적(Syntax)으로는 옳바른 문장이라도 의미적으로 올바르지 않은 경우가 있음
 - 목적어-동사 관계가 맞지 않은 경우 : 나는 밥을 먹었다. / ~~나는 자동차을 먹었다.~~
 - 명사-형용사 관계가 맞지 않은 경우 : 나는 맛있는 밥을 먹었다. / ~~나는 화질좋은 밥을 먹었다.~~

구분	내용
원문	식탁이 골을 작렬하며 토크 넘 홫스퍼의 식사를 이끌었다.

구문분석

형태소	식탁	이	골	을	작렬	하	며	토크	넘	핫스퍼	의	식사	를	이끌	었	다	.

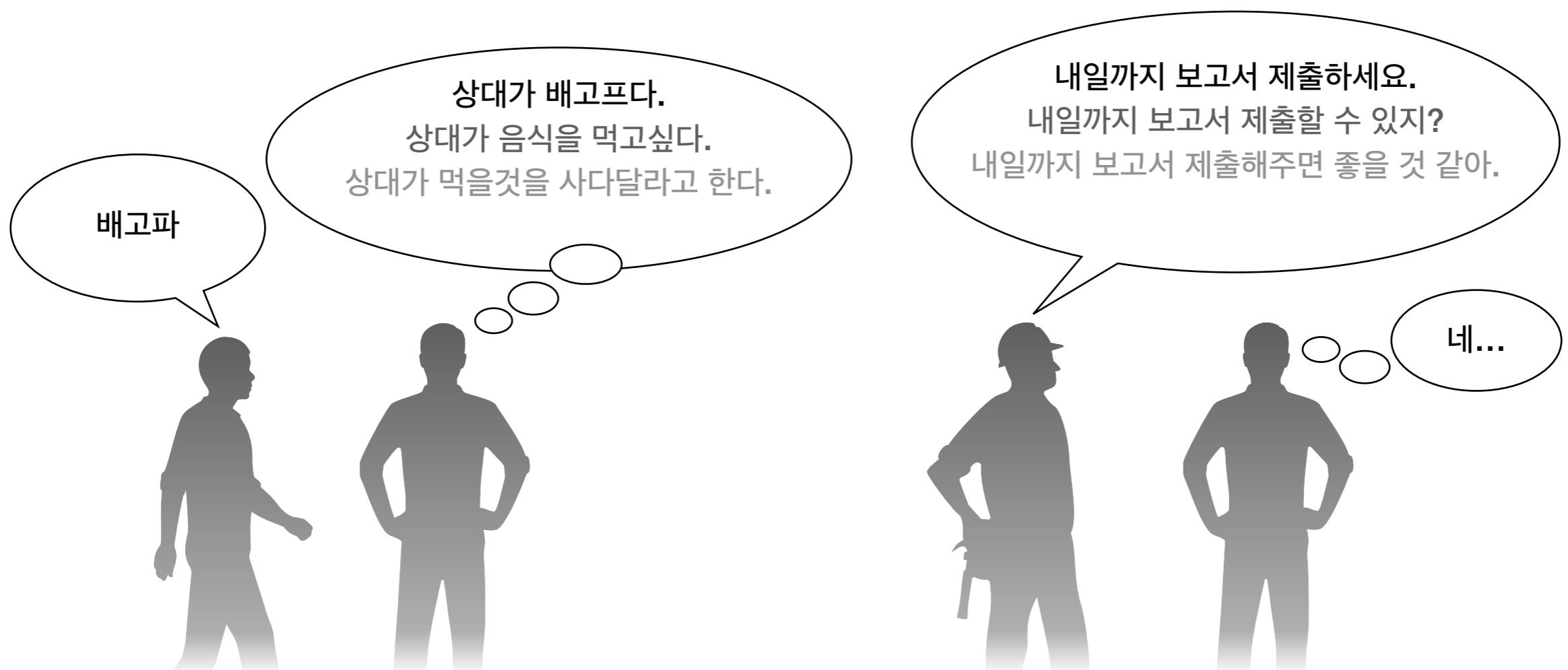
자연어처리 유형

화용론적 분석 (Pragmatics Analysis)

- 언어가 특정 목적을 달성하기 위해 어떻게 사용되는지 분석하는 방법
 - 의사소통 과정에서 대화상대나 문맥을 고려하여 어떤 표현을 사용해야하는지 연구하는 분야
- "죄송하지만 힘들 것 같습니다." ≈ "도와드릴 수 없다"

양해

결론



자연어처리 유형

담화/대화 분석 (Discourse & Dialogue analysis)

- 문서 또는 대화 내 여러개의 문장 속에서 문맥에 따른 문장의 의미를 분석하는 방법
- 문맥에 따라 달라질 수 있는 의미들 중 가장 가능성이 높은 의미를 선정하는 과정



HOW?

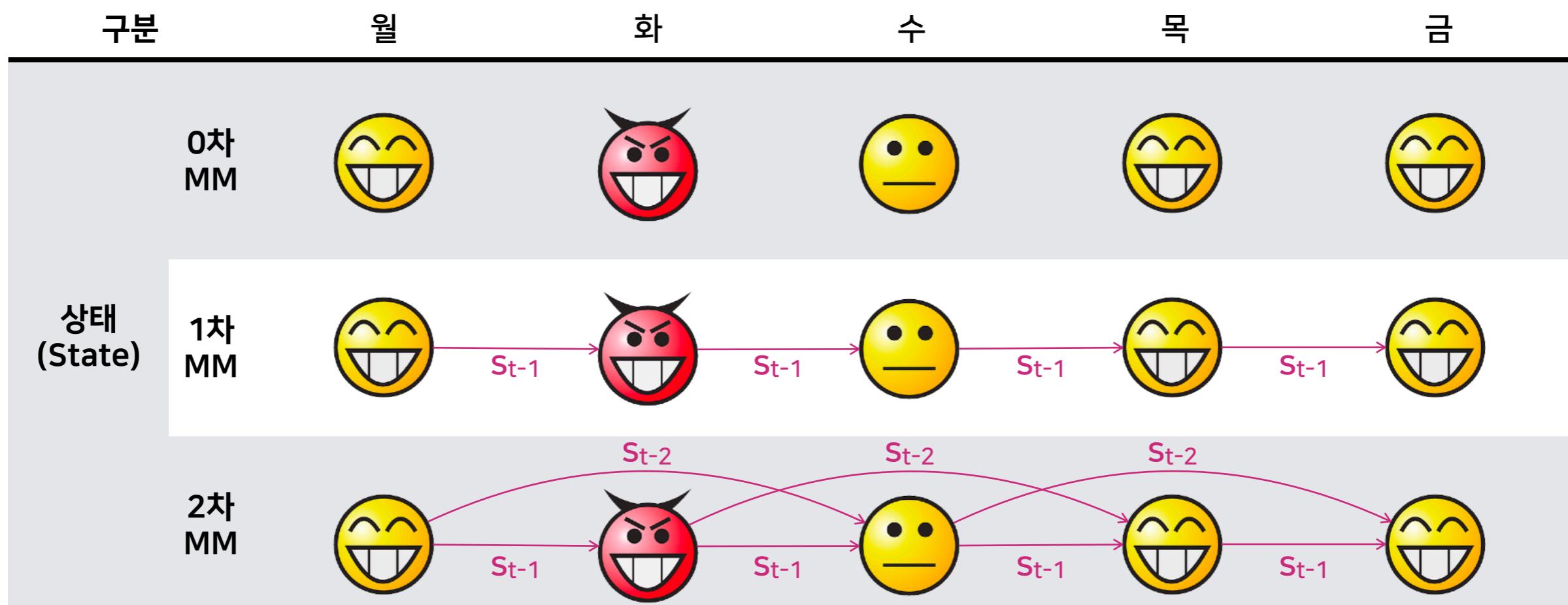
자연어처리 대부분의 문제는
순서에 따른 분류 문제

순차 레이블링

(Sequence Labeling)

마르코프 모델 (Markov Model, Markov Chain)

- 순차적으로 출현하는 상태를 예측하는 문제에서 과거와 현재의 상태(state)가 주어졌을 때, 미래 상태가 과거에 출현한 N개 상태에서만 영향을 받는다는 가정을 바탕으로 한 순차적 확률분포 모델
- 0차 마르코프 가정 (Markov Assumption) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t)$
- 1차 마르코프 가정 (1st Order Markov Assumption, bigram model) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t|S_{t-1})$
- 2차 마르코프 가정 (2nd Order Markov Assumption) : $P(S_t|S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t|S_{t-1}, S_{t-2})$



순차 레이블링

(Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델

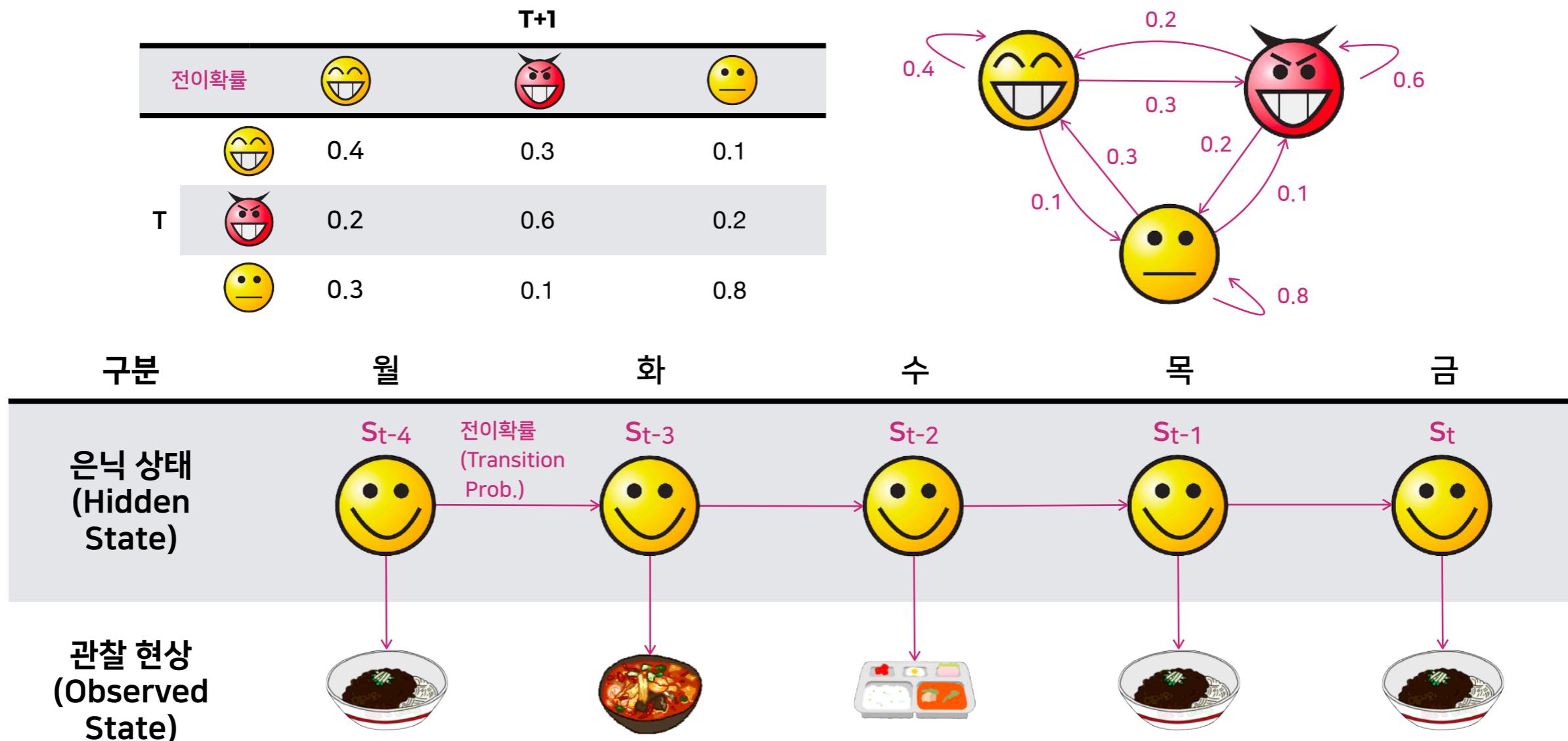


$$\begin{aligned}
 P(S_t | S_{t-4}, S_{t-3}, S_{t-2}, S_{t-1}) &= \prod P(S_{t-4}) P(S_{t-3} | S_{t-4}) P(S_{t-2} | S_{t-3}) P(S_{t-1} | S_{t-2}) P(S_t | S_{t-1}) \\
 &= \prod 0.4 \times 0.3 \times 0.2 \times 0.3 \times 0.4 = 0.00288\pi = 2.88 \times 10^{-3}\pi \text{ (1차 마르코프 가정)}
 \end{aligned}$$

순차 레이블링 (Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델

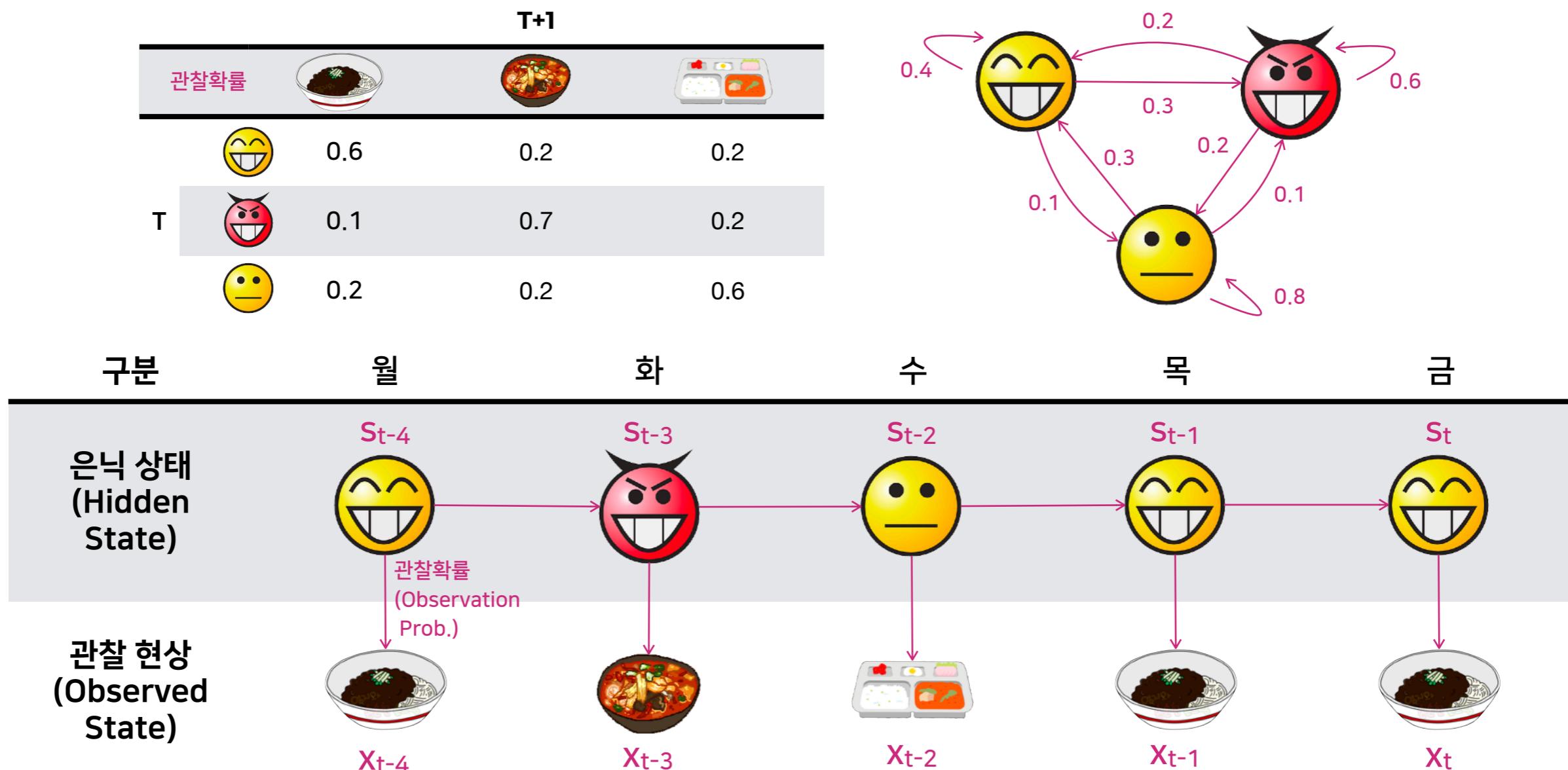


순차 레이블링

(Sequence Labeling)

은닉 마르코프 모델 (Hidden Markov Model, HMM)

- 순차적으로 출현하는 상태를 예측하는 문제에서 상태를 직접 관찰할 수 없는 경우, 미래 상태가 오직 상태의 영향을 받는 과거에 관찰된 N개 현상에서만 영향을 받는다(독립가정)는 가정을 바탕으로 한 순차적 확률분포 모델



순차 레이블링

(Sequence Labeling)

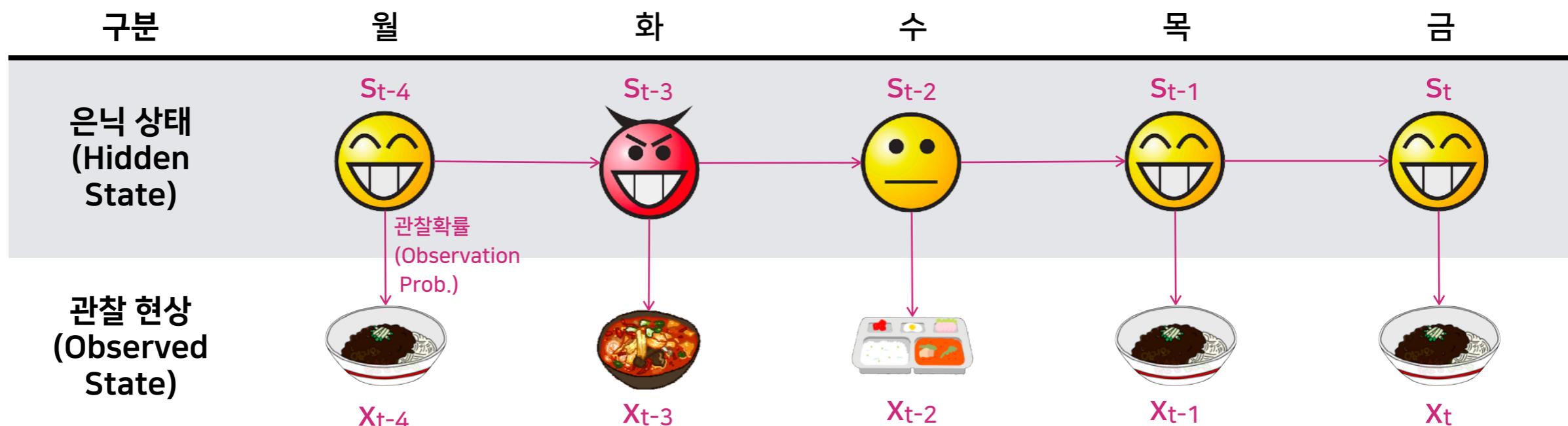
은닉 마르코프 모델 (Hidden Markov Model, HMM)

$$P_T = P(S_t | S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t | S_{t-1}) \quad (\text{1차 마르코프 가정})$$

$$P_{O|T} = P(X_t | S_{t-N}, X_{t-N}, \dots, S_{t-2}, X_{t-2}, S_{t-1}, X_{t-1}) = P(X_t | S_t) \quad (\text{관찰확률 반영})$$

$$\begin{aligned} P_{HMM} &= \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) = \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) P(s_{t-N,n}) \\ &\approx \underset{s_{t-N,n}}{\operatorname{argmax}} \sum_{t=1}^n P(x_t | s_t) \underset{\substack{\text{관찰확률} \\ \text{초기확률}}}{P(s_t | s_{t-1})} \quad (\text{1차 마르코프 가정, 독립가정 반영}) \end{aligned}$$

Independance Assumption



순차 레이블링

(Sequence Labeling)

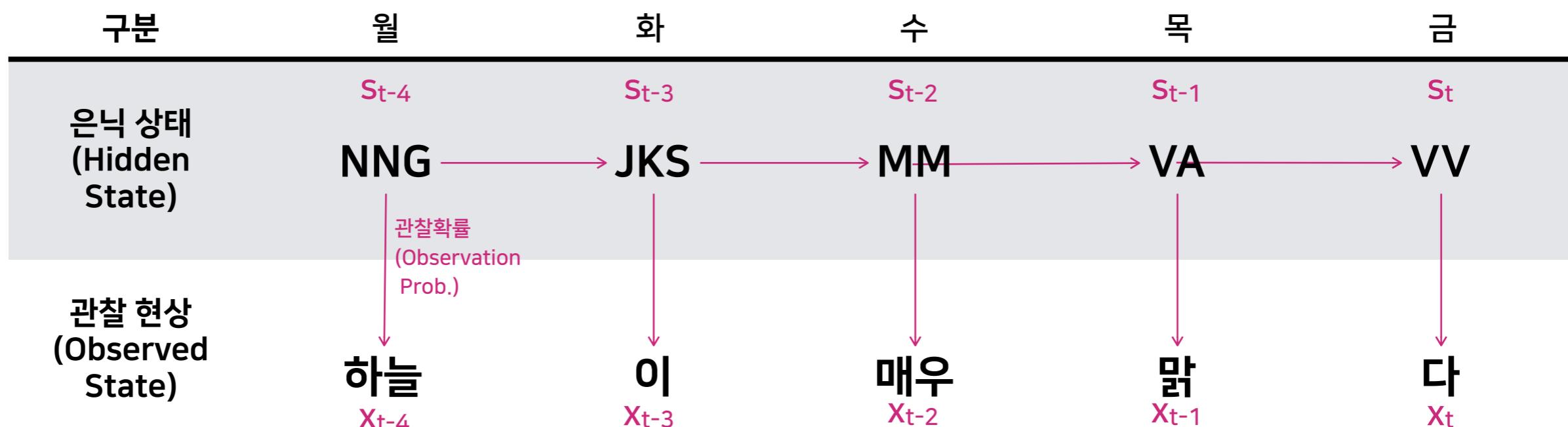
은닉 마르코프 모델 (Hidden Markov Model, HMM)

$$P_T = P(S_t | S_{t-N}, \dots, S_{t-2}, S_{t-1}) = P(S_t | S_{t-1}) \quad (\text{1차 마르코프 가정})$$

$$P_{O|T} = P(X_t | S_{t-N}, X_{t-N}, \dots, S_{t-2}, X_{t-2}, S_{t-1}, X_{t-1}) = P(X_t | S_t) \quad (\text{관찰확률 반영})$$

$$\begin{aligned} P_{HMM} &= \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) = \underset{s_{t-N,n}}{\operatorname{argmax}} P(x_{t-N,n}, s_{t-N,n}) P(s_{t-N,n}) \\ &\approx \underset{s_{t-N,n}}{\operatorname{argmax}} \sum_{t=1}^n P(x_t | s_t) \underset{\substack{\text{관찰확률} \\ \text{초기확률}}}{P(s_t | s_{t-1})} \quad (\text{1차 마르코프 가정, 독립가정 반영}) \end{aligned}$$

Independance Assumption

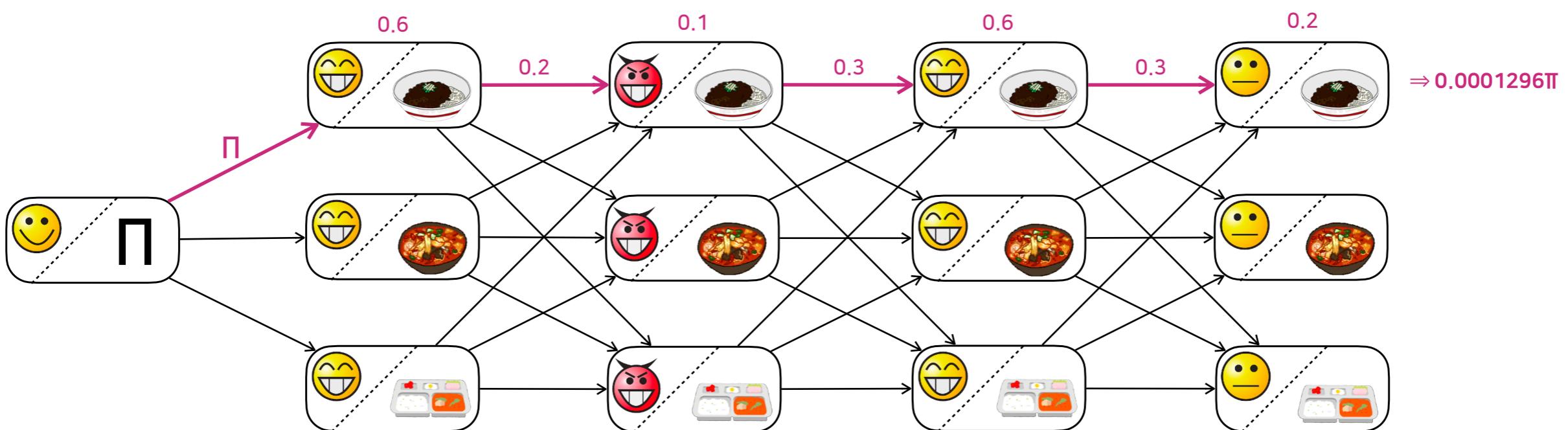


순차 레이블링 (Sequence Labeling)

순차 레이블링 경로 분석 (Sequence Labeling Path Analysis)

		T+1		
		전이확률	웃음표	노는 표
			0.4	0.3
T		0.2	0.6	0.2
	느낌표	0.3	0.1	0.8

		T+1		
		관찰확률	국밥	탕수육
			0.6	0.2
T		0.1	0.7	0.2
	느낌표	0.2	0.2	0.6



순차 레이블링

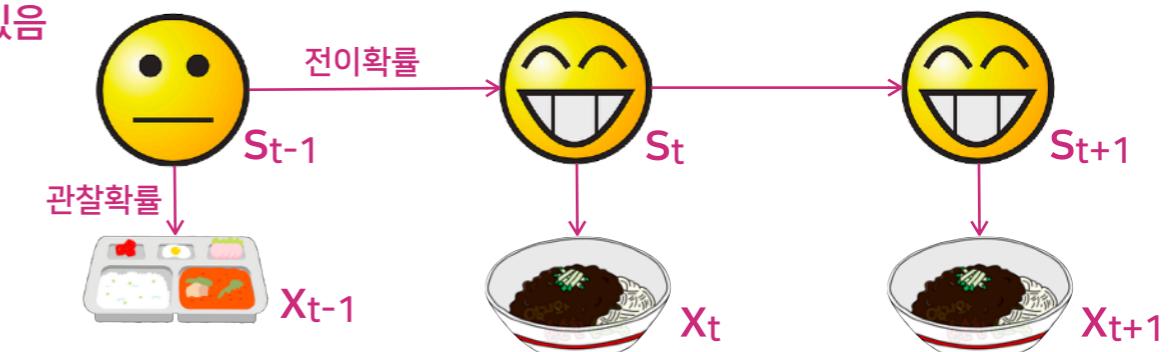
(Sequence Labeling)

HMM & MEMM (Maximum Entropy MM) & CRFs (Conditional Random Fields)

(HMM) "독립가정"에 의해 단어-단어 간 관계(문맥)을 반영하지 못한다는 단점이 있음

$$P_{\text{HMM}} \approx \operatorname{argmax} \prod_{t=1}^n P(x_t|s_t) P(s_t|s_{t-1})$$

(1차 마르코프 가정, 독립가정 반영)



(MEMM) 현재 단어에서 추출된 자질(feature)로 상태(State)를 예측함 ('90 중반)

$$P_{\text{MEMM}}(y, x) = \frac{1}{Z(x)} \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x, y) \right\}$$

$$Z(x) = \sum_y \exp \left\{ \sum_{i=1}^n \lambda_i f_i(x, y) \right\}$$

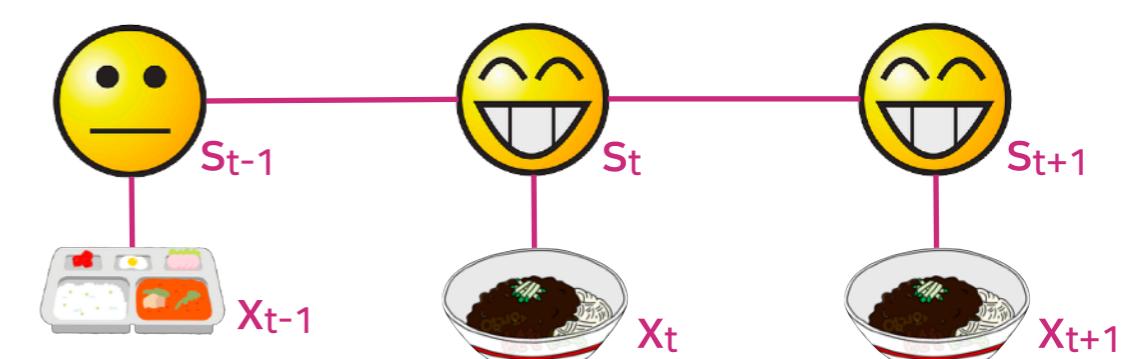
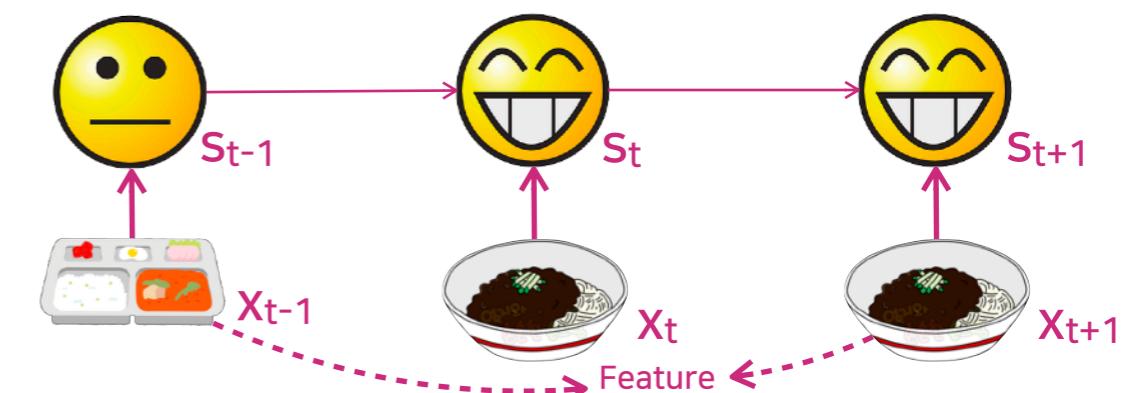
(CRFs) 전이확률을 계산할 때 앞뒤 순서를 모두 고려하여 상태(State)를 예측함 ('90 후반)

$$F_j(y, x) = \sum_{i=1}^n f_i(y_{i-1}, y_i, x, i)$$

$$P(y|x, \lambda) = \frac{1}{Z(x)} \exp \sum_j \lambda_j f_j(x, y)$$

Global Feature Function

Global Normalization



순차 레이블링: RNN

RNN을 활용한 순차 레이블링

순차 레이블링: RNN

RNN을 활용한 순차 레이블링

순차 레이블링: RNN

RNN을 활용한 순차 레이블링

구분

내용 (B: Beginner, I: Inner, O: Outer)

원문 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 꼴 었 다 .

띄어쓰기 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 꼴 었 다 .

토큰화 손흥민 이 골 을 작렬 하 며 토트넘 홋스퍼 의 승리 를 이꼴 었 다 .

품사태깅 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 꼴 었 다 .

B-NNP	I-NNP	I-NNP	B-JKS	B-NNG	B-JKO	B-NNG	I-NNG	B-XSV	B-EC	B-NNG	I-NNG	I-NNG	B-NNG	I-NNG	I-NNG	B-JKG	B-NNG	I-NNG	B-JKO	B-VV	I-VV	B-EP	B-EF	B-SF
-------	-------	-------	-------	-------	-------	-------	-------	-------	------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	------	------	------	------	------

순차 레이블링: RNN

RNN을 활용한 순차 레이블링

구분

내용 (B: Beginner, I: Inner, O: Outer)

원문 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홫 스 퍼 의 승 리 를 이 꼴 었 다 .

띄어쓰기	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홧	스	퍼	의	승	리	를	이	꼴	었	다	.
------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

토큰화	손흥민	이	골	을	작렬	하	며	토트넘	홋스퍼	의	승리	를	이끌	었	다	.
-----	-----	---	---	---	----	---	---	-----	-----	---	----	---	----	---	---	---

품사태깅	NNP	JKS	NNG	JKO	NNG	XSV	EC	NNG	NNG	JKG	NNG	JKO	VV	EP	EF	SF
------	-----	-----	-----	-----	-----	-----	----	-----	-----	-----	-----	-----	----	----	----	----

개체명인식	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	훗	스	퍼	의	승	리	를	이	꼴	었	다	.
	B-PER	I-PER	I-PER	O	O	O	O	O	O	O	B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	O	O	O	O	O	O	O	O	

HOW?

다시 형태소 분석과 개체명인식
문제로 돌아와서...

한국어의 언어학적 특징

언어를 구성하는 단위

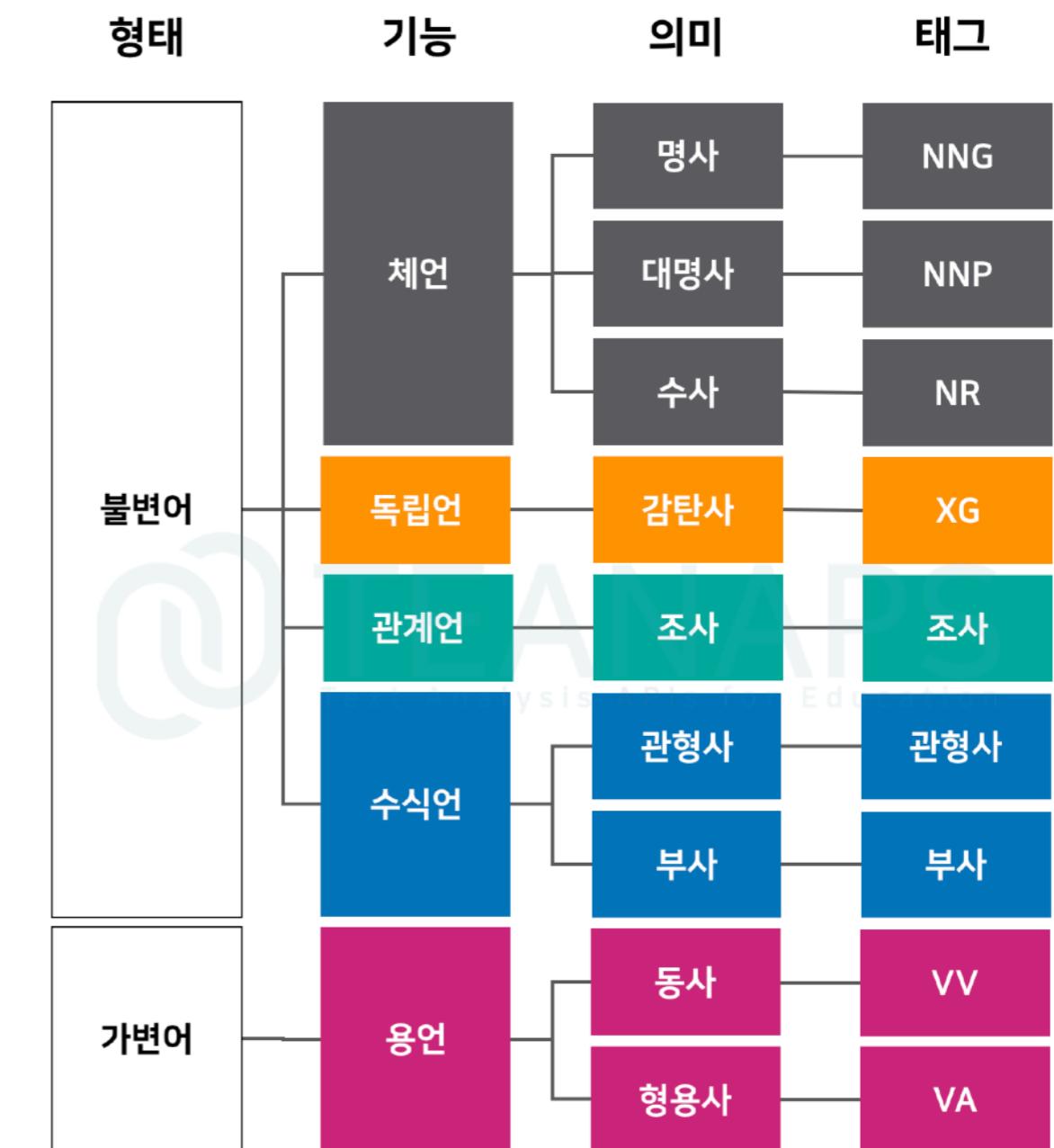
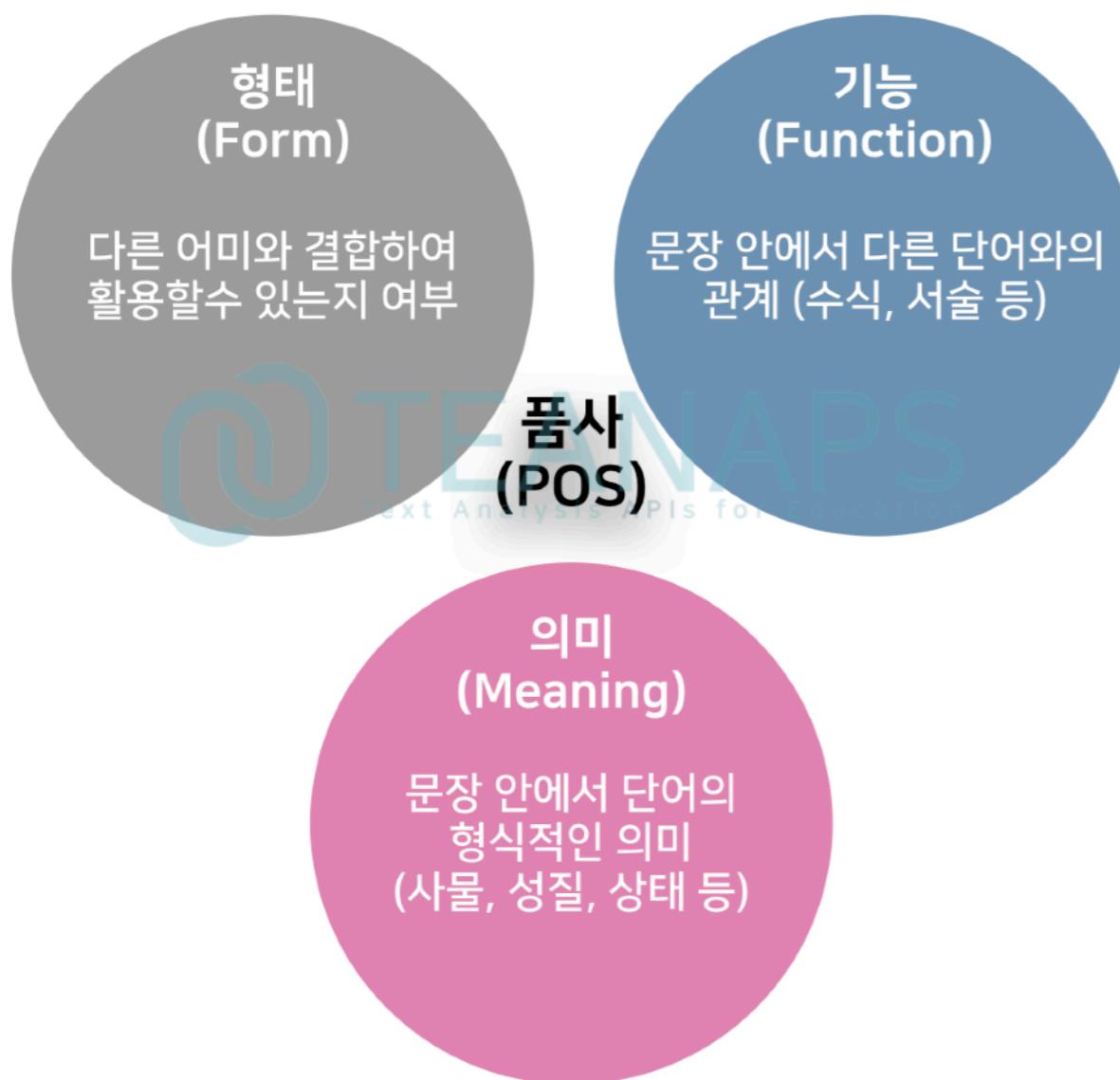
- **음절 (syllable)** : 언어를 말하고 들을 때 하나의 덩어리로 여겨지는 가장 작은 발화의 단위로, 한국어 음절은 자음(consonant)과 모음(vowel)으로 구성된 초성(onset), 중성(nucleus), 종성(coda)의 조합으로 이루어짐
- **형태소 (Morpheme)** : 언어에서 의미를 가지는 가장 작은 단위로, 실제 의미를 가지는 실질 형태소(어휘 형태소)와 문법적 역할을 하는 형식 형태소(문법 형태소)로 구분할 수 있음
- **어절 (?)** : 한 개 이상의 형태소가 모여 구성된 단위로, 발화 시 끊어서 말하거나 대부분의 띄어쓰기를 구분함

구분	내용												
원문	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 흛 스 퍼 의 승 리 를 이 꼴 었 다 .												
음절	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 흛 스 퍼 의 승 리 를 이 꼴 었 다 .												
형태소	손흥민 이 골 을 작렬 하 며 토트넘 흛스퍼 의 승리 를 이끌 었 다 .												
어절	손흥민이 골을 작렬하며 토트넘 흛스퍼의 승리를 이끌었다 .												

한국어의 언어학적 특징

한국어 5언 9품사

- 한국어는 단어를 기능(function), 의미(meaning), 형태(form)의 세 가지 기준에 의해 분류함



자연어처리 유형

Review

형태소 품사 태그표

구분	품사태그	설명	구분	품사태그	설명
체언	NNG	일반명사	선어말 어미	EP	선어말어미
	NNP	고유명사		EF	종결어미
	NNB	의존명사		EC	연결어미
	NR	수사		ETM	명사형 전성어미
	NP	대명사		ETD	관형형 전성어미
용언	VV	동사	접두사	XPN	체언접두사
	VA	형용사		XSN	명사파생 접미사
	VX	보조용언		XSV	동사파생 접미사
	VCP	긍정지정사		XSA	형용사 파생 접미사
	VCN	부정지정사		XR	어근
관형사	MM	관형사	부호	SF	마침표, 물음표, 느낌표
부사	MAG	일반부사		SP	쉼표, 가운뎃점, 콜론, 빗금
	MAJ	접속부사		SS	따옴표, 괄호, 줄표
감탄사	IC	감탄사		SE	줄임표
조사	JKS	주격조사		SO	물결표, 숨길표, 빠짐표
	JKC	보격조사		SW	기타 기호
	JKG	관형격조사		NF	명사추정범주
	JKO	목적격조사		NV	용언추정범주
	JKB	부사격조사		NA	분석불능범주
	JKV	호격조사	미식별	SL	외국어
	JKQ	인용격조사		SH	한자
	JX	보조사		SN	숫자
	JC	접속조사			

한국어의 언어학적 특징

교착어, 굴절어, 그리고 고립어

- **교착어** (*agglutinative language*) : 어근에 접사가 결합되어 각 단어의 기능을 나타내는 언어
- **굴절어** (*inflectional language*) : 단어 자체의 형태변화로 그 단어의 문법성을 나타내는 언어
- **고립어** (*isolating language*) : 단어의 형태변화 없이 문법적 관계는 어순에 의해 정해지는 언어

어근	피동	높힘	과거	추측	전달	어미	파생된 단어
일어나						+다	일어나다
일어나	+지					+다	일어나지다
일어나	+지	+시				+다	일어나지시다
일어나	+지	+시	+었			+다	일어나지셨다
일어나				+았		+다	일어났다
일어나					+겠	+다	일어나겠다
일어나					+더라		일어나더라
일어나		+지	+었			+다	일어나졌다
일어나		+지	+었	+겠		+다	일어나졌겠다
일어나	+지	+었	+겠		+더라		일어나졌겠더라
일어나				+았	+겠	+다	일어났겠다
일어나	+지	+시	+았	+겠	+더라		일어나지셨겠더라

한국어의 언어학적 특징

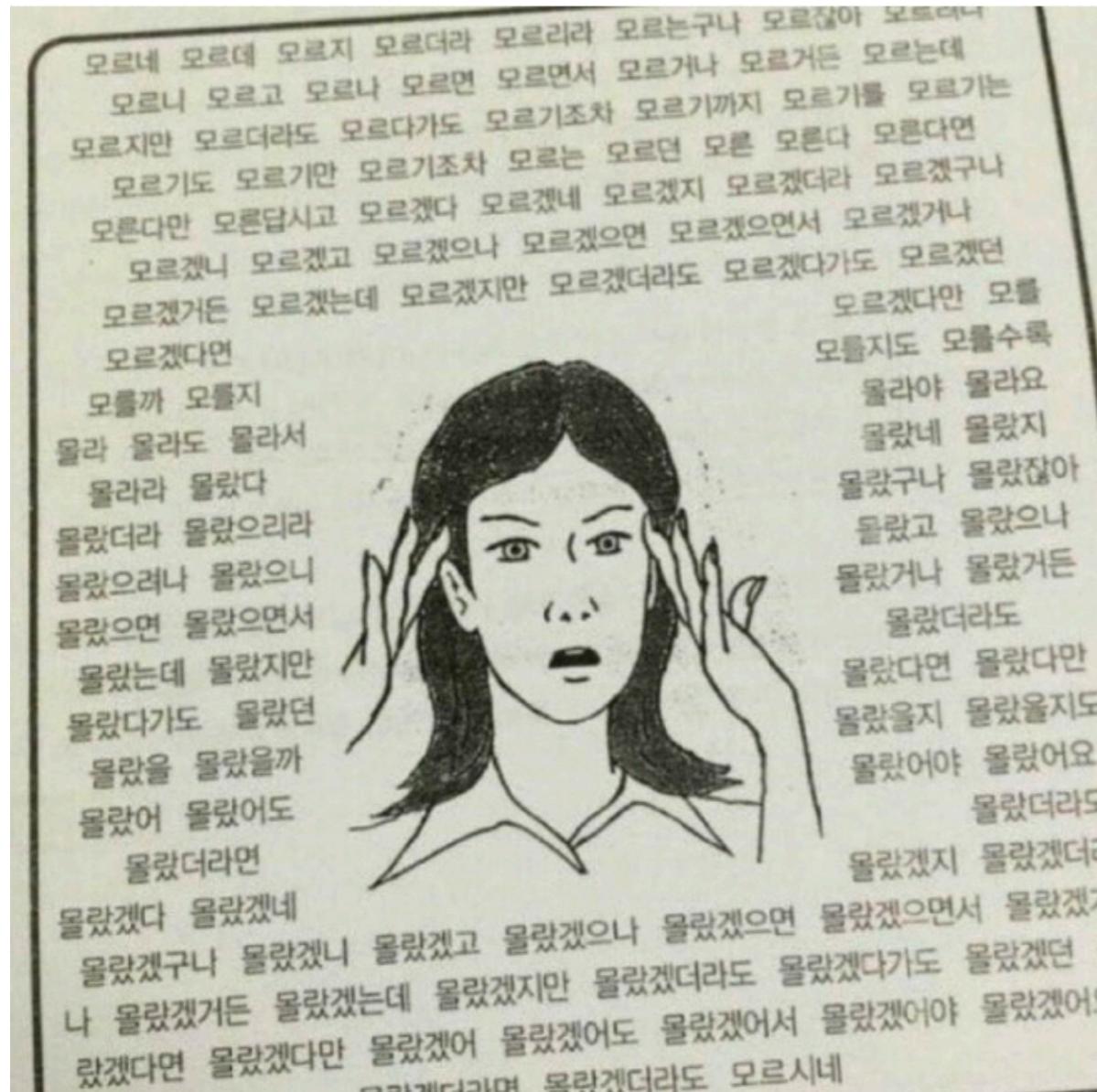
교착어 텍스트 분석이 어려운 이유

- 동일한 단어임에도 불구하고 어근에 따라 단어가 다양한 형태로 파생되어 표현됨
- 단어가 다양하게 생겨나므로 하나의 어근에서 비롯된 비슷한 의미의 단어가 매우 많이 생성됨
- 형태소 분석 시 수많은 경우의 수를 다르게 처리하거나 추가적인 토큰화가 필요함
- 어근에 따라 단어의 역할이 정의되기 때문에, 어순이 전체적인 문장의 의미에 미치는 영향이 상대적으로 매우 적음

구분	한국어	영어
원문	나는 오전수업을 들으러 학교에 간다.	I morning classes to take to school go.
문장 1	간다 나는 오전수업을 들으러 학교에.	Go I to morning classes take to school.
문장 2	학교에 간다 나는 오전수업을 들으러.	To school go I morning classes to take.
문장 3	오전수업을 들으러 학교에 간다 나는.	Morning classes to take to school go I.
문장 4	나는 간다 학교에 들으러 오전수업을.	I go to school to take morning classes.
문장 5	간다 학교에 오전수업을 들으러 나는.	Go to school morning classes to take I.
문장 6	학교에 들으러 나는 간다 오전수업을.	To school to take I go morning classes.
문장 7	나는 간다 들으러 학교에 오전수업을.	I go to take to school morning classes.
문장 8	들으러 학교에 나는 오전수업을 간다.	To take school I morning classes go to.

텍스트 마이닝이 어려운 이유

Review



I just got here.

상기 문장은 영어로 "나 막 도착했어" 가 된다. 자연스럽게 위 문장을 바꿀 수 있는

I have just arrived 하나 정도다.

한국어에서 저 just라는 표현은 대체 수십 가지로 가능하다.

나 막 왔어.

나 방금 왔어.

나 지금 왔어.

나 금방 왔어.

나 온 지 조금/좀 됐어. (조금에 강세)

나 온 지 별로/얼마 안 됐어.

나 이제 왔어.

나 바로 막 왔어.

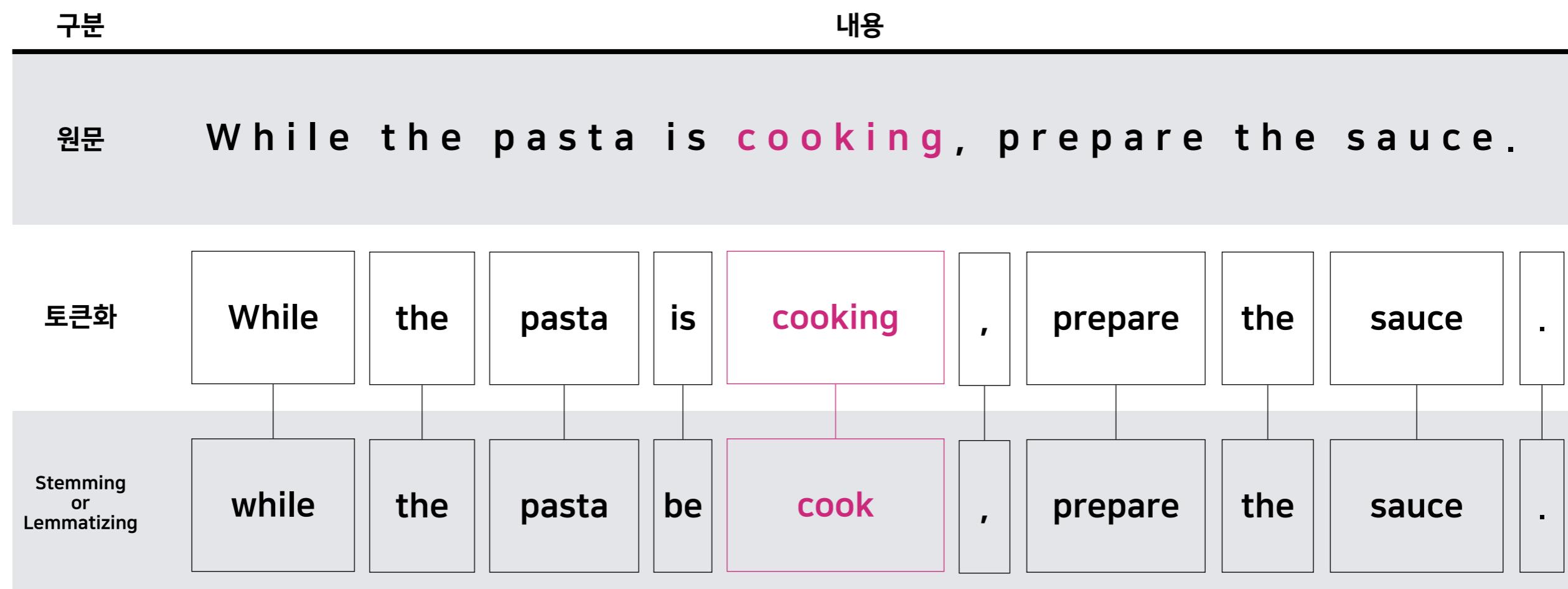
게다가 위의 모든 표현의 '왔어'를 "도착했어"로 바꿔도 말이 된다.

- 시발ㅋ, 시발ㅋㅋ : 웃김
- 오 시발 : 놀라움
- 마 시발 : 마쉬움
- 시발... : 슬픔
- 시발! : 분노
- 시발; : 어이없음
- 시발ㅠㅠ : 격한슬픔
- 시발;; : 당황스러움
- 시바ㄹ : 급함
- 시ㅂ : 더욱 급함
- tlqkf : 정말로 급함

영어 형태소 분석

원형복원: 어간추출과 표제어추출 (Stemming & Lemmatization)

- 원형에서 다른 형태로 변형되어 활용된 단어를 원형으로 복원하는 과정
- **어간추출 (Stemming)** : 규칙 기반으로 단어의 변형된 형태를 제거하거나 치환하여 원형으로 복원하는 방법
cooking (v) → cook, cooking (n) → cook, cookbooks → cookbook, believes → believ, using → us
- **표제어추출 (Lemmatizing)** : 단어의 형변환 사전을 기반으로 대상 단어의 품사에 맞는 단어의 원형으로 복원하는 방법
cooking (v) → cook, cooking (n) → cooking, cookbooks → cookbook, believes → believe, using → use



한국어 형태소 분석

토큰화 (Tokenization)

- 문서 또는 문자열이 주어졌을 때 이를 의미있는 최소단위의 토큰(형태소) 단위로 분리하는 작업
- 경우에 따라 구두점 등 불필요하거나 의미를 담고 있지 않은 토큰은 제외하기도 함
- 영어는 언어학적 특성상 단어에 조사가 붙지 않아 토큰화가 비교적 쉬우나 (어절단위 분리과정과 동일함), 한국어는 언어학적 특성으로 다른 언어에 비해 토큰화가 매우 어려움

구분	내용																
원문	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 끌 었 다 .																
토큰화	손흥민	이	골	을	작	렬	하	며	토트넘	홋스퍼	(?)	의	승리	를	이끌	었	다.

한국어 형태소 분석

품사태깅 (Part of Speech Tagging)

- 토큰화된 문장의 각 토큰에 대해 앞뒤 문맥상 적합한 품사를 태깅하는 작업
- 토큰화가 잘못되거나, 동일한 토큰임에도 문맥에 따라 다양한 의미를 가지는 토큰이 존재하는 경우 품사태깅 결과가 잘못될 가능성이 큼
- 도메인에 특화 문서는 별도의 형태소 사전을 통해 토큰의 유형과 형태소 태그를 관리하여 품사태깅에 반영해야함



자연어처리 유형

Review

형태소 품사 태그표

구분	품사태그	설명	구분	품사태그	설명
체언	NNG	일반명사	선어말 어미	EP	선어말어미
	NNP	고유명사		EF	종결어미
	NNB	의존명사		EC	연결어미
	NR	수사		ETM	명사형 전성어미
	NP	대명사		ETD	관형형 전성어미
용언	VV	동사	접두사	XPN	체언접두사
	VA	형용사		XSN	명사파생 접미사
	VX	보조용언		XSV	동사파생 접미사
	VCP	긍정지정사		XSA	형용사 파생 접미사
	VCN	부정지정사		XR	어근
관형사	MM	관형사	부호	SF	마침표, 물음표, 느낌표
부사	MAG	일반부사		SP	쉼표, 가운뎃점, 콜론, 빗금
	MAJ	접속부사		SS	따옴표, 괄호, 줄표
감탄사	IC	감탄사		SE	줄임표
조사	JKS	주격조사		SO	물결표, 숨길표, 빠짐표
	JKC	보격조사		SW	기타 기호
	JKG	관형격조사		NF	명사추정범주
	JKO	목적격조사		NV	용언추정범주
	JKB	부사격조사		NA	분석불능범주
	JKV	호격조사	미식별	SL	외국어
	JKQ	인용격조사		SH	한자
	JX	보조사		SN	숫자
	JC	접속조사			

한국어 형태소 분석

형태소 분석 (Morphological Analysis)

- 문장을 형태소 단위로 구분하고 품사를 구별하여 태깅하고 용언의 변형으로 탈락한 형태소를 복원하는 과정
- 형태소 분석기마다 형태소를 구분하는 방식이 다르기 때문에 특성에 맞는 형태소 분석기를 선택해야함
- 형태소 분석의 활용 범위
 - 1) 언어학적 측면 : 특정 언어현상의 생성과정을 설명하는 데 용이하게 쓰일 수 있음
 - 2) 전산학적 측면 : 정보검색이나 자연어 처리 자동 처리시스템의 구문 분석의 전 단계 등의 용도로 쓰일 수 있음



오픈소스 형태소 분석기 특징

한나눔 형태소 분석기 (Hannanum)

- 한국과학기술원(KAIST)의 SWRC(Semantic Web Research Center)에서 개발한 형태소 분석기
- 자동 띄어쓰기 모듈을 제공해 형태소 분석 결과를 활용하여 한글 문장에 대한 자동 띄어쓰기 수행 가능
- 사전 기반의 맞춤법 교정 모듈로 형태소 분석 결과를 활용하여 한글 단어에 대한 맞춤법 교정 수행 가능



오픈소스 형태소 분석기 특징

Okt (Twitter)

- 트위터에서 개발한 한국어 형태소 분석기
- SNS에서 발생하는 언어에서 자주 발생하는 인물명, 신조어 등을 잘 인식하는 장점이 있음
- 타 형태소 분석기 대비 속도가 빠른 편이지만 형태소 태그 구분이 명확하지 않고 품질이 상대적으로 낮음



오픈소스 형태소 분석기 특징

꼬꼬마 형태소 분석기 (KKMA)

- 서울대 IDS(Intelligent Data Systems) 연구실에서 자연어 처리를 위한 모듈구축과제로 개발한 형태소 분석기
- Java 언어를 기반으로 하며, Python-Java 연동을 통해 Python에서 사용 가능하도록 배포됨
- 가능한 모든 형태소 후보와 조합을 모두 찾아 그 중 가장 적합한 형태소를 판단함 → 매우느림



오픈소스 형태소 분석기 특징

Mecab (은전, 은전한닢)

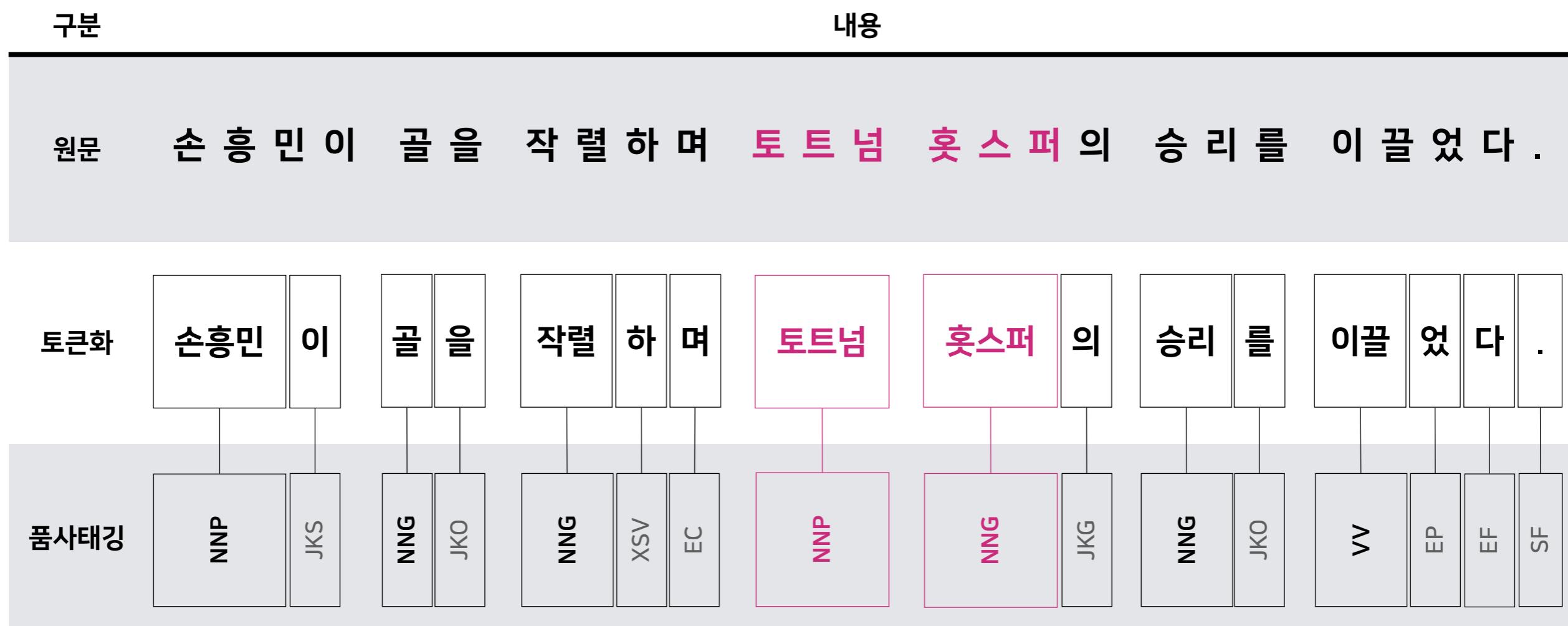
- 검색에서 쓸만한 오픈소스 한국어 형태소 분석기를 목적으로 개발된 한국어 형태소 분석기
- 오픈소스 검색엔진 Elasticsearch에 형태소 분석 모듈로 적용되어 활용되고 있음
- 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌



오픈소스 형태소 분석기 특징

카카오 Khaiii

- 카카오에서 DHA2(Daumkakao Hangul Analyzer 2)를 계승하여 개발하고 2018년 공개한 형태소분석기
- 속도를 매우 중요시하며, 신경망 알고리즘 중에서 CNN(Convolutional Neural Network)을 사용하여 개발됨
- 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌



형태소 분석기 성능비교

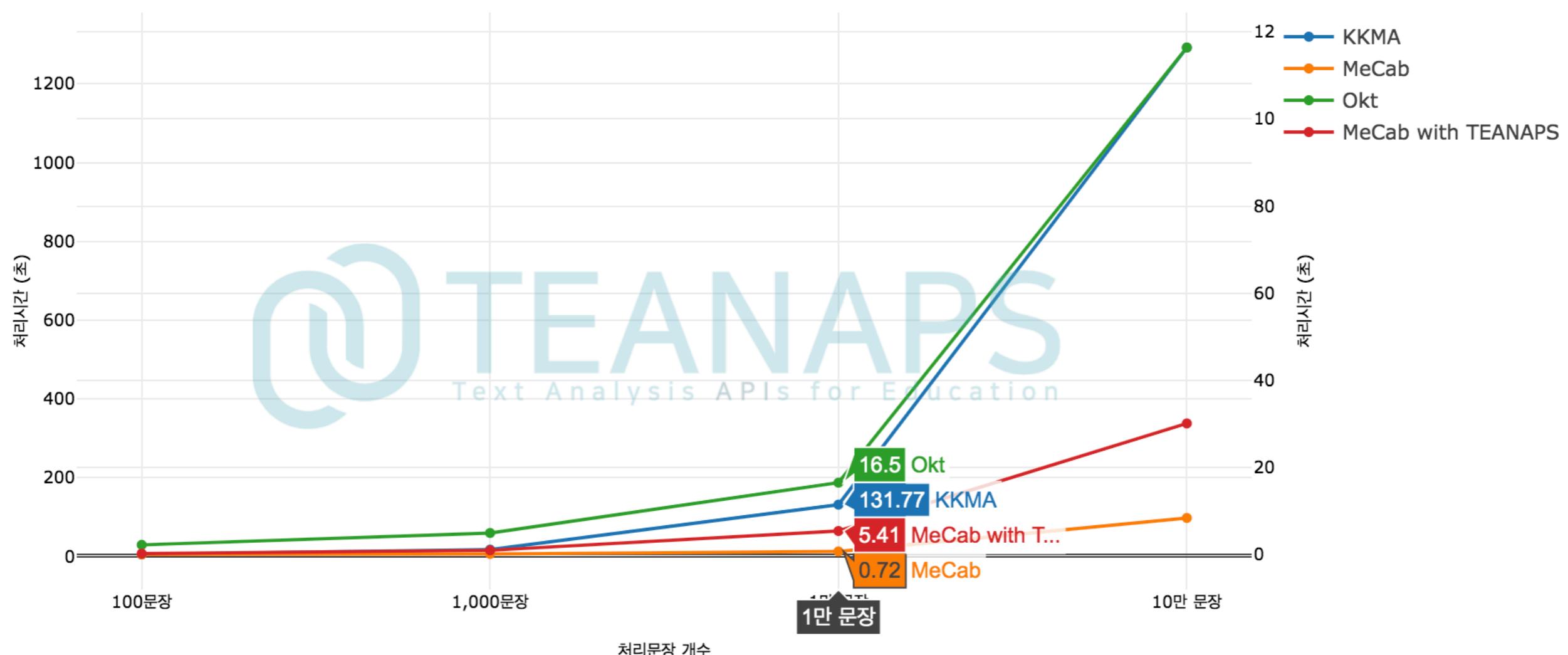
형태소 분석기 별 결과 비교

구분	내용																							
원문	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다
한나눔	손흥민		이		골	ㄹ	작렬하	이	며		토트넘			홋스퍼		의		승리	를	이끌	었다		.	
Okt	손흥민	이		골	을		작렬		하며		토트넘			홋스퍼		의		승리	를	이끌었다			.	
꼬꼬마	손	흥	민	이	골	을	작렬	하	며		토트	넘		홋스퍼		의		승리	를	이끌	었	다	.	
Mecab	손흥민	이		골	을		작렬	하	며		토트넘			홋스퍼의				승리	를	이끌	었	다	.	
Khaiii	손흥민	이		골	을		작렬	하	며		토트넘			홋스퍼		의		승리	를	이끌	었	다	.	

형태소 분석기 성능비교

형태소 분석기 별 수행시간 비교 (Time Analysis)

형태소 분리 및 품사태깅 평가결과



한국어 NLP 데이터셋 모음

코퍼스 명	용도	설명	링크
Naver sentiment movie corpus v1.0	분류	네이버 영화 리뷰 (긍정, 부정) 분류 라벨링 됨	https://github.com/e9t/nsmc
Chatbot_data	분류	채팅 대화 (일상, 긍정, 부정) 분류 라벨링 됨	https://github.com/songys/Chatbot_data
청와대 국민청원 사이트의 만료된 청원 데이터 모음	RAW	일자, 카테고리, 제목, 내용 등 만료된 청원 Raw 데이터	https://github.com/akngs/petitions
Korean NER Corpus	NER	한국어 NER 용 데이터 (NER, 형태소)	https://github.com/machinereading/KoreanNERCorpus
Korean Parallel corpora	번역	번역용 한국어/영어, 한국어/불어 병렬 데이터	https://github.com/j-min/korean-parallel-corpora
KorQuAD 1.0	MRC	MRC 용 Wikipedia에 대한 질문 답변 데이터	https://korquad.github.io/category/1.0_KOR.html
KorQuAD 2.1	MRC	MRC 용 Wikipedia에 대한 질문 답변 데이터 (1.0 보다 데이터가 큼)	https://korquad.github.io/
AI허브 AI데이터	다양	법률, 특허, 상식, 대화 등 다양한 분야의 학습용 데이터 제공 (데이터 신청 별도 해야함)	http://www.aihub.or.kr/ai_data
국립국어원 언어정보나눔터	다양	말뭉치, 대화 자료등등 방대한 한국어 데이터 제공 (학습을 위해서는 전처리가 많이 필요함)	https://ithub.korean.go.kr/user/total/database/corpusManager.do

문서를 표현하는 방법

단어 주머니 (Bag of Words)

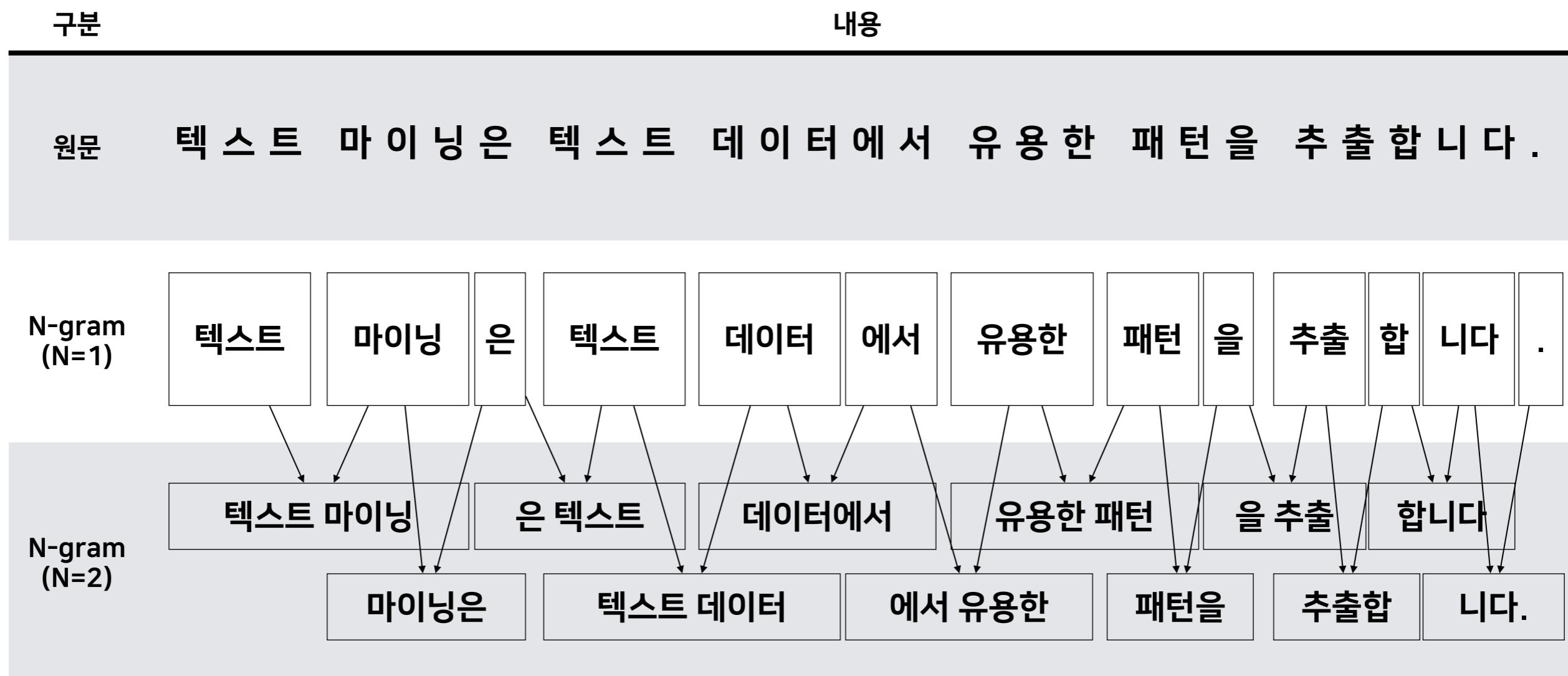
- 문서에 함께 사용된 단어의 집합
- 중복된 단어는 하나로 취급하며, 순서에 의미를 고려하지 않음
- “아버지가 방에 들어가신다.” → [“아버지”, “가”, “방”, “에”, “들어가다”]



문서를 표현하는 방법

N-그램 (N-gram)

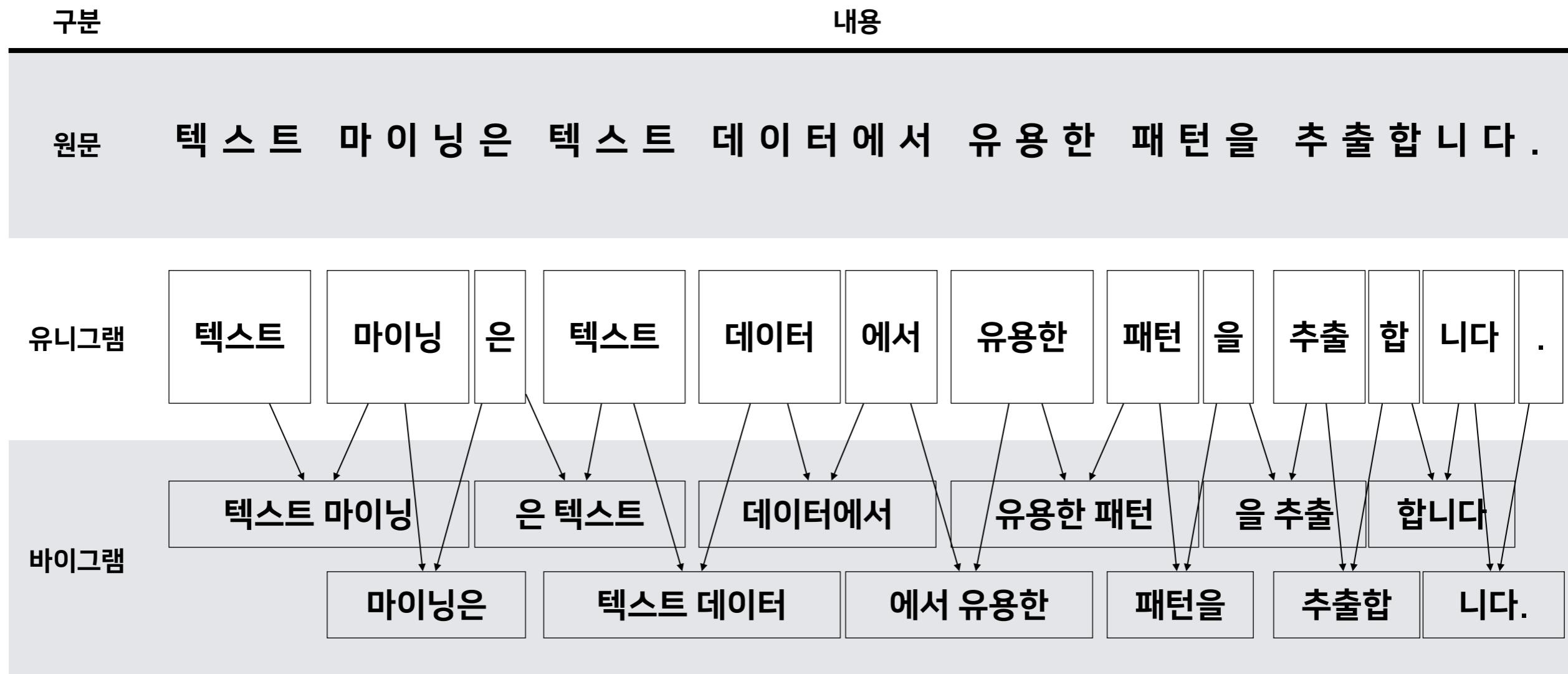
- N 개의 단어 또는 형태소를 하나의 토큰으로 취급하여 문서를 표현하는 방법
- N이 증가할 수록 텍스트 처리에 필요한 연산량이 기하급수적으로 증가하기 때문에 보통 N은 3을 넘지 않음
- 경우에 따라 불용어를 제외한 (명사+명사), (형용사+명사) 등 조합을 가진 N-gram만 사용하기도 함



문서를 표현하는 방법

유니-그램 & 바이-그램 (Uni-gram & Bi-gram)

- 유니그램 (uni-gram) : 독립된 하나의 단어 또는 형태소를 토큰으로 취급하여 문서를 표현하는 방법
- 바이그램 (bi-gram) : 두 개의 단어 또는 형태소를 토큰으로 취급하여 문서를 표현하는 방법
- N-그램에서 N이 클수록 하나의 토큰에 더 많은 정보를 내포할 수 있음



문서를 표현하는 방법

불용어 (Stopword)

- 텍스트 데이터를 분석함에 있어, 분석 결과에 출현하더라도 큰 의미가 없거나 방해되는 단어 또는 형태소
(그거, 여기, 이제, 은, 는, 이, 등, a, an, the, of, the, !, @, #, \$, %, & 등)
- 정보를 포함하기 보다는 주로 기능적인 역할(조사, 어미 등)을 하는 단어 또는 형태소에 해당하는 경우가 많음
- 일반적으로 한국어는 외자, 영어는 알파벳 두 자 단어 또는 형태소를 대부분 불용어 취급함
- 빈출어 (Common-word) : 너무 많이 출현하여 분석 결과에서 의미 또는 중요도가 떨어지는 단어
(기사, 기자, 제목, 사진, 네이버, 검색, 보다, 연기, 평점, 공감, 비공감 등)



개체명 인식

개체명 인식 (Named Entity Recognition, PLO Tagging > PLODT Tagging > NER)

- 문서에서 하나의 개체로써 인식되어야하는 단어를 추출하고 분류를 태깅하는 과정 (지명, 인물명, 회사명 등)
- 개체명 태그 유형은 응용영역에 따라서 서로 다르게 지정할 수 있음 (개체명 사전 또는 개체명 태깅 학습셋 활용)
- 정보추출(Information Extraction), 정보검색(Information Retrieval) 분야에서 대부분의 지식(Who, Where, ...)은 개체명으로 표현될 수 있음



개체명 인식

개체명 사전 (Named Entity Dictionary)

구분	의학용어	인물명	회사명
1	불량 식품	사나	오피스디포
2	진행 암	쯔위	아마존
3	전진 피판	정연	구글
4	유해 효과	나연	한국수력원자력
5	무력증	황민현	코레일
6	유산소 운동	강다니엘	애플
7	산소 호흡	옹성우	깨끗한나라
8	공기 삼킴증	전병진	금호아시아나
9	분무제	진상형	몽블랑
10	에어로졸	서지석	샤넬
11	분무 주입법	배현진	티파니
12	대기 요법	현빈	신한생명
13	정동 장애	진세연	신한카드
14	정감성	남지현	삼성생명
15	정동성	주상욱	삼성전자
16	들신경	김태희	롯데칠성음료
17	협력 병원	허맹호	롯데백화점
18	친화력	유아인	인프라웨어
19	친화 크로마토그래피	이승기	카카오
20	무섬유소원 혈증	한예슬	네이버

구분	지역명	영문 지역명	광역시/구
1	서울	Seoul	Metropolitan
2	종로	Jongno	district
3	중	Jung	district
4	용산	Yongsan	district
5	성동	Seongdong	district
6	광진	Gwangjin	district
7	동대문	Dongdaemun	district
8	중랑	Jungnang	district
9	성북	Seongbuk	district
10	강북	Gangbuk	district
11	도봉	Dobong	district
12	노원	Nowon	district
13	은평	Eunpyeong	district
14	서대문	Seodaemun	district
15	마포	Mapo	district
16	양천	Yangcheon	district
17	강서	Gangseo	district
18	구로	Guro	district
19	금천	Gumcheon	district
20	영등포	Yeongdeungpo	district

개체명 인식

개체명 인식 (Named Entity Recognition, PLO Tagging > PLODT Tagging > NER)

세계일보

현대·기아차 전기차 판매 세계 4위 쾌속질주

A17면 1단 | 기사입력 2020.10.04. 오후 8:25 | 기사원문 | 스크랩 | 본문듣기 | 설정

공감 댓글

요약본 가 드 드

2020년 들어 7월까지 6만707대 팔아
2019년比 25%↑... 테슬라 19만대로 1위



현대·기아차가 올해 7월까지 전 세계 자동차 브랜드 중 네 번째로 전기차를 많이 판매한 것으로 나타났다.

4일 자동차연구원이 SNE리서치 통계를 분석한 내용에 따르면 현대·기아차는 올해 7월 말 기준으로 6만707대의 전기차를 판매해 세계 4위를 차지했다.

세계 최대 전기차 기업 테슬라가 19만1971대로 1위를 차지했으며, 르노-닛산과 폴크스바겐이 각기 8만6189대로 7만5228대로 뒤를 이었다. 5위는 중국 BYD(4만2340대)였다.

현대·기아차는 작년 같은 기간(4만8570대)에 비해 판매량이 25% 늘어난 점이 눈에 띈다. 모델3를 앞세운 테슬라는 판매량이 4% 증가하는 데 그쳤고, 주력 모델인 리프의 판매가 급감한 르노닛산은 5% 감소했다. 폴크스바겐은 지난해에 비해 110% 늘어나며 현대·기아차를 앞질렀다. BYD는 지난 해(11만3409대)에 비해 63% 감소하며 순위가 하락했다.

개체명 인식 결과

<현대:OG>·<기아차:OG>가 올해 <7월까지:DT> 전 세계 자동차 브랜드 중 네 번째로 전기차를 많이 판매한 것으로 나타났다.

<4일:DT> 자동차연구원이 SNE리서치 통계를 분석한 내용에 따르면 <현대:OG>·<기아차:OG>는 올해 <7월:DT> 말 기준으로 <6만707대의:QT> 전기차를 판매해 세계 <4위:QT>를 차지했다.

세계 최대 전기차 기업 <테슬라:OG>가 <19만1971대:QT>로 <1위:QT>를 차지했으며, <르노-닛산:OG>과 <폴크스바겐:OG>이 각기 <8만6189대:QT>로 <7만5228대:QT>로 뒤를 이었다.

<5위:QT>는 <중국 BYD:OG>(<4만2340대:QT>)였다.

<현대:OG>·<기아차:OG>는 작년 같은 기간(<4만8570대:QT>)에 비해 판매량이 <25%:QT> 늘어난 점이 눈에 띈다.

<모델3:AF>를 앞세운 <테슬라:OG>는 판매량이 <4%:QT> 증가하는데 그쳤고, 주력 모델인 <리프:UN>의 판매가 급감한 <르노닛산:OG>은 <5%:QT> 감소했다.

<폴크스바겐:OG>은 지난해에 비해 <110%:QT> 늘어나며 <현대:OG>·<기아차:OG>를 앞질렀다. <BYD:OG>는 지난해 (<11만3409대:QT>)에 비해 <63%:QT> 감소하며 순위가 하락했다.

개체명 인식

개체명 인식 학습데이터 구성

- 1 맨유는 2일(한국시간) 영국 바이탈리티 스타디움에서 본머스와 2019-2020시즌 잉글랜드 프리미어리그(EPL) 11라운드 원정경기에서 0-1로 졌다.
2 게티이미지코리아잉글랜드 프로축구팀 맨체스터 유나이티드(이하 맨유)가 부진한 경기력과 함께 낮은 평점대를 기록했다.
3 맨유는 2일(한국시간) 영국 바이탈리티 스타디움에서 본머스와 2019-2020시즌 잉글랜드 프리미어리그(EPL) 11라운드 원정경기에서 0-1로 졌다.
4 최근 3연승 상승세를 탔던 맨유이기에 더욱 빼아픈 패배다.
5 맨유는 3승 4무 4패(승점 13점)로 한 단계 하락한 8위에 랭크됐다.
6 하지만 맨유의 순위는 다른 중위권 팀 경기 결과에 따라 최악의 경우 15위까지 떨어질 수 있다.
7 맨유는 전반 45분 요슈아 킹에게 선제골을 내주며 약 4년 만에 본머스에 무너졌다.
8 여기엔 35년 만에 본머스전 무득점이란 굴욕까지 따라왔다.
9 유럽축구 통계사이트 후스코어드닷컴은 대부분의 맨유 선수에게 평점 6점대를 부여했다.
10 최고점을 받은 프레드(7.5점)만 6점대를 피해갔다.
11 선방으로 대패를 막은 골키퍼 다비드 데헤아와 미드필더 스콧 맥토미니가 6.9점으로 뒤를 이었고, 안드레아스 페레이이라와 빅토르 린델로프가 6.1점으로 최저점을 받았다.
12 양팀 통틀어 최고점은 맨유를 격추시킨 주인공 킹(7.8점)이다.
13 이준혁 온라인 기자 2jh@kyunghyang.com▶ [스포츠경향 인기 무료만화 보기]▶ [지금 옆사람이 보고있는 뉴스]©스포츠경향(sports.khan.co.kr), 무단전재 및 재배포 금지 기사제공 스포츠경향
14 [인터풋볼] 득점력 부재에 시달리며 최근 5경기 1승 4무에 그친 아틀레티코 마드리드가 세비야 원정길을 떠난다.

- 1 <맨유:ORG>는 <2일:DAT>(<한국:LOC>시간) 영국 <바이탈리티 스타디움:LOC>에서 <본머스:ORG>와 <2019-2020시즌:DAT> <잉글랜드 프리미어리그:CVL>(<EPL:CVL>) <11라운드:QTY> 원정경기에서 <0-1:QTY>로 졌다.
2 <게티이미지코리아:ORG><잉글랜드:LOC> 프로축구팀 <맨체스터 유나이티드:ORG>(이하 <맨유:ORG>)가 부진한 경기력과 함께 낮은 평점대를 기록했다.
3 <맨유:ORG>는 <2일:DAT>(<한국:LOC>시간) <영국:LOC> <바이탈리티 스타디움:LOC>에서 <본머스:ORG>와 <2019-2020시즌:DAT> <잉글랜드 프리미어리그:CVL>(<EPL:CVL>) <11라운드:QTY> 원정경기에서 <0-1:QTY>로 졌다.
4 최근 <3연승:QTY> 상승세를 탔던 <맨유:ORG>이기에 더욱 빼아픈 패배다.
5 <맨유:ORG>는 <3승 4무 4패:QTY>(승점 <13점:QTY>)로 <한 단계:QTY> 하락한 <8위:QTY>에 랭크됐다.
6 하지만 <맨유:ORG>의 순위는 다른 중위권 팀 경기 결과에 따라 최악의 경우 <15위:QTY>까지 떨어질 수 있다.
7 <맨유:ORG>는 <전반 45분:TIM> <요슈아 킹:PER>에게 선제골을 내주며 약 <4년:QTY> 만에 <본머스:ORG>에 무너졌다.
8 여기엔 <35년:QTY> 만에 <본머스:ORG>전 무득점이란 굴욕까지 따라왔다.
9 유럽축구 통계사이트 <후스코어드닷컴:ORG>은 대부분의 <맨유:ORG> 선수에게 평점 <6점:QTY>대를 부여했다.
10 최고점을 받은 <프레드:PER>(<7.5점:QTY>)만 <6점:QTY>대를 피해갔다.
11 선방으로 대패를 <막은 골키퍼:CVL> <다비드 데헤아:PER>와 <미드필더:CVL> <스콧 맥토미니:PER>가 <6.9점:QTY>으로 뒤를 이었고, <안드레아스 페레이이라:PER>와 <빅토르 린델로프:PER>가 <6.1점으로:QTY> 최저점을 받았다.
12 양팀 통틀어 최고점은 <맨유:ORG>를 격추시킨 주인공 <킹:PER>(<7.8점:QTY>)이다.
13 <이준혁:PER> 온라인 <기자:CVL> <2jh@kyunghyang.com:TRM>▶ [<스포츠경향:ORG> 인기 무료만화 보기]▶ [지금 옆사람이 보고있는 뉴스]©<스포츠경향:ORG>(<sports.khan.co.kr:TRM>), 무단전재 및 재배포 금지 기사제공 <스포츠경향:ORG>
14 [<인터풋볼:ORG>] 득점력 부재에 시달리며 최근 <5경기:QTY> <1승:QTY> <4무:QTY>에 그친 <아틀레티코 마드리드:ORG>가 <세비야:ORG> 원정길을 떠난다.

순차 레이블링: RNN

Review

RNN을 활용한 순차 레이블링

구분

내용 (B: Beginner, I: Inner, O: Outer)

원문 손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홫 스 퍼 의 승 리 를 이 꼴 었 다 .



E.O.D

Contact

 <http://www.teanaps.com>

 fingeredman@gmail.com