Embedding 기법의 활용

* Word2Vec와 Doc2Vecc

임베딩?

- 자연어처리에서 많이 활용
- 사람이 사용하는 자연어의 단어를 수치화하여 벡터로 표현한 결과가 그 과정
- 다양한 방법이 사용

워드2벡터(Word2Vec)

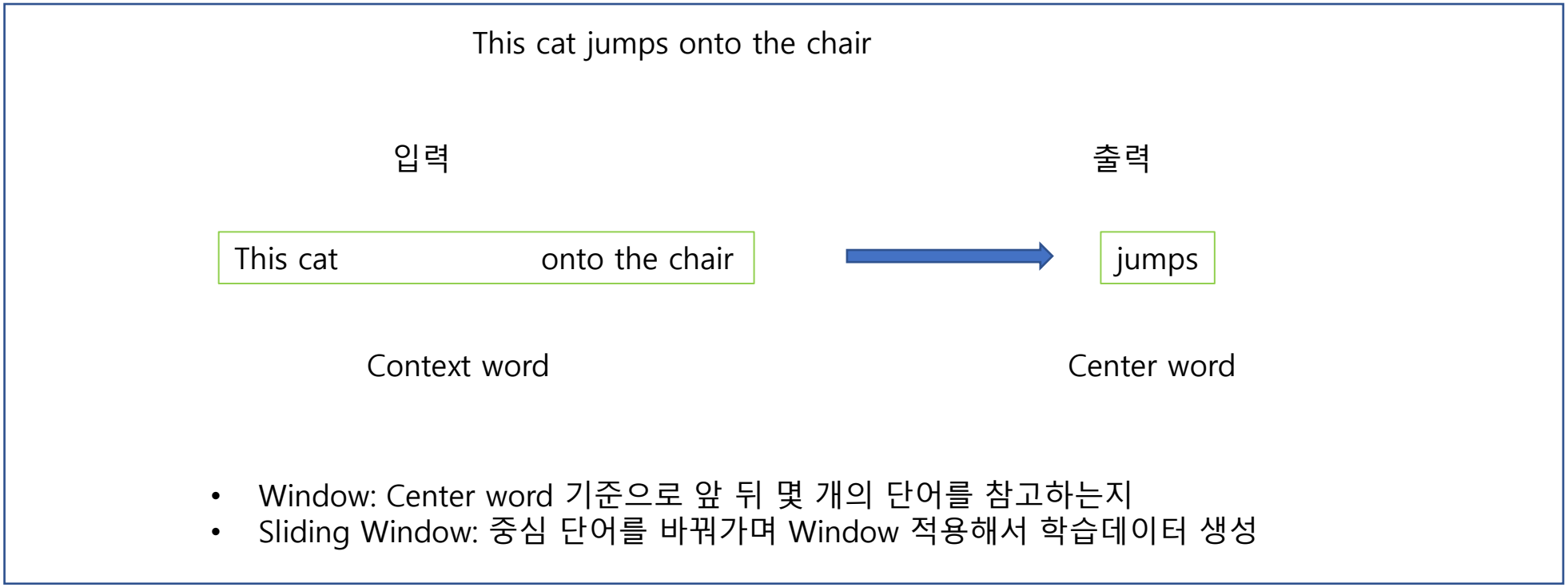
- 워드2벡터는 단어를 벡터로 바꾸는 대표적인 임베딩 모형
- 단어를 벡터로 변환하는 알고리즘으로 인공신경망 언어모형을 따르면서 학습 속도와 성능을 개선
- 단어 유사도 계산/ 단어와 특정 쿼리의 유사도 계산 / 문장 분류

* Word2Vec와 Doc2Vecc

워드2벡터(Word2Vec)의 대표적인 알고리즘: CBOW

Continuous Bag of Words

주변 단어(Context Word)로 중심의 단어(Center Word)를 예측



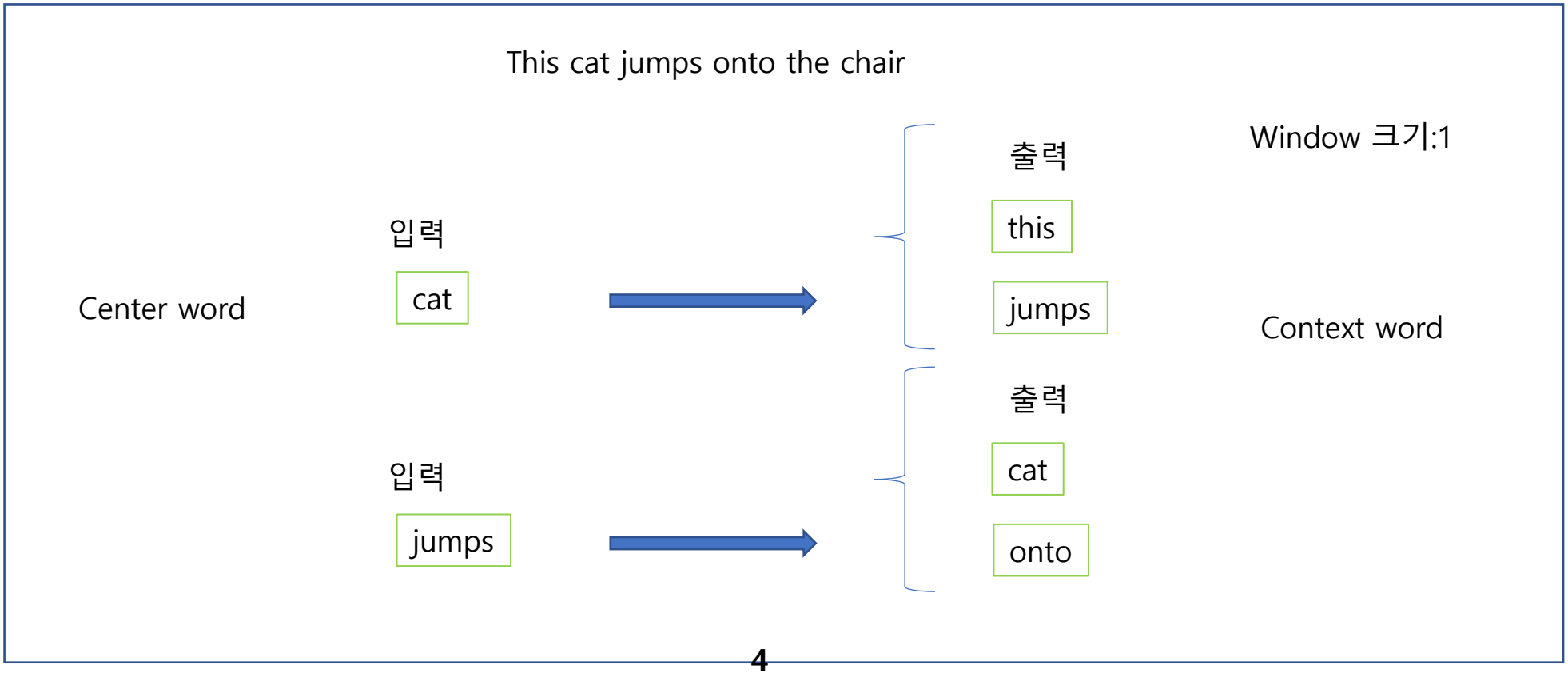
- Window: Center word 기준으로 앞 뒤 몇 개의 단어를 참고하는지
- Sliding Window: 중심 단어를 바꿔가며 Window 적용해서 학습데이터 생성

* Word2Vec와 Doc2Vecc

워드2벡터(Word2Vec)의 대표적인 알고리즘: Skip Gram

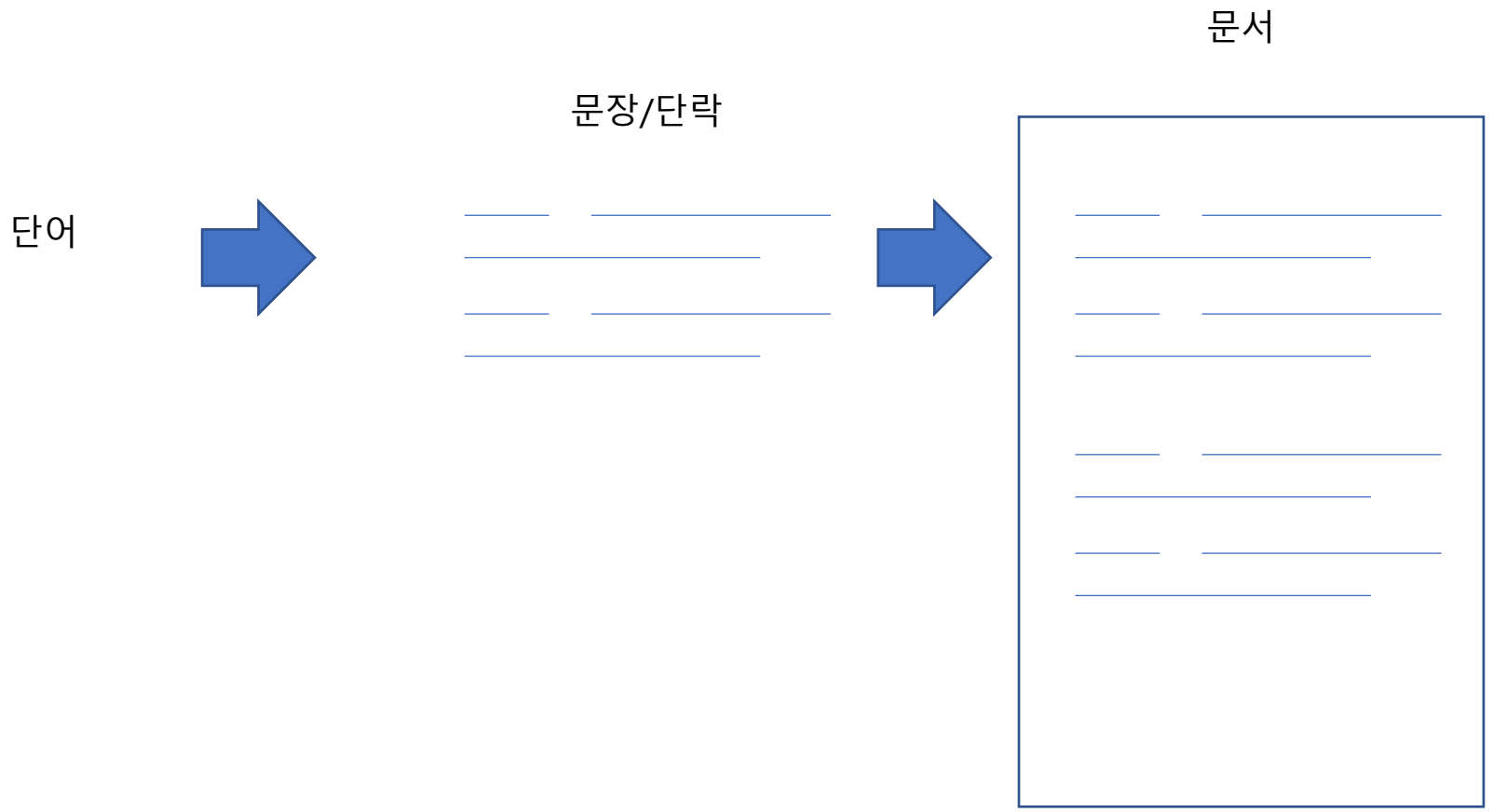
Skip Gram

중심 단어(Center Word)로 주변 단어(Context Word)를 예측



* Word2Vec와 Doc2Vecc

Word2Vec에서 Doc2Vec으로



* Word2Vec와 Doc2Vecc

Doc2Vec이란?

- *Word2Vec에 이어 2014년 구글에서 개발한 모형*
- *다음 단어를 예측하며 로그 확률 평균을 최대화하고 문장/단락/문서의 임베딩 표현을 수행함.*

* Word2Vec와 Doc2Vecc

PV(Paragraph Vector)-DM(Distributed Memory), 성능 우수!

Paragraph ID: para_1

Sentence: The dog sleep on the sofa

Y: Target 단어

X: Target 이전의 k개 단어+Paragraph ID

D2v는 단락에서 단어를 예측하며, 로그확률평균을 최대화하는 학습

예: k=2

X		Y
[para_1, the, dog]		sleep
[para_1, dog, sleep]	on	
[para_1, sleep, on]	the	
[para_1, on, the]		sofa

* Word2Vec와 Doc2Vecc

PV(Paragraph Vector)- DBOW(Distributed Bag of Words)

Paragraph ID: para_1

Sentence: The dog sleep on the sofa

Paragraph ID가 입력, 문장의 단어들이 Target

X	Y
[para_1]	the
[para_1]	dog
[para_1]	sleep
[para_1]	on
[para_1]	the
[para_1]	sofa

QnA