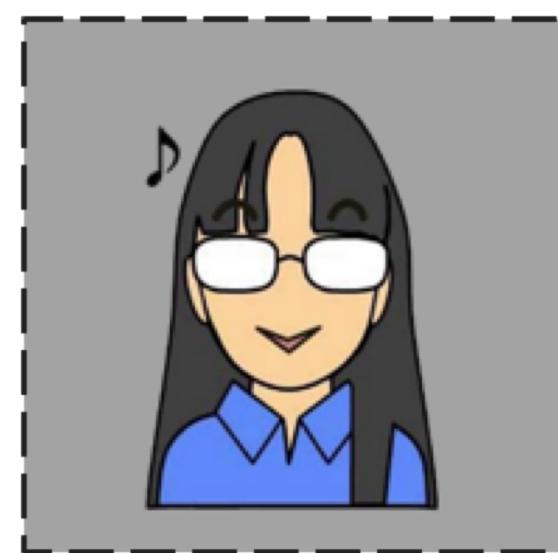


Why I learn, How I learn

어학 전공자가 새로운 영역에 접근하는 방법

송영숙(*<https://songys.github.io/#portfolio>*)



10/2014

자연어처리 스터디



바벨피쉬

9/2015

경희대학교 박사과정

1/2017

Company.A.I.

NAVER 입사

NAVER

9/2017

NER 부문 참가

S_MCH

NP_OBJ

VP

NNP+副/JKO

AP

명사어/NNG

S_VPO

VP

이동/Ah+副/EP+V2

VP



9/2018

의존구문분석 참가중

Why I learn

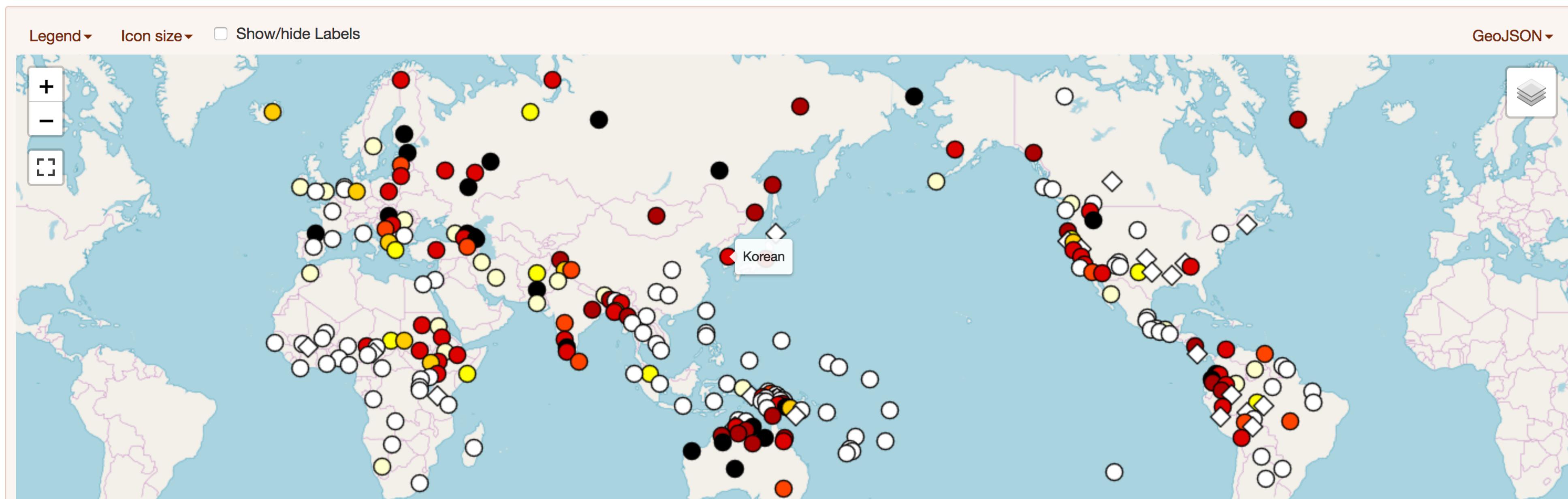
Feature 49A: Number of Cases

This feature is described in the text of chapter 49 [Number of Cases](#) by Oliver A. Iggesen [cite](#)

You may combine this feature with another one. Start typing the feature name or number in the field below.

Values

	No morphological case-marking	100
	2 cases	23
	3 cases	9
	4 cases	9
	5 cases	12
	6-7 cases	37
	8-9 cases	23
	10 or more cases	24
	Exclusively borderline case-marking	24

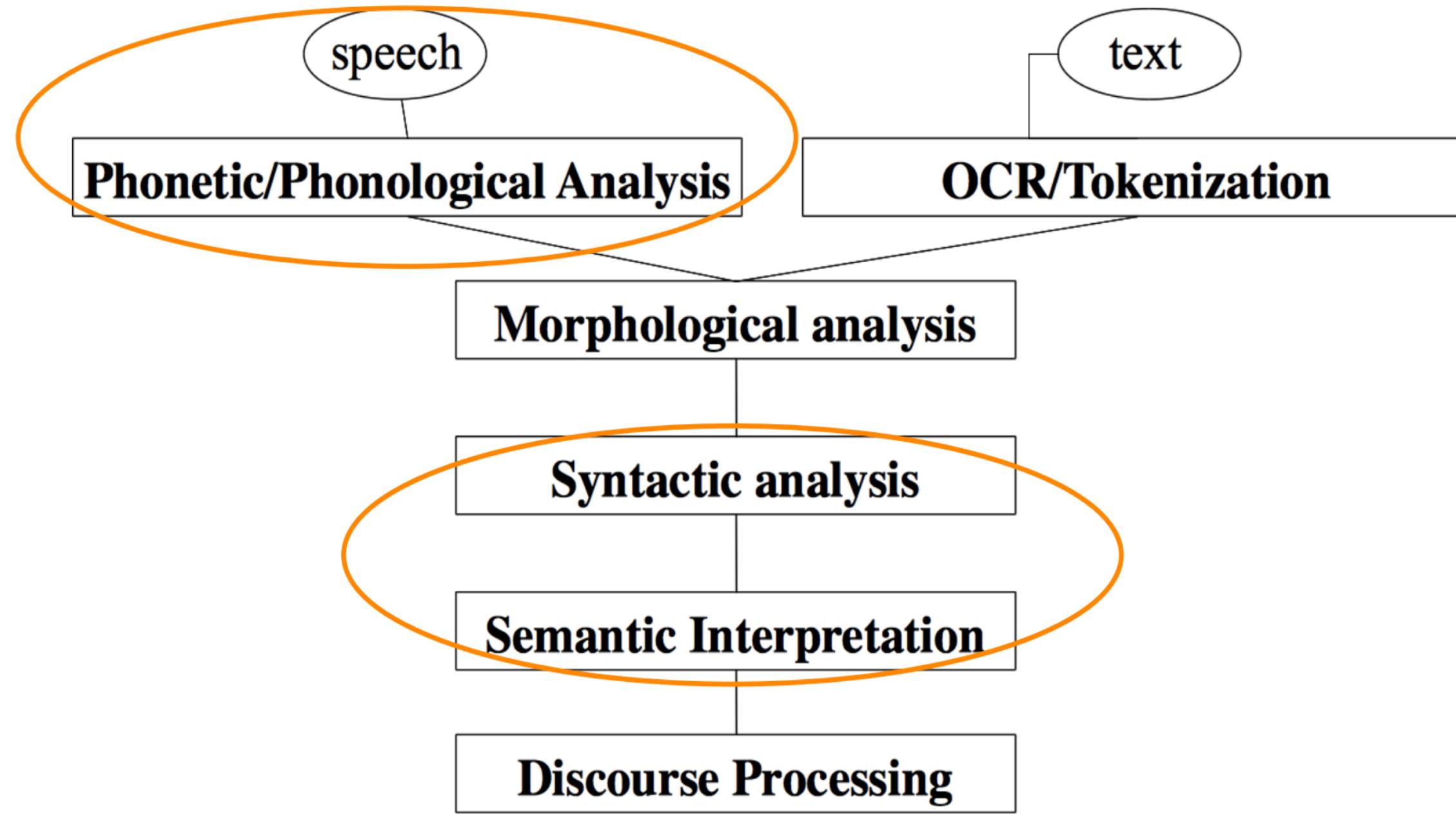


- 국어학 자료의 연대 추정
- 고전 번역 기간의 단축
- 특정 작가, 특정 시대의 문체로 창작하는 인공지능

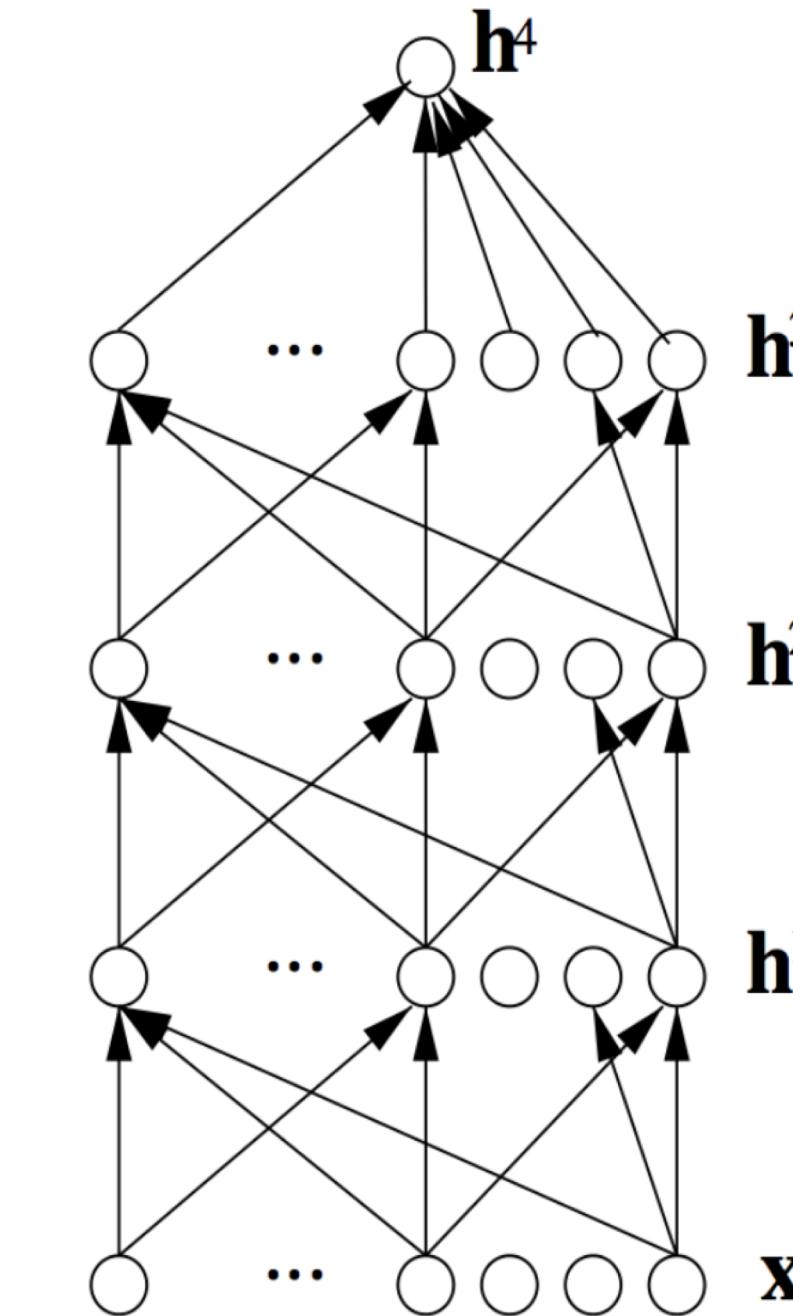
구분	승정원일기번역(1994년 ~)	조선왕조실록현대화(2012년 ~)	일성록번역(1998년 ~)
번역대상	2386권	574권	515권
진척률(%)	20.9	10.8	44.9
2017년 사업량	46권	13권	18권
완역 예상기간 (년)	41	40	16

How I learn

자연어 처리의 일반적 과정



딥러닝 방법



“raw” inputs x (소리, 글자, 단어 등)

기계 번역

음성 인식

감성 분석

대화 시스템



데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가



**데이터 수집과 정제에서 고생한 만큼
모델링에서 좋은 결과를 얻을 수 있습니다.**

데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가

C	D	E
형태소분석 후 (모델 입력)	label	label(english)
비행 청소년 급증	0	There are words that can not be translated.
사회 문제 되다 비행 청소년	0	There are words that can not be translated.
뉴욕행 비행기 예약 하다 싶다	1	I'd like to book a flight to New York
항공권 예약 하다	1	I'd like to book a flight to New York
비행기 취소 하다	2	I need to cancel my flight.
티켓 취소 하다	2	I need to cancel my flight.

데이터의 왜곡이 없다면 한 문장이라도 학습을 시작할 수 있습니다.

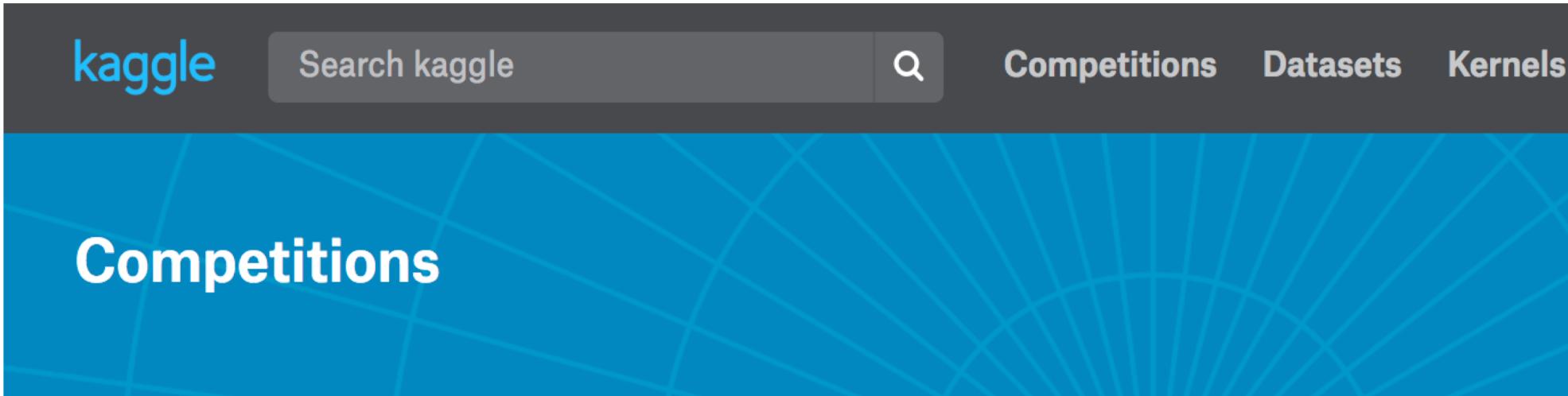
데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가



Lecture	Description	kaggle	code&data	etc
1	word2vec nlp tutorial	위드투 백 튜토 리얼	English_data code1_description code2_More_fun.ipynb	python 3 으로 작업
2	sentiment analysis on movie reviews	영화평 감성분 석	English_data Korean_data English_code Korean_code	긍정과 부정으로 레이블
3	jigsaw toxic comment classification challenge	약평분 류	English_data Korean_data English_code Korean_code	toxic, severe_toxic, obscene, threat, insult, identity로 레이블 되어 있는 위키피디아 데이터를 분류하는 대회입니다. 한국어는 Naver sentiment movie corpus에서 약 2만개를 핸드 레이블로 toxic, obscene, threat, insult, identity로 5 단계로 분류하였습니다.
4	quora- question- pairs	두 개의 질문이 같은가, 다른가	English_data Korean_data English_code korean_code	네이버 지식인과 비슷한 사이트인 Quora에 올라오는 질문들을 두 개씩 비교해서 같은 질문인지 아닌지 맞추는 대회입니다. 한 국어 데이터는 6888개의 train 데이터와 688개 test data를 질문 페어로 만들어서 분석했습니다.

그런데 국어는?!

없는 데이터만 만들어서
분석했습니다!

<https://www.kaggle.com/>

데이터 1 : 영화평에 나오는 악평 레이블링

데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가

Naver sentiment movie corpus v1.0_감성 분석에 부정적 감성의 레이블링 상세화
label : 'toxic', 'obscene', 'threat', 'insult', 'identity_hate'

Data description

train : 9,999 test : 9,999

		id		document	toxic	obscene	threat	insult	identity_hate
9994	7448293	혹시나 그래도 카메룬디아즈니까 하고봤는데...먼이런영화를..이도저도아닌 .아암튼 시...			1	0	0	0	0
9995	5824024		10점주는것들 한국영화는 1점주네. M창		1	0	0	0	0
9996	6420437		영상도 아름답고 뭘 말하는지 알겠지만 그렇기 때문에 짜증나고 답답하다		1	0	0	0	0
9997	6777278		영화를 왜 영화라고 하는지 모르는 애들이 나왔네.		1	0	0	0	0
9998	9238637	내가 이걸보고 며칠내내 화가났네 별이 아깝 어쩜이리쓰래기가... 내용잡다한거 다섞여...			1	0	0	0	0

https://github.com/songys/Toxic_comment_data

데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가

kor_Pair_test.csv kor_pair_train.csv**train 6,888 개 test 688개****label****같은 질문 0****다른 질문 1**

id	qid1	qid2	question1	question2	is_duplicate
1	1	2	1000일 만난 여자친구 와 이별	10년 연애의 끝	1
2	3	4	10년 연애사 되돌아보 니 다 부질없네	10년이라는 시간이 참 무색하다 싶네	1
1451	2901	2902	다리가 덜덜 떨려	다리가 후덜덜	0
1452	2903	2904	다리가 통통 부었어	다리가 부은 거 같아	0
1453	2905	2906	다시 꼭 잡겠다는 결심	다시 만나자는 말 하기 전까지 연락 안 하게	1

데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가

Chatbot_data_for_Korean v1.0

챗봇 트레이닝용 문답 페어 11,876개

중립 0, 이별(부정) 1, 사랑(긍정) 2로 레이블링

https://github.com/songys/Chatbot_data

1	그만 울어야지	툭툭 털고 일어나세요.	1
2	그만 힘들어하구 싶다	그랬었지 하고 추억하는 날이 올 거예요.	1
3	그만 울어야지	툭툭 털고 일어나세요.	1
4	그만 힘들어하구 싶다	그랬었지 하고 추억하는 날이 올 거예요.	1
5	3박4일 놀러가고 싶다	여행은 언제나 좋죠.	0
6	3박4일 정도 놀러가고 싶다	여행은 언제나 좋죠.	0
7	PPL 심하네	눈살이 찌푸려지죠.	0

데이터 수집

데이터 정제

모델링

파라미터 튜닝

성능평가



Archived: CS224n: Natural Language Processing with Deep Learning (2016-2017)

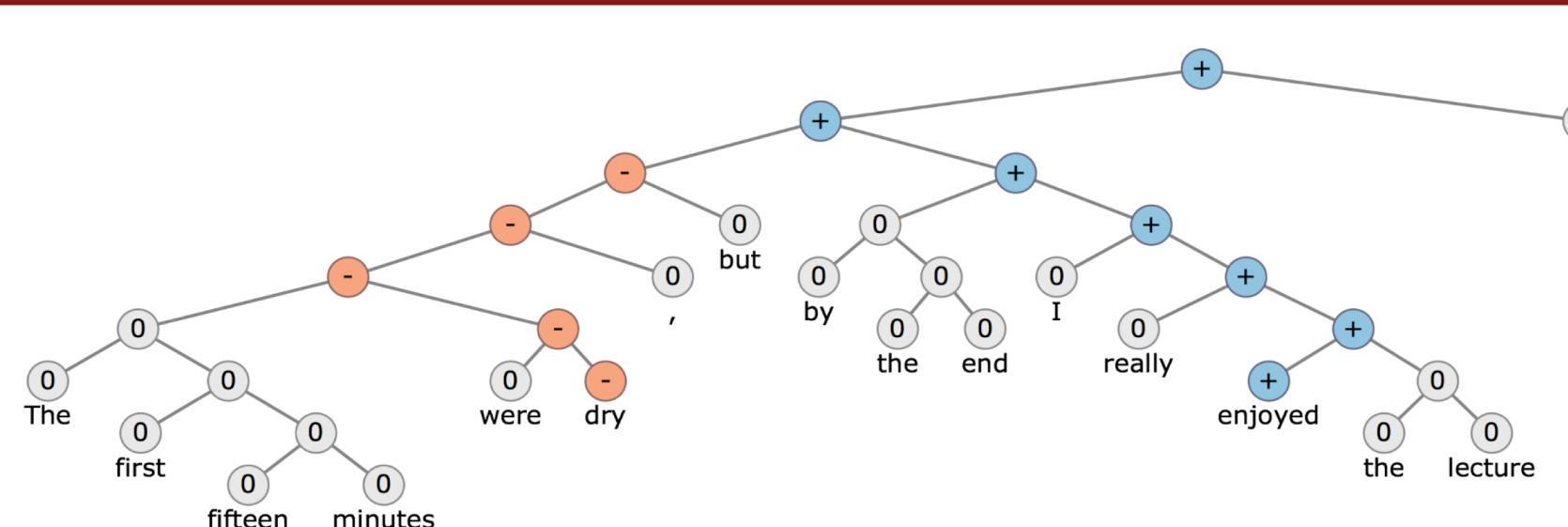


Table of Contents (CS224d 2017)

Lecture(keras_code)	Description	Videos(youtube)	Slides	Subtitles
1	Intro to NLP and Deep Learning	video	slide	english korean
2	Word Vector Representations: word2vec	video	slide	english korean
3 code	GloVe: Global Vectors for Word Representation	video	slide	english korean
4	Word Window Classification and Neural Networks	video	slide	english korean
5	Backpropagation and Project Advice	video	slide	english [korean]
6	Dependency Parsing	video	slide	english korean

- 선택 기준 : 자연어 처리 이론으로 강의영상과 강의 자료가 모두 잘 올라와 있는 최근 자료
- 대상 : CS224n: Natural Language Processing with Deep Learning (2016-2017)
- 번역 작업 : 강좌 당 80분 정도, 약 1,000라인. 지금 5강좌 약 5,000라인 번역

https://github.com/songys/CS224N_17_KOR_SUB

Pain Points

자연어 처리가 어려운 이유

(1) 긴 명사

- 니코틴아마이드 아데닌 다이뉴클레오타이드가뭐야 ?

(2) 잘못 된 띄어쓰기 또는 띄어쓰기로 해결할 수 없는 명사

- 아버지 가방에 들어가신다.

(3) 모호한 문장

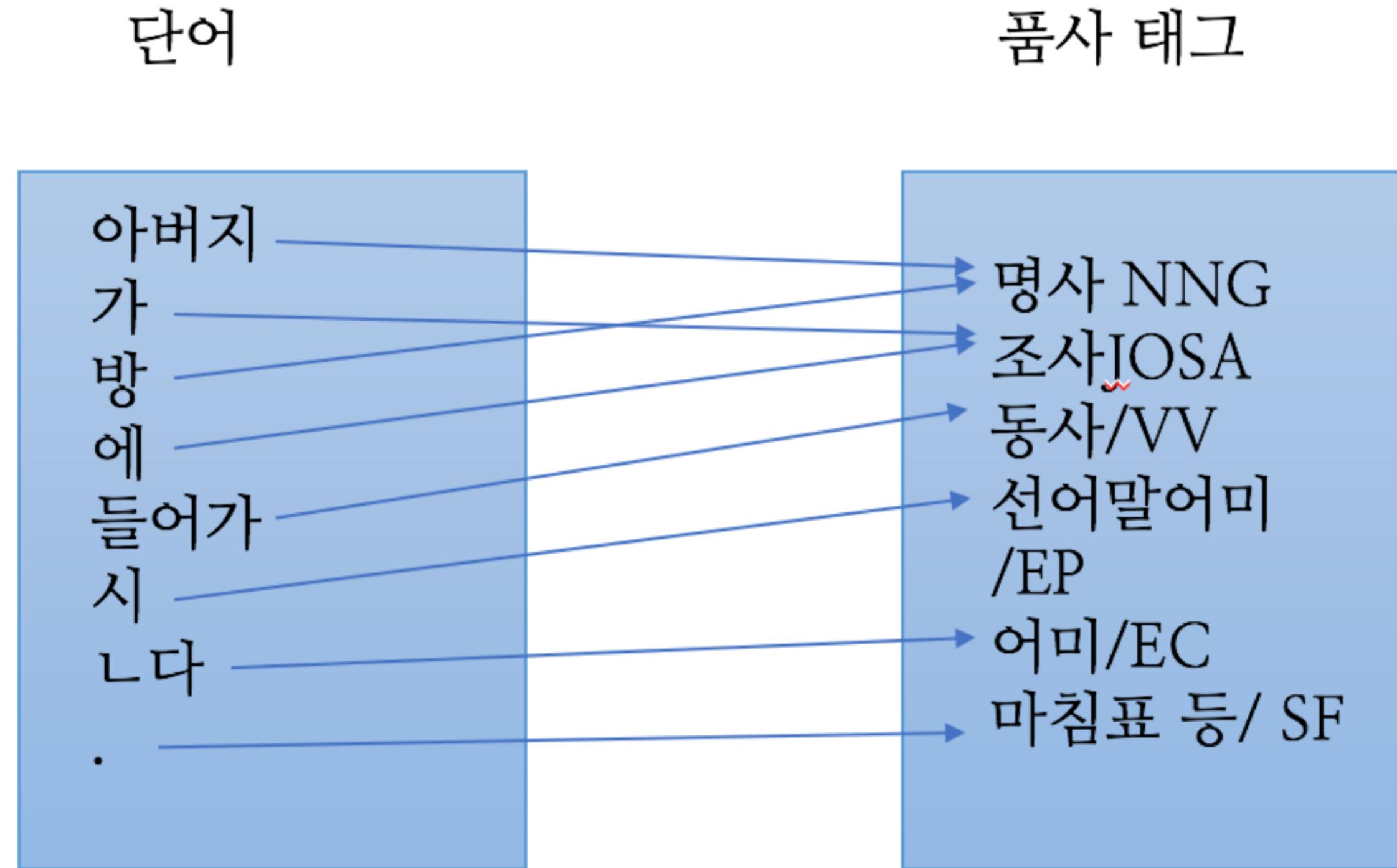
- 철수와 순이의 결혼식 / 거만한 친구의 남편

(4) 중심이 없이 긴 문장

- 게시물 신고하려고 하는데 이상한건 아니고 그냥 문제가 있는 것 같아서 그러는데 어떻게 신고할 수 있어?

(5) 짧은 순간 발생했다 사라지는 비속어와 유행어

- 옛총, ㄷ ㄷ ㄷ



1) 의미어 : 아버지, 방, 들어가- 는 의미 분석하고 싶은 의미어

2) 기능어 : '-가, -에, -시-, -ㄴ다' 의미가 잘 전달되도록 도와주는 기능어로 국어에 기능어는 의미어에 붙어 있는 형태로 되어 있다. (교착어)

3) 띠어쓰기가 안 되어 있는 경우 의미적 중의성 발생

<예> ㄱ. 아버지가 방에 들어가신다.

ㄴ. 아버지 가방에 들어가신다.

종결	빈도	비율
어	6117	13.71%
야	4429	9.93%
지	3953	8.86%
아	3545	7.94%
다	2245	5.03%
ㅓ	1920	4.30%
ㅋ	1871	4.19%
ㅏ	1802	4.04%
어요	1005	2.25%
냐	994	2.23%
애	823	1.84%
아요	766	1.72%
거든	738	1.65%
네	671	1.50%

21세기 세종 계획 구어 전사
말뭉치(EF로 태깅) 172만 어절 중
형태가 다른 종결 어미 종류 472개 !!

‘맛있다’ 라고만 말하고 ‘맛있당’이라고
말하지 말라고 할 수도 없는데 ….

구문 : 문장 단위로 의미를 구분하고자 할 때
가장 많이 쓰는 방법

의존문법 : 어느 요소가 어느 요소에
의존하는지와 이들 의존 관계의 유형은

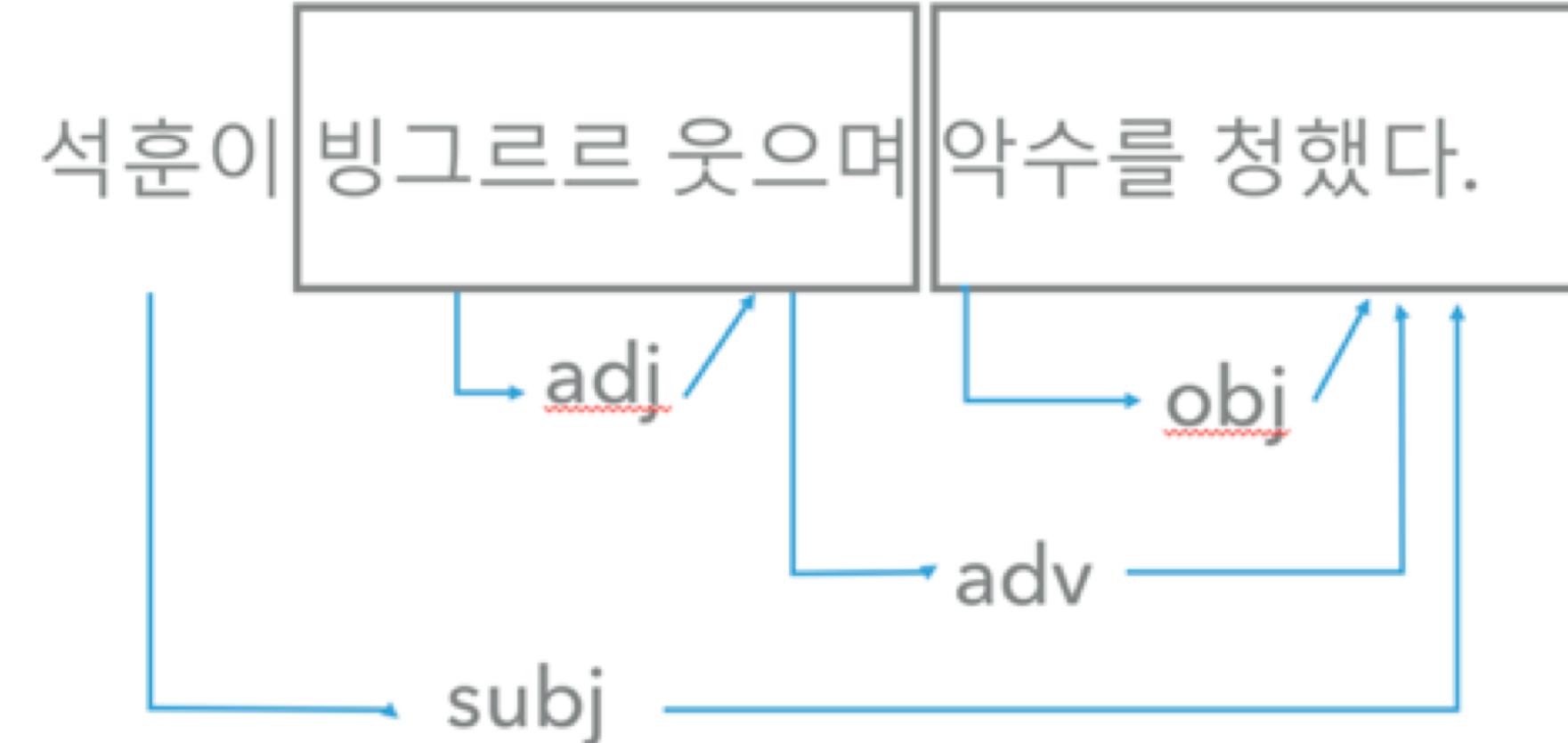
무엇인지를 확정하는 작업

- 각각의 단위를 나누어서 화살로 의존 관계를
나타냄

- ML, 규칙기반 모두 사용 가능

파서의 종류 : 기본적으로는 모두 의존 문법
기반의 파서이고 규모, 최적화 등에서

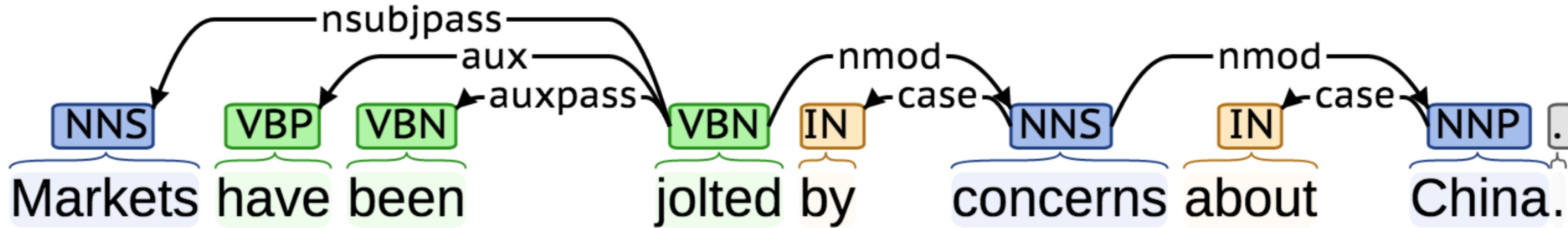
차이가 있는 파서들이 다수 존재



subj : 주어, obj : 목적어, adj: 관형어, adv: 부사어

박진호(2004), 한국어 정보화와 구문분석, 월인

스탠포드 파서(<https://stanfordnlp.github.io/CoreNLP/>)



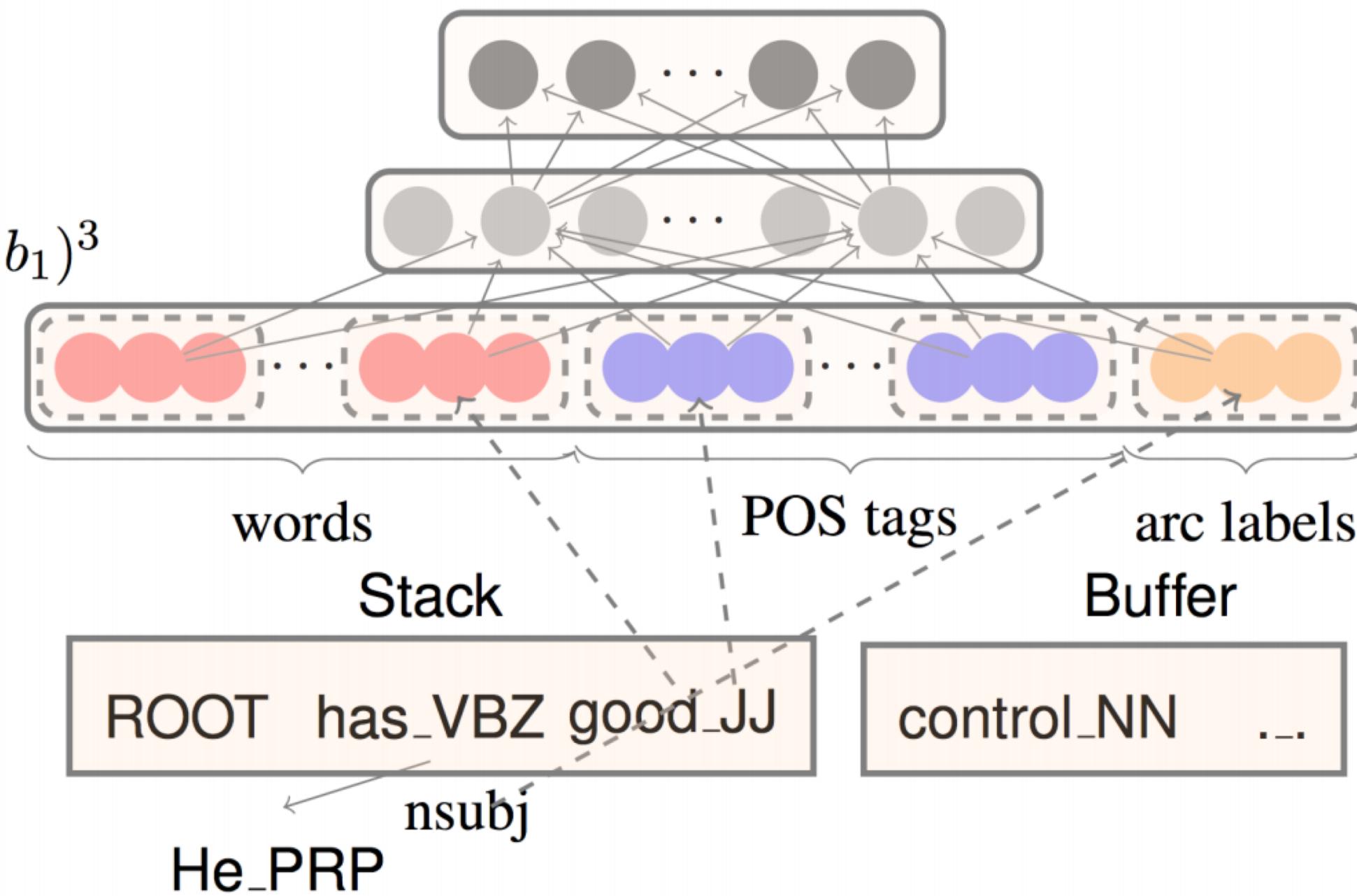
Softmax layer:

$$p = \text{softmax}(W_2 h)$$

Hidden layer:

$$h = (W_1^w x^w + W_1^t x^t + W_1^l x^l + b_1)^3$$

Input layer: $[x^w, x^t, x^l]$



One More

추천 홈페이지

초급용NLP

<https://tacademy.sktechx.com/live/player/onlineLectureDetail.action?seq=123>

중급용NLP

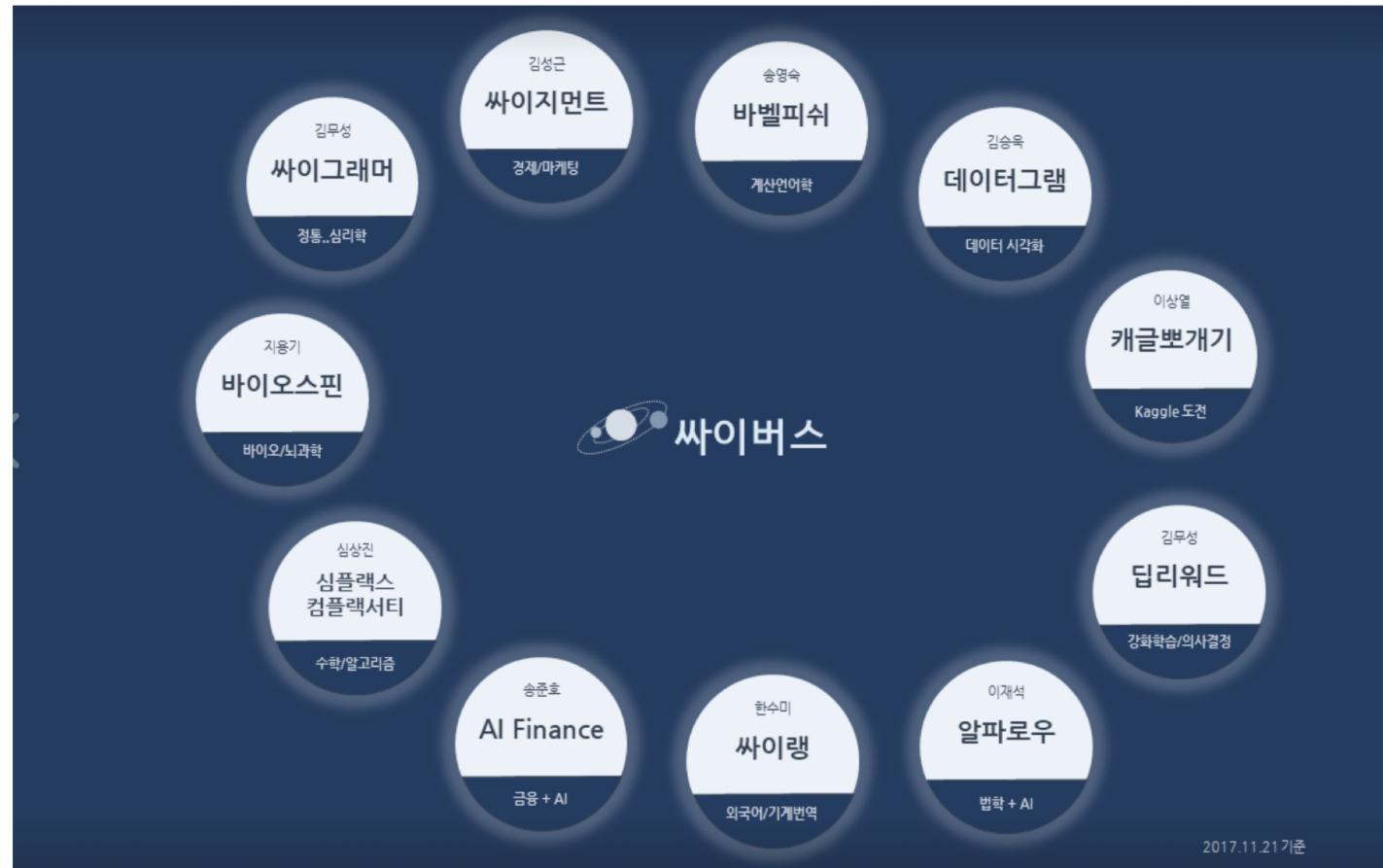
<https://www.coursera.org/learn/language-processing/>

<https://www.edwith.org/>



상시 수강
조경현 교수 | edwith

추천 커뮤니티



GDG Korea
gdgkr.slack.com

스터디를 원하시면 싸이버스에 진행 중인 스터디 소개 참고:

https://docs.google.com/spreadsheets/u/1/d/1kDHaZ4rr3tVsLXx0_WCFSAlv_-rcxsxdYe-05MCPz4/edit#gid=1745243154

모여서 각자 코팅하는 PyJog에 참여를 원하시면 파이조그 페이지 이벤트 참고 :

<https://www.facebook.com/pyjog/?fref=mentions>

Google Developer Group Korea 참여는 <https://www.facebook.com/gdgkorea/> 또는 슬랙을 통해 meet up 또는 스터디, 네크워킹을 겸한 밥모임

감사합니다

