Word2Vec:

Efficient estimation of word representations in vector space

산업융합학부 정보융합전공
2020058995 엄정훈

Introduction

❖ One-Hot Encoding

- 한 단어에 대해, 표현하고자 하는 단어에 1, 나머지 단어는 0을 부여한 vector 표현 방식

Example

귤 →

사과	포도	귤	망고	…	오렌지	참외
0	0	1	0	…	0	0

오렌지 →

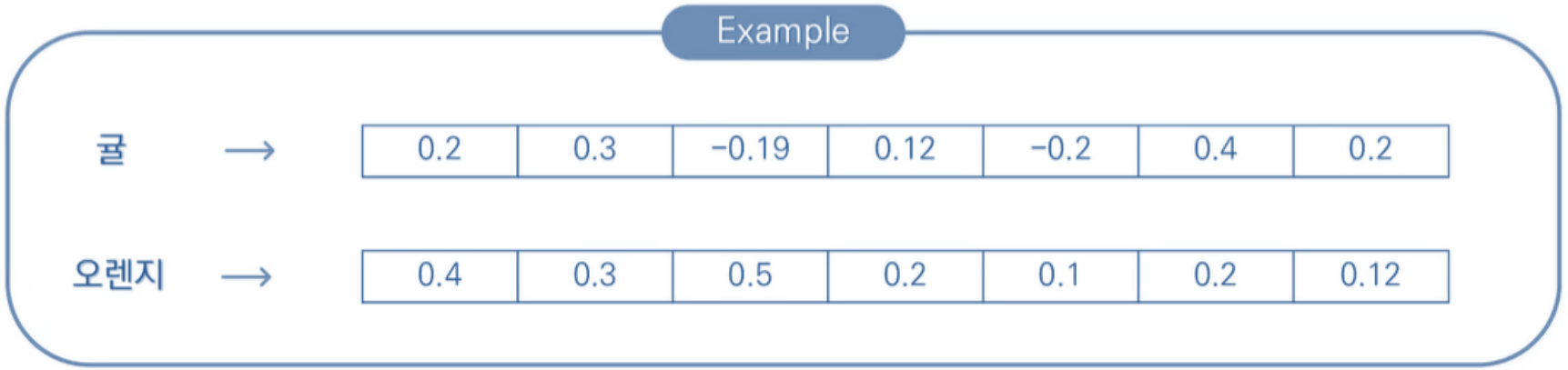
사과	포도	귤	망고	…	오렌지	참외
0	0	0	0	…	1	0

- 한계점 : 단어 간 유사성에 대한 정보를 담고 있지 않다.

Introduction

❖ Distributed Representation (분산 표현)

- 단어의 의미를 여러 차원의 벡터에 분산시켜 표현

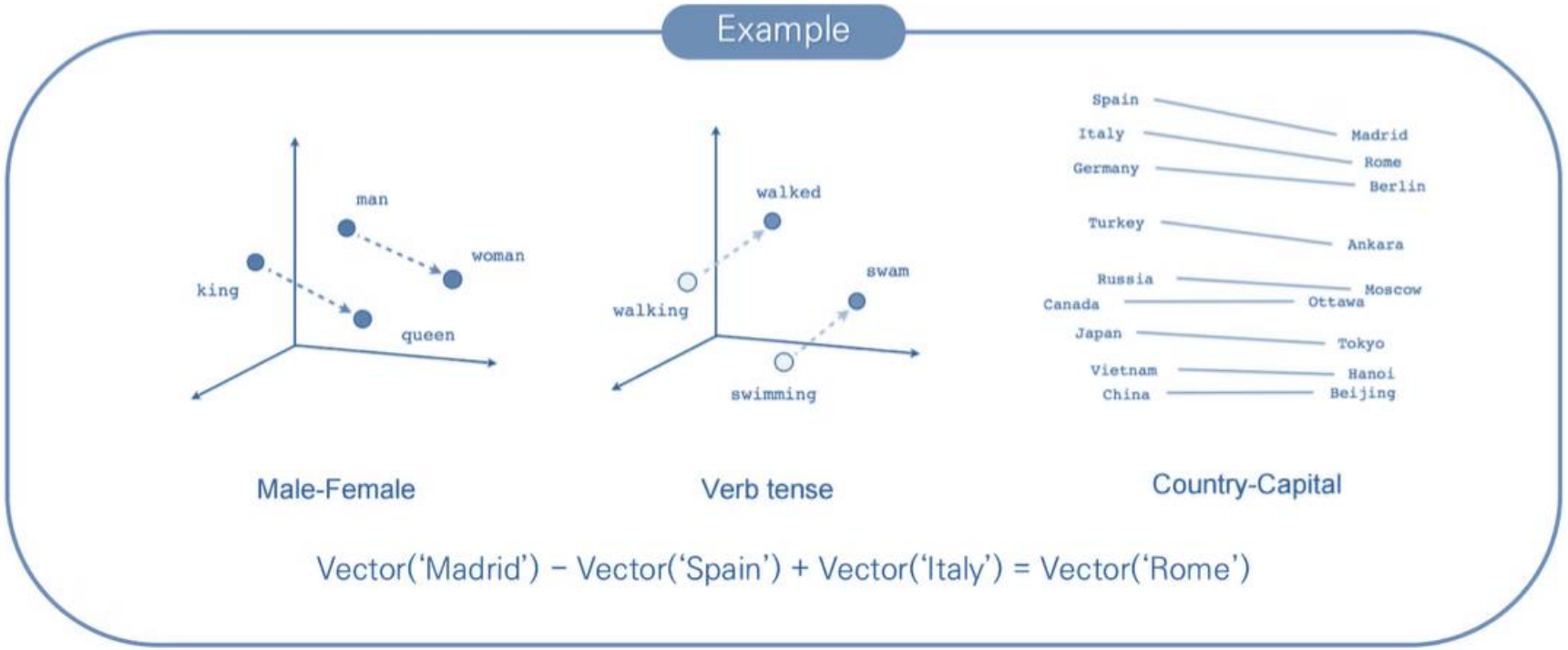


- 분산 가설 : 비슷한 문맥에서 나타나는 단어는 비슷한 의미를 가진다

Goals of the Paper

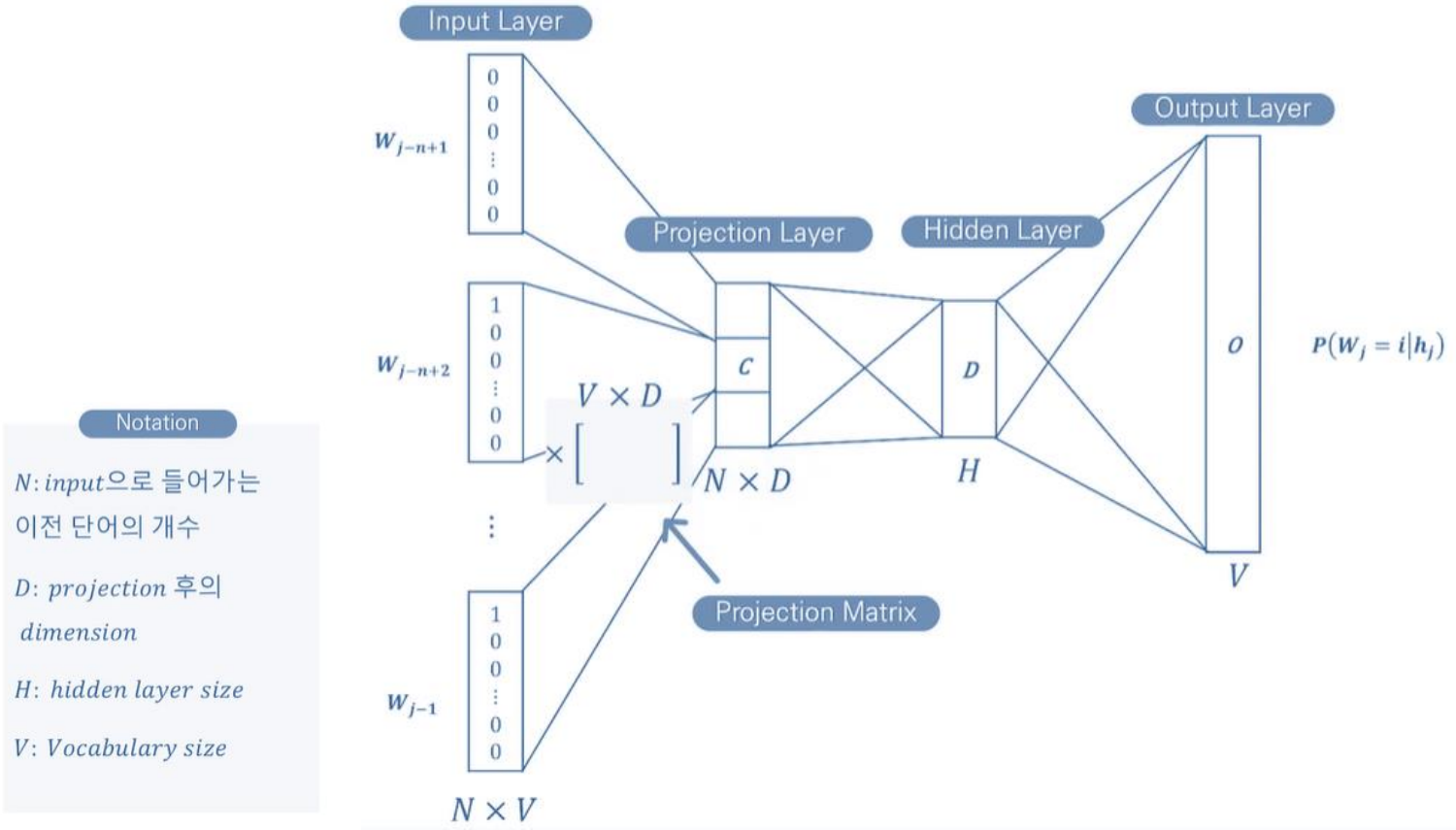
❖ Word2Vec의 이점

- 단어간 유사성 계산 가능



Previous Work

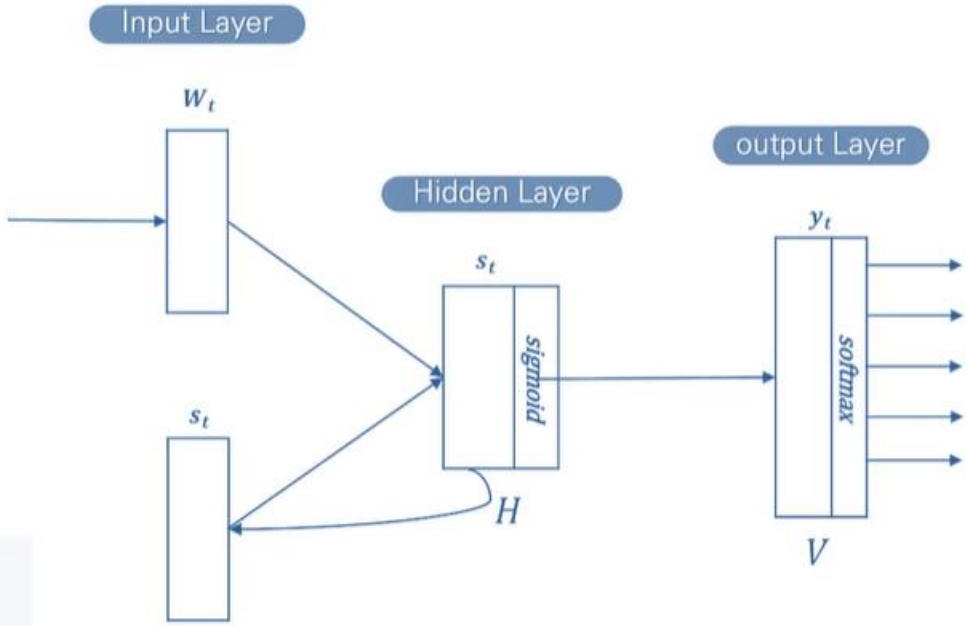
❖ Feedforward Neural Net Language Model (NNLM)



- 한계 : N의 크기가 고정되어 있다.

Previous Work

❖ Recurrent Neural Net Language Model (RLLLM)



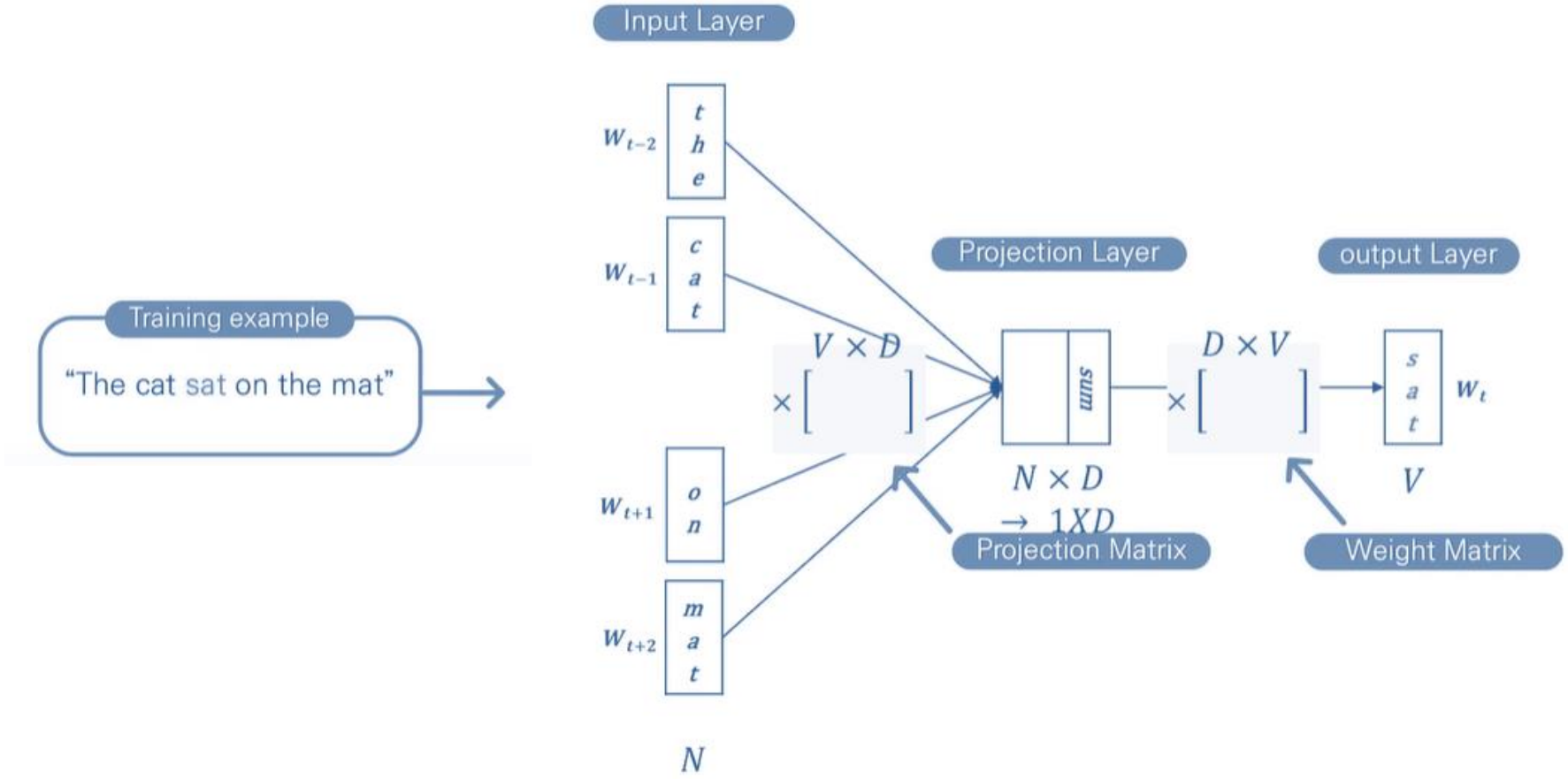
Notation

H: *hidden layer size*

V: *Vocabulary size*

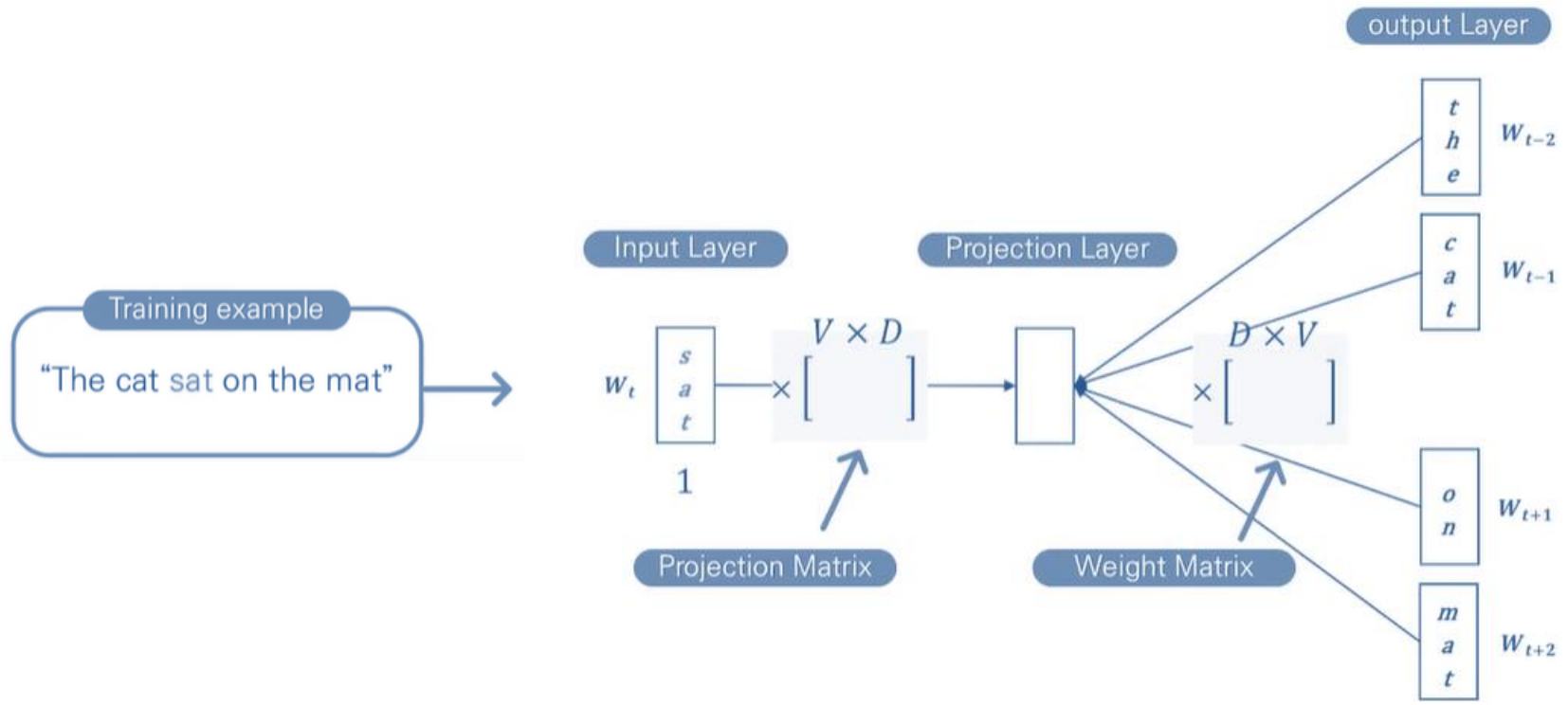
Word2Vec

❖ Continuous Bag of Words Model



Word2Vec

❖ Continuous Skip-gram Model



Results

❖ Task Description

- Semantic question 5가지 (어휘 질문)
- Syntatic question 5가지 (구문 질문)

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Results

❖ Maximization of Accuracy

- Continuous Bag of Words Model 사용
- Data : Google News corpus (약 60억개의 token 포함)

- Word vector dimensionality 변경

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

- Word vector의 차원과 data size를 동시에 크게 가질 때 성능이 향상됨

Results

❖ Comparison of Model Architecture

- Word vector dimensionality=640 고정
- Training data 동일 (3만개의 어휘 테스트셋에 있는 semantic, syntactic 질문 모두 사용)

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

- NNLM이 RNNLM보다 성능이 좋다. 이는 NNLM이 projection layer를 거쳐 hidden layer로 들어가기 때문
- CBOW는 syntactic task 에서 NNLM보다 큰 성능 향상을 보임
- Skip-gram은 syntatic 질문에 대해서는 CBOW보다 약간 성능이 떨어지지만, semantic에서는 가장 좋은 성능 보임

Results

❖ Comparison of Model Architecture

- 동일한 training dataset에 대해, epoch = 1 or 3으로 실험한 결과
- Full Semantic, Syntatic dataset 사용

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

- 같은 data size로 3번의 epoch을 도는 것 보다 두 배 이상의 data로 epoch 한 번을 도는 것이 수행시간이 더 짧고 비슷하거나 더 좋은 결과를 냈다.
- Training data size를 두 배 이상 늘리는 것 보다 vector 차원을 두 배 늘리는 것이 더 큰 성능 향상 가져옴

Results

❖ Word Relationships

- Best word vector를 사용해 도출한 word pair relationship 예시

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

- 두 vecto를 뺌으로서 관계를 정의할 수 있다. 예) Paris – France + Italy = Rome
- Accuracy = 약 60%
- Relationship example 10개 이상 주어지면, accuracy 약 10% 향상

결론

- ❖ 매우 간단한 model 구조 사용, 높은 질의 word vector를 학습할 수 있다는 것 발견
- ❖ 다양한 질문을 제시해 word vector가 다양한 의미들에 여러 유사성 반영 확인
- ❖ 이전 model보다 낮은 계산 복잡도로 많은 dataset으로 부터 높은 차원의 정확한 계산 가능
- ❖ Pre-trained embedding model로써 다양한 NLP task(자연어처리 작업)에 사용 가능