



BOAZ 공동세션

구어체 텍스트 데이터와 토픽 모델링

2018.09.20

김태희



BOAZ 공동세션

~~구어체 텍스트 데이터와 토픽 모델링~~

본격 자연어 처리(NLP) 홍보 발표

2018.09.20

김태희



같이 NLP 공부하실 분 찾습니다~

자연어 처리?

Natural Language Process?

자연어 처리가 뭘까...

인터넷 좀 찾아볼까?

오... 이미지를 설명하네?



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

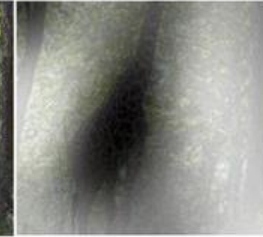


Image Caption Generation [Xu et al., 2015]

이미지 캡션 생성(Image Caption Generation)

영상도 설명하네?



+Local+Global: A **man** and a **woman** are **talking** on the **road**

Ref: A man and a woman ride a motorcycle



+Local+Global: **Someone** is **frying** a **fish** in a **pot**

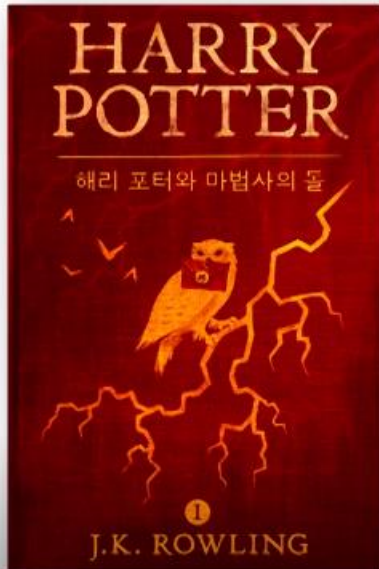
Ref: A woman is frying food

Video Caption Generation [Li et al., 2015]

영상 캡션 생성(Video Caption Generation)

해리포터도 읽어준다고?

바쁘면 31분부터



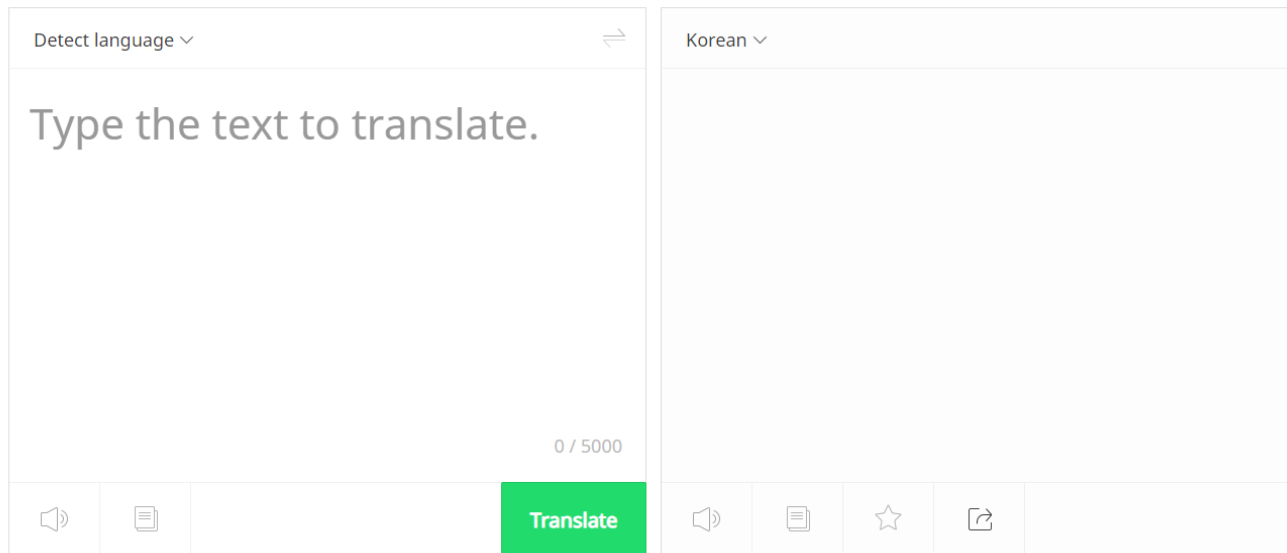
그것은 은빛 라이터처럼 보였다.

<https://github.com/carpedm20/multi-speaker-tacotron-tensorflow>

음성 합성(Speech Synthesis)

영어로 과제를 해야 한다...?

파파고 쓰세요 두 번 쓰세요



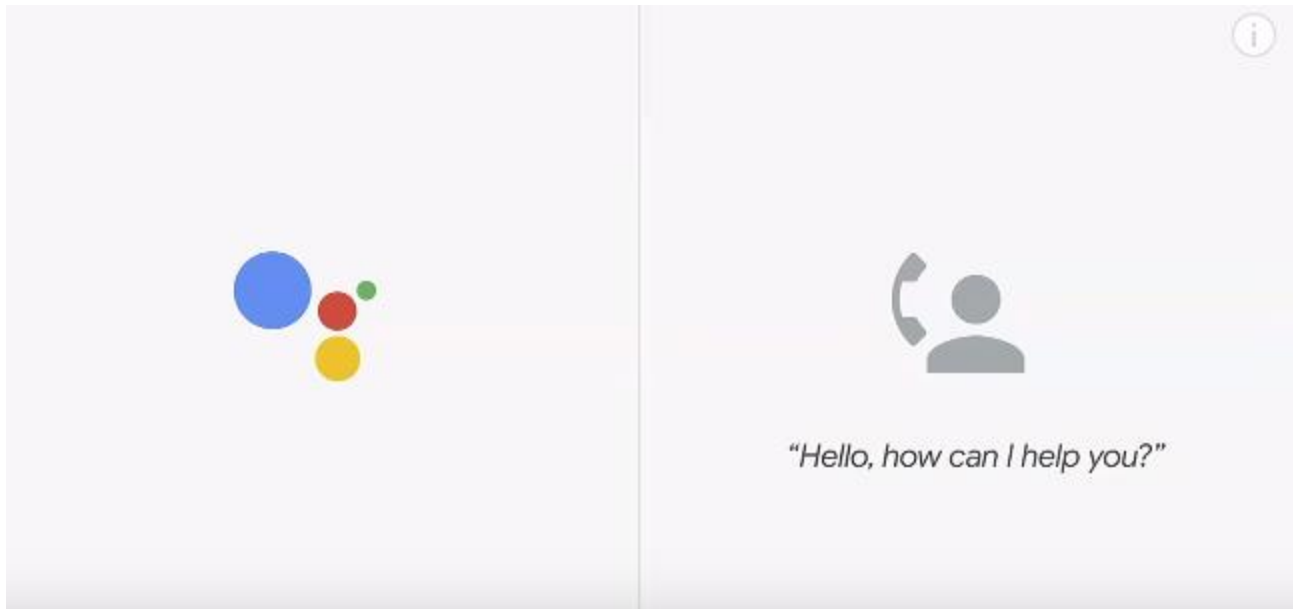
The image shows a screenshot of the Papago web interface. On the left, there is a text input area with the placeholder text "Type the text to translate." and a character count "0 / 5000". Above the input area is a dropdown menu labeled "Detect language". To the right of the input area is a dropdown menu labeled "Korean". Below the input area is a green button labeled "Translate". There are also icons for voice input, document input, and other functions.

<https://papago.naver.com/>

기계번역(Machine Translation)

구글은 미용실 예약도 합니다

바쁘면 35분 45초부터



<https://www.youtube.com/watch?v=ogfYd705cRs>

인공지능(Google Assistant)

삼성도 당연히 합니다

**'Bixby' 를
누르세요**

'Bixby Home 바로가기' 누르고
'Bixby에게 말하기' 길게 누르세요.



근데 잘 안돼요... (아직) 쓰지 마세요

언젠가는 잘 될 거예요...

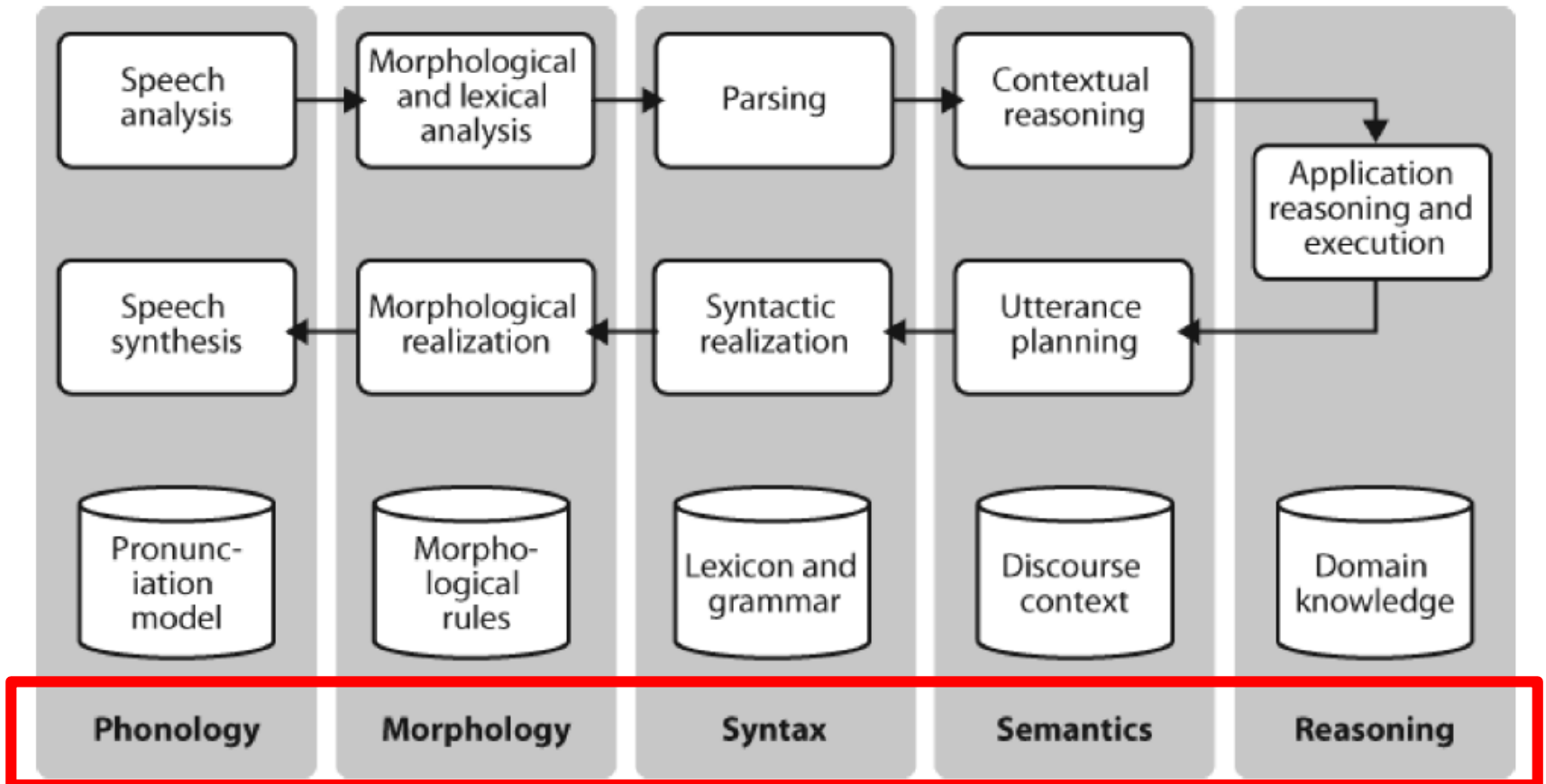
이외에도 문법 검사, 문서 요약

토픽 모델링, 감성 분석 등등... 많습니다

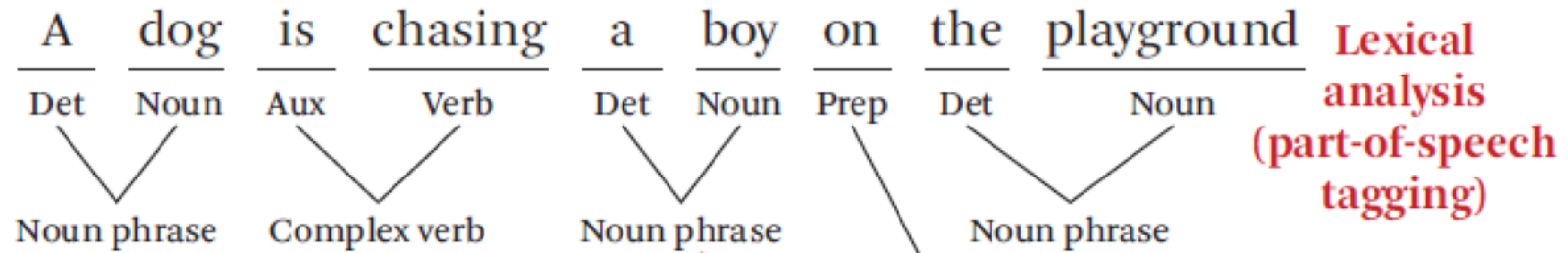
아무튼 신기하네...

그래서 어떻게 하는건데???

음운 → 형태 → 통사(구문) → 의미



.....???



Semantic analysis

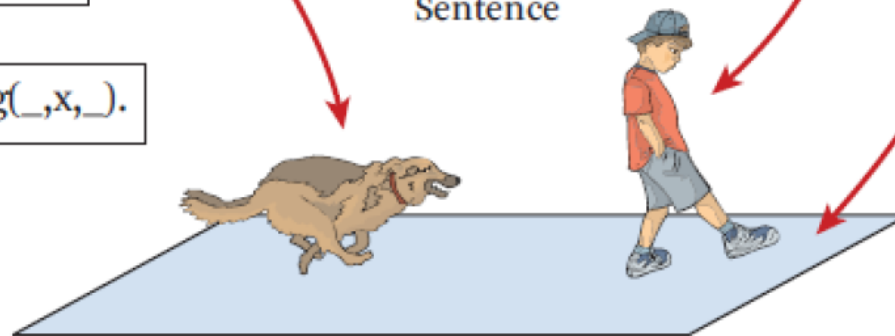
Dog (d1).
Boy (b1).
Playground (p1).
Chasing (d1, b1, p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

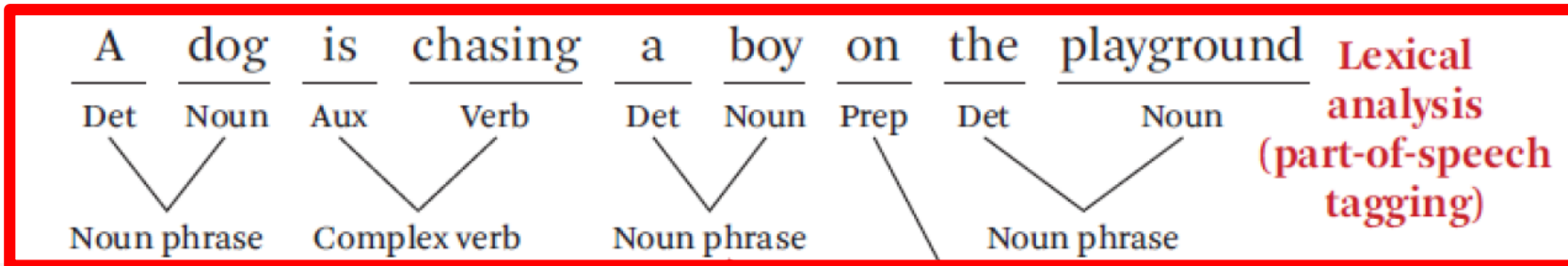


Syntactic analysis
(parsing)

A person saying this may be reminding another person to get the dog back.

Pragmatic analysis
(speech act)

일단 여기 !



Semantic analysis

Dog (d1).
Boy (b1).
Playground (p1).
Chasing (d1, b1, p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

명사, 동사, 관형사, 전치사 등에 대한 구분 !

Syntactic analysis (parsing)



A person saying this may be reminding another person to get the dog back.

Pragmatic analysis (speech act)

형태소 분석(POS-tagging)

원시 말뭉치를 형태소 단위로 쪼개고 각 형태소에 품사 정보를 부착하는 작업

A dog is chasing a boy on the playground

↗
넌 명사

↗
넌 관형사

↗
넌 전치사

언어학... 어렵다...

오픈소스로 공개된 형태소 분석기가 많이 있습니다.

꼬꼬마, 코모란, 트위터, 한나눔, 온전한늬 등...



무료 공개 감사합니다

꼬꼬마 형태소 분석기 예시

영화가 아니라 영상으로 예술을 만든 작품 이다지도 소박한 주제에 숨이 막힐 듯 한 우아함이라니



영화/NNG, 가/JKC, 아니/VCN, 라/ECD, 영상/NNG, 으로/JKM,
예술/NNG, 을/JKO, 만들/VV, ㄴ/ETD, 작품/NNG, 이/VCP,
다/ECS, 지도/NNG, 소박/NNG, 하/XSV, ㄴ/ETD, 주제/NNG,
에/JKM, 숨/NNG, 이/JKS, 막히/VV, ㄹ/ETD, 듯/NNB, 한/MDN,
우/NNG, 아함/NNP, 이/VCP, 라니/EFQ

분석기 성능은 자료에 따라 서로 다를 수 있어요!

그다음 단어를 숫자로 바꿔야 하는데...

나는 발표 중이다.

내용이 점점 산으로 가는 것 같다.

재밌는 발표 해야 하는데 솔직히 더 들어가면 재미가 없어진다.

강민이랑 연식이가 짜려볼까 무섭다.

몇 번 나왔는지 세어보자 (Count-based methods)

Doc 1: 나는 밥을 먹었다.
Doc 2: 나는 잠을 잤다.
Doc 3: 나는 배가 고프다.

	Doc 1	Doc 2	Doc 3
나	1	1	1
는	1	1	1
밥	1	0	0
을	1	1	0
먹-	1	0	0
쌌	1	1	0
다	1	1	1
잠	0	1	0
자	0	1	0

⋮

문서-단어 행렬(Document-Term Matrix)

문서 - 단어 행렬로 해볼 수 있는 분석은?

LSA, LDA !

잠재 의미 분석 **(LSA, Latent Semantic Analysis)**





문서 - 단어 행렬에 0이 너무 많으니까 좀 줄여볼까?

잠재 디리클레 할당 **(LDA, Latent Dirichlet Allocation)**

토픽~단어, 토픽~문서 분포에서 단어가 만들어질거야

일단 표시를 해보자(One-hot vector)




보아즈 여러분 안녕하세요. 저는 지금 구어체 텍스트와 토픽 모델링이라는 주제로 발표하고 있는 김태희입니다. 파파고 번역기의 성능은 아주 좋습니다. 영작을 잘 못해도 여기에다가 한국말로 집어넣으면 알아서 이렇게 변환해주니까 그냥 갖다가 쓰면 좋겠죠?

보아즈		[1,0,0,0]
주제		[0,1,0,0]
토픽		[0,0,1,0]
성능		[0,0,0,1]

One-hot vector로
바꿨더니 아주 큰 문제가
발생했다




전부 같은 거리에 있네...

의미 구분이 안된다 !

보아즈		$[1,0,0,0]$	← 다른 단어
주제		$[0,1,0,0]$	↙ 비슷한 단어 ↘
토픽		$[0,0,1,0]$	

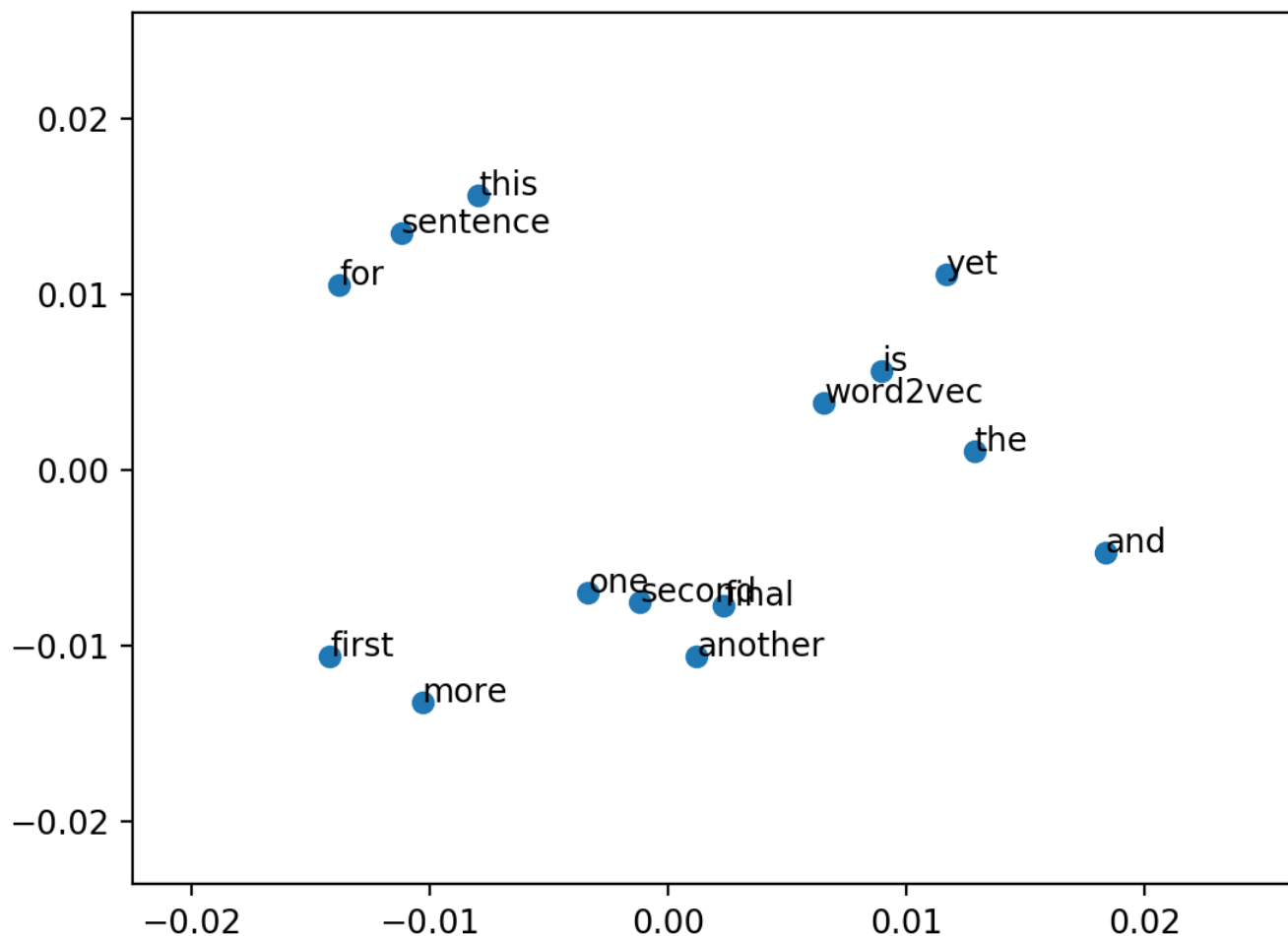
의미를 분산해서 나타내 보자

$$[0 \quad 0 \quad 0 \quad \textcolor{green}{1} \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \textcolor{green}{10} & \textcolor{green}{12} & \textcolor{green}{19} \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

보아즈		$[1, 0, 0, 0.7]$	← 다른 단어
주제		$[0.3, 1, 0.5, 0]$	↙ 비슷한 단어
토픽		$[0.5, 0.7, 1, 0]$	↘

의미 구분이 된다 !

그림으로도 그릴 수 있다 ! (feat. PCA)

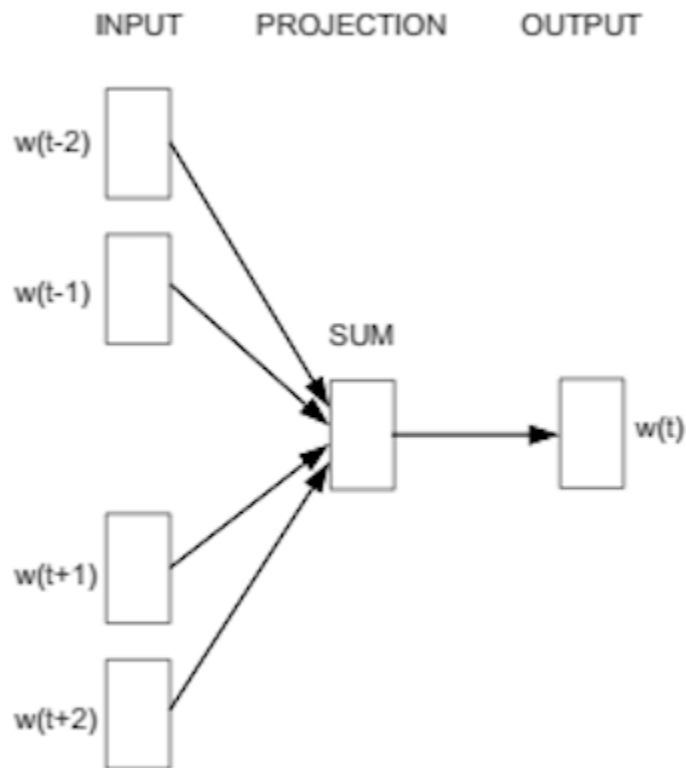


무엇을 기준으로 단어 벡터를 분산 표현할까?

단어의 동시 등장 정보 !
feat. Co-occurrence matrix

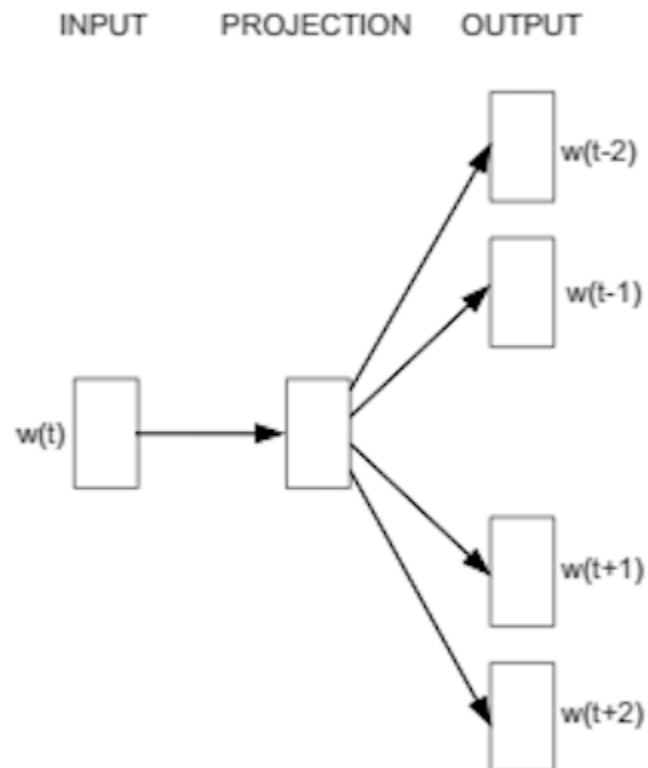
Word2vec, Glove, Fasttext
다 이거 씁니다

Word2vec만 잠깐 봅시다



CBOW

주변단어는 중심 단어의 벡터와
가까워 지도록 weight를 크게



Skip-gram

중심단어 주변에 안 나오는
단어의 weight는 작게

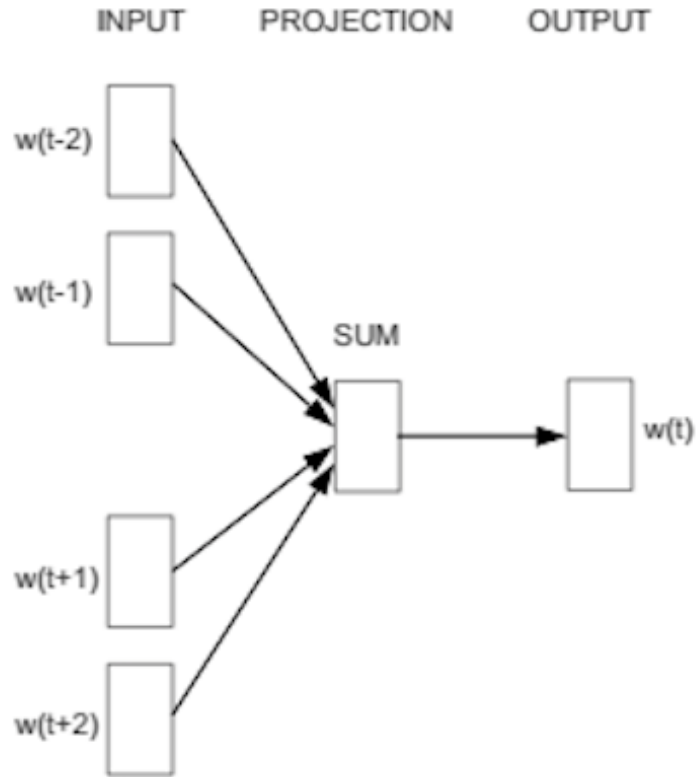
Skip-gram, Window size = 2인 경우

Source Text

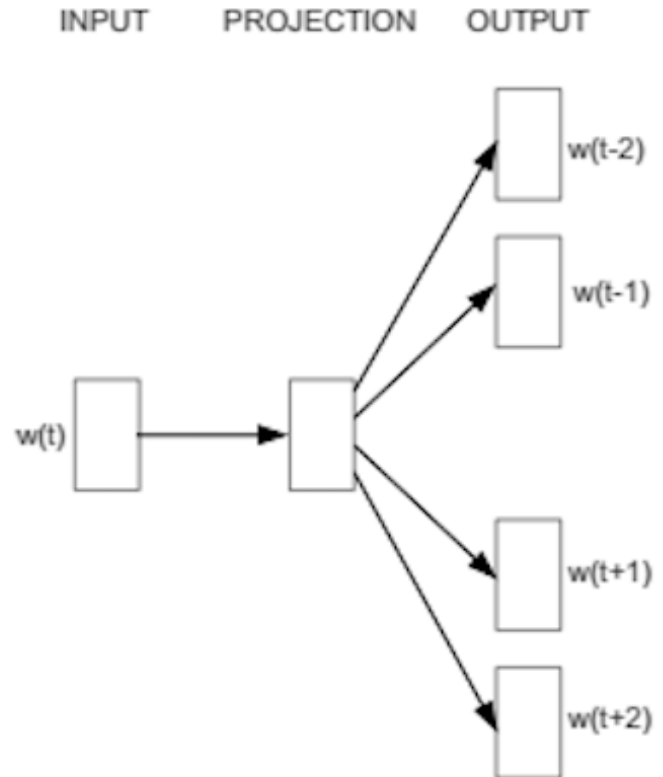
Training
Samples

The quick brown fox jumps over the lazy dog. →	(the, quick) (the, brown)
The quick brown fox jumps over the lazy dog. →	(quick, the) (quick, brown) (quick, fox)
The quick brown fox jumps over the lazy dog. →	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown fox jumps over the lazy dog. →	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

CBOW보단 Skip-gram이 좀 더 좋아요 왜 일까요?



CBOW



Skip-gram

Glove

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k ice)/P(k steam)$	8.9	8.5×10^{-2}	1.36	0.96

단어 동시 등장 행렬 구해서 단어끼리 서로 내적하면 전체 말뭉치의 동시등장 확률이 되게 하자 !

Fasttext

해
달

(a) chosung

해
달

(b) joongsung

해
달

(c) jongsung

$\{<, \square, \vdash\}, \{\dot{\vdash}, \neg, \circ\}, \{\neg, \circ, \vdash\},$
 $\{\text{ㅍ}, \sqsubset, \vdash\}, \{\vdash, \text{ㅍ}, \sqsubset\}, \{\vdash, e, >\}$

단어 단위 말고 철자 단위로 잘라서 단어 벡터 표현 해보면 어때?

Glove

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k ice)/P(k steam)$	8.9	8.5×10^{-2}	1.36	0.96

단어 동시 등장 행렬 구해서 단어끼리 서로 내적하면 전체
말뭉치의 동시등장 확률이 되게 하자 !

Fasttext

해
달

(a) chosung

해
달

(b) joongsung

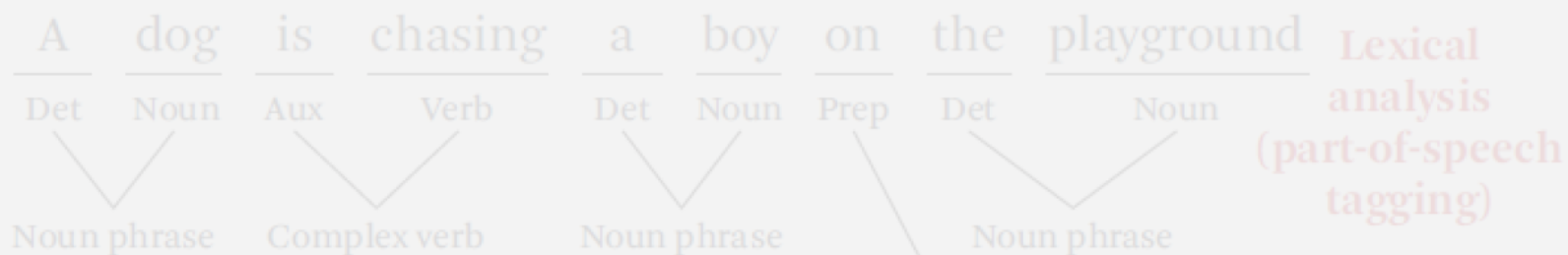
해
달

(c) jongsung

$\{<, \square, \vdash\}, \{\dot{\vdash}, \neg, \circ\}, \{\neg, \circ, \vdash\},$
 $\{\text{ㅍ}, \sqsubset, \vdash\}, \{\vdash, \text{ㅍ}, \sqsubset\}, \{\vdash, e, >\}$

단어 단위 말고 철자 단위로 잘라서 단어 벡터 표현 해보면 어때?

여기서부터 잘 몰라요 저도...



Semantic analysis

Dog (d1).
Boy (b1).
Playground (p1).
Chasing (d1, b1, p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

Verb phrase

Prep phrase

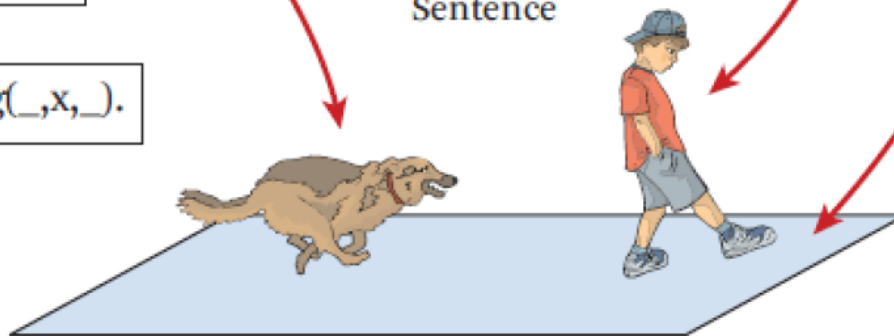
Verb phrase

Sentence

Syntactic analysis
(parsing)

A person saying this may be reminding another person to get the dog back.

Pragmatic analysis
(speech act)



같이 공부하실 분 찾아요!

이제 LSA, LDA 실습해봅시다