MINDs Lab

maum.ai

NLP

[NLP]

① **Natural Language (자연어) 정의 및 이해**

② **한국어의 이해**

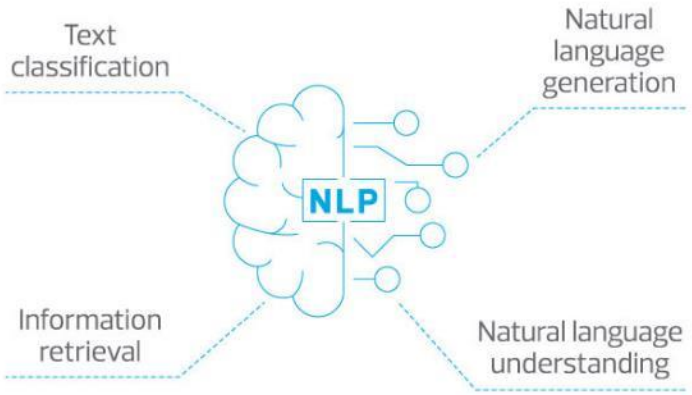
③ **NLU Processing(Natural Language Understanding)**

④ **Vectorization**

Natural Language(자연어) 정의

Natural Language(자연어):

1. 정보전달의 수단
2. 인간 고유의 능력
3. 인공언어에 대응되는 개념
4. 특정 집단에서 사용되는 모국어의 집합



Natural Language Process(자연어 처리):

1. 문서를 문장으로 분할하고,
2. 문장을 최소의 의미 단위인 형태소로 나누어 품사를 부여하고,
3. 문장에 포함된 인물명, 기업명, 장소, 숫자 표현 등을 인식함

Natural Language Understanding(자연어 이해):

1. 자연어 처리에서 한단계 진화된 형태, 단순히 자연어를 처리하는 것이 아닌 실제로 인간의 언어를 이해 하기 위하여 처리하는 것.
2. 형태소 분석, 개체명 인식 뿐만 아니라 구문 분석, 화자 의도 분석, 감성분석까지 결합시켜서 인간의 언어를 의도를 완벽히 이해하기 위함.

언어의 이해

언어의 특징

인간은 두 가지 방식으로 언어를 사용한다.

첫 번째는 **언어로 인간의 경험을 표상**하는 것이고,

두 번째는 인간의 지각이 부분적으로 **모델이나 표상에 의해 결정**된다는 것이다.

언어로 인간의 경험을 표현할 때, 추리 사고, 상상, 리허설 등을 사용한다.

이때, 인간은 자신의 경험 모델을 지각을 기반으로 창조된다.

인간은 자신의 모델이나 표상을 바탕으로 사고하기 때문에 지각이 부분적으로 결정되는 특징을 지닌다.

언어의 이해

Meta Model

Meta Model이란 대상을 직접 서술하는 언어 그 자체를 **고차적으로 언급하는 언어 모델**이다.

Meta Model을 사용하면, **대화 시 누락된 정보를 명확하게 전달**할 수 있다.

Meta Model에는 세 가지 메커니즘인 **일반화, 삭제, 왜곡** 이 있다.

Meta는? 영어의 접두사로, 다른 개념으로 부터의 추상화를 가리키며, 후자를 완성하거나 추가하는 데에 쓰인다.

언어의 이해

Meta Model - 일반화

일반화

전체적인 경험으로부터 분리된 개인 모델의 **부분적인 경험이 하나의 모본이 되어 전체의 범주로 표상**되는 과정이다.

인간은 어릴 때부터 세상에서 생존하기 위해 세상의 경험을 일반화하는 능력이 필수적이었다.

일반화의 예시

“검정 정장을 쫙 빼입은 사람은 보수일거야.”

“바짝 마른 것을 보니 성깔이 있을 것 같군.”

“경상도 남자는 다 그지 뭐.”

“전라도 사람은 좀 그래.”

“말투를 보니 목사님이시군요.”

언어의 이해

Meta Model - 삭제

삭제

자신의 경험 중에서 어떤 특정한 부분에 **선택적으로 주의를 기울이고 다른 것은 배제**하는 과정.

삭제는 사람들이 많은 소란스러운 공간에서 특정인의 이야기를 듣기 위해 다른 소리를 걸러 내거나 차단하는 것과 동일.(칵테일 효과)

사람들은 그와 동일한 과정으로 자신에게 중요한 사람들의 관심 있는 메시지를 스스로 차단 가능.

삭제의 예시

“모두가 그 일은 실현되기 어렵다고 말해요.”

“그 일은 김 주임에게 도와달라고 하면 돼.”

“이 기계를 쓰면 훨씬 능률적이야.”

“그것 참 훌륭한 아이디어야.”

“노력을 하는데 잘 안 되는 것 같아요.”

언어의 이해

Meta Model - 삭제

왜곡

왜곡은 우리 **경험 안에서 감각 데이터를 바꾸는 과정**이다.

인류의 문명과 문화에서 나타나는 모든 것 특히 건축, 예술, 과학은 현재의 사실을 왜곡하고 다르게 표현하는 상상의 능력을 통해 구현된 것이다.

왜곡의 예시

“상사가 계속 보고하라고 해요. 저를 못 믿는 것 같아요.”

“그 사람에게 계속 이 일을 맡기는 것은 무리야.”

“오늘 비가 오니까 업무 진행이 잘 안 될 거야.”

“요즘 시간을 거의 내주지 않아요. 이 업무에 관심이 없네요.”

[NLP]

① Natural Language (자연어) 정의 및 이해

② 한국어의 이해

③ NLU Processing(Natural Language Understanding)

④ Vectorization

한국어의 이해

한국어의 형태적 특징

한국어의 대표적인 **형태적 특징**은

풍부한 어휘 생성이 가능한 **단어 형성법** 과

다양한 허사가 발달된 **교착어** 라는 것이다.

허사란? 단어가 그 본래 목적이 퇴화되어 자립적으로 쓰이지 않고, 다른 단어의 문법적, 의미적 보충 역할을 하는 단어를 말한다

한국어의 이해

한국어의 형태적 특징 – 단어 형성법 발달

풍부한 어휘 생성이 가능

한국어의 형태적 특징은 **단어 형성법이 발달**되었다는 것이다.
어근끼리 결합하는 합성어나 어근에 파생 접사가 붙어 만들어진 다양한 파생어가 있다.
단어의 결합 형태에 따라 **합성어, 파생어, 어미가 붙은 경우** 로 나뉜다.

합성어(어근+어근)	:	나무상자, 밤나무, 우리나라
파생어		
- (접두 파생 접사 + 어근)	:	햇감자, 날생선, 드높다
- (어근 + 접미 파생 접사)	:	구경꾼, 가난하다, 새롭다
연결 어미로 연결된 경우	:	벗어나다, 작은집,
연결 어미가 생략된 경우	:	늦봄, 오르내리다

한국어의 이해

한국어의 형태적 특징 - 교착어

다양한 허사 발달

한국어는 세계의 언어 유형 구분 형태인 첨가어, 굴절어, 포합어, 독립어 중 일본어, 터키어와 함께 **교착어 즉 첨가어에 속한다.** 영어는 굴절어, 중국어는 독립어이며, 포합어에는 에스키모어, 바스크어 등이 있다. 포합어는 어근에 붙는 요소가 다양하고 음운 규칙이 복잡하다.

한국어가 첨가어인 이유는 조사, 어미, 접사와 같은 **다양한 허사가 발달** 되어있기 때문이다.

조사는 체언(주어, 목적어, 보어 자리에 오는 명사, 대명사, 수사) 뒤에 붙어 체언의 문장 성분과 성격을 결정한다.

어미는 용언에 여러가지 문법적, 의미적 형태소로 붙어 단어의 뜻을 구체적이고 풍부하게 한다.

접사는 접두사와 접미사로 어근에 첨가되어 새로운 단어를 생성한다.

교착어에서는 하나의 형태소가 하나의 기능을 가지기 때문에 하나의 어간에 여러 개의 문법 형태소들이 결합한다.

수녀님은 늘 아이들에게 성경을 읽히셨습니다.

수녀/님/은/늘/아이/들/에게/성경/을/읽/히/시/었/습니다.

한국어의 이해

한국어의 통사적 특징

한국어의 대표적인 **통사적 특징**은

주어 + 목적어 + 서술어의 문장 구조를 지니며,

어순 변화가 자유롭고,

조사와 어미가 문장 형성의 기능을 갖는 것이다.

나는 너를 사랑한다. **I love you.**

나는 사랑한다. 너를

너를 사랑한다. 나는

너를 나는 사랑한다.

사랑한다. 나는 너를

사랑한다. 너를 나는

통사란? 언어의 구조를 말하는 것으로, 주로 문장 구조 상의 특성

한국어 분석

형태소 분석 (POS Tagging)

1. POS는 Part-Of-Speech에 약자로, 형태소의 뜻과 문맥을 고려하여, 품사를 태깅.
2. 형태소란? '뜻을 가진 가장 작은 말의 단위
3. 형태소 분석은 문장을 형태소 단위로 쪼개서 표현한 것을 이야기 함.
4. 쪼갠 후 명사 및 동사 등의 품사를 태깅해 주는 역할까지 하게 됨.

ex) 미국 환율 알려줘
- 미국/미국/**NN**/조직_지명 환율/환율/**NN**.용언불가능/양
  알리/알리/**VB**/+어주/어주/**EV**/+어/어라/**EE**.종결/명령형

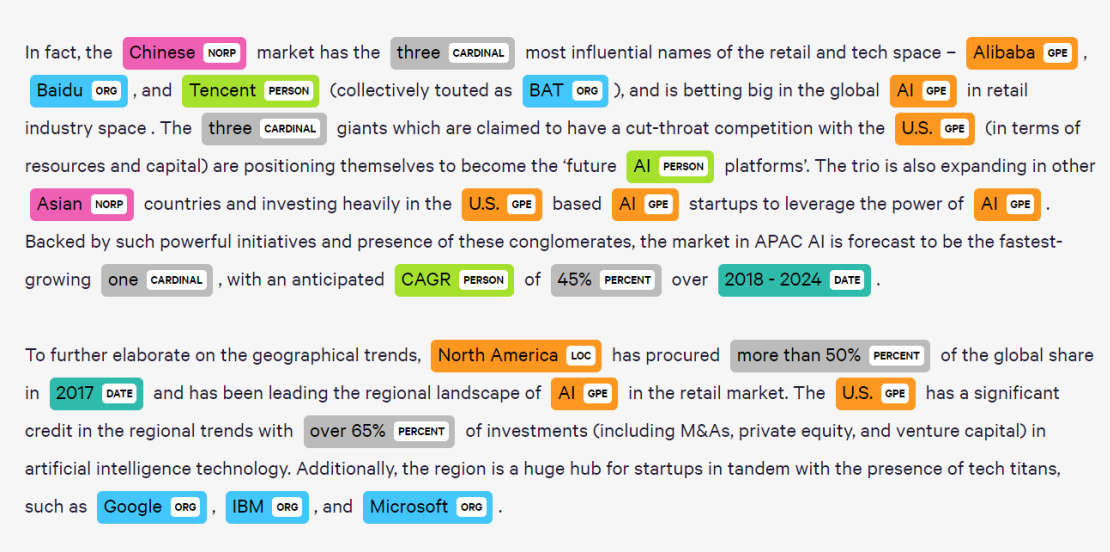
순번	품사번호	품사명		품사명	
1	0	nc	자립명사	NN	명사
2	1	np	대명사	NR	고유명사
3	2	nb	의존명사	NU	수사
4	3	nn	수사	NP	대명사
5	4	pv	동사	SN	접미사
6	6	pa	형용사	PF	접두사
7	8	mag	일반부사	NX	의존명사
8	11	ii	감탄사	AD	부사
9	12	jc	격조사	VV	동사
10	13	co	지정사	AJ	형용사
11	14	ef	종결어미	PP	조사

한국어 분석

개체명 분석(Named-entity Recognition)

1. 단어의 정보에서 추출된 결과 값.
2. 사람, 조직, 지명, 시간 등의 값들을 미리 학습시켜 놓은 사전에 의해서 인식하여 추출하는 방법.
3. 전처리와 후처리 방식 두가지 존재.
4. 전처리 사전과 규칙을 적용하여 개체 모호성을 사용자 주도로 해소.
5. 모델 학습을 통하여 필터링 사전, 후처리 규칙과 후처리 사전을 적용하여 신조어 형태의 개체 유형도 효과적으로 탐지.

ex) 미국 환율 알려줘
- 미국/미국/NN/**조직_지명** 환율/환율/NN.용언불가능/양
  알리/알리/VB/+어주/어주/EV/+어/어라/EE.종결/명령형



출처 : https://medium.com/@b.terryjack/nlp-pretrained-named-entity-recognition-7caa5cd28d7b

[NLP]

① Natural Language (자연어) 정의 및 이해

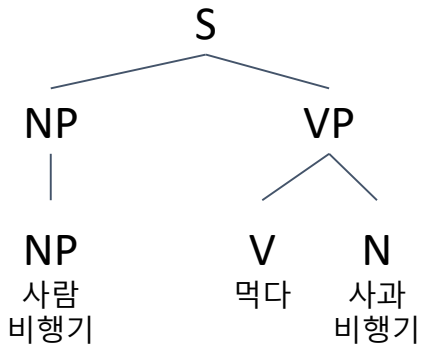
② 한국어의 이해

③ NLU Processing(Natural Language Understanding)

④ Vectorization

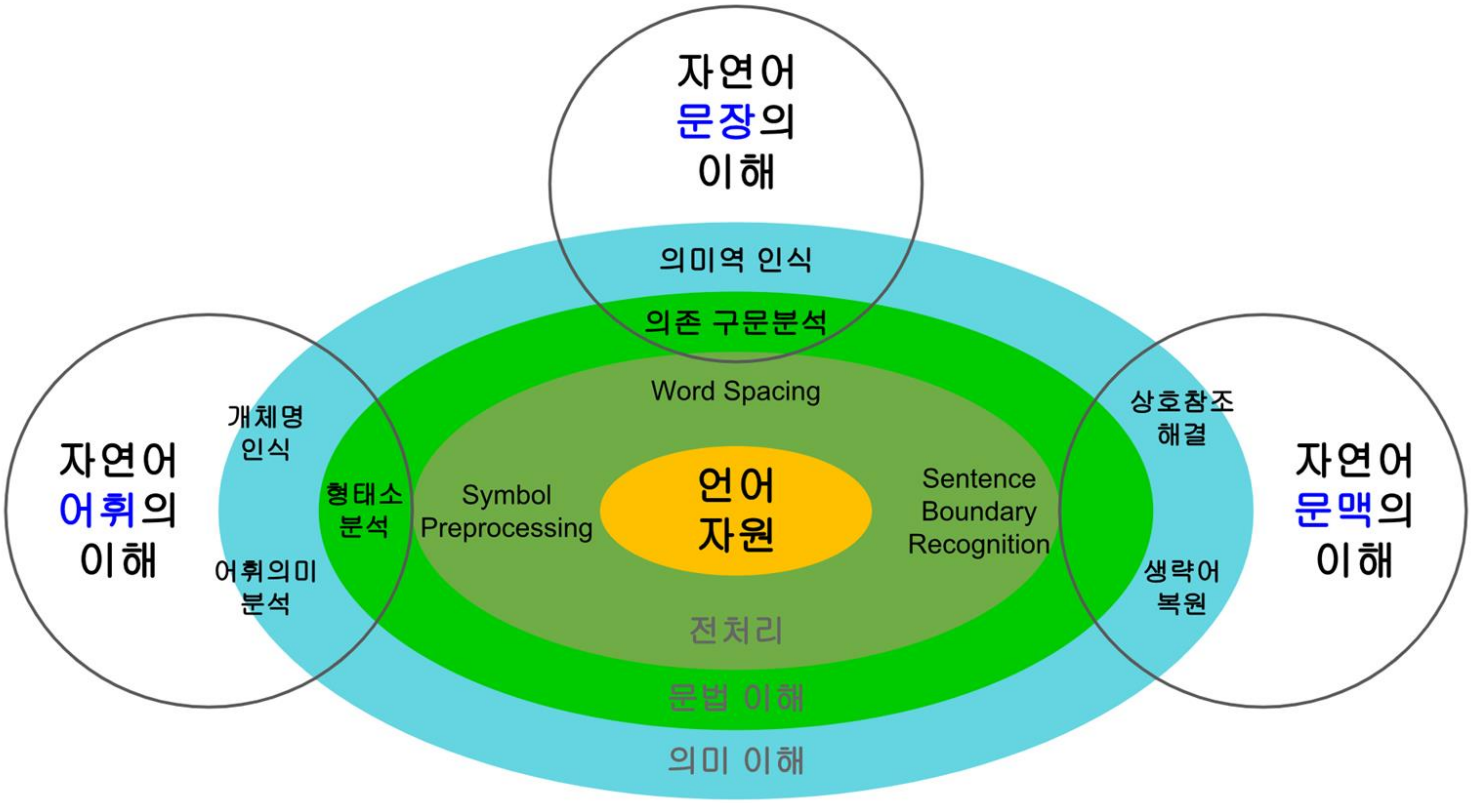
NLU Processing(Natural Language Understanding)

자연어 이해의 순서

형태소 분석	구문 분석	의미 분석	담화 분석
Input을 형태소 단위로 분할하고 품사를 부착	주어, 목적어, 서술어와 같은 구문 단위를 찾음	문장이 의미적으로 올바른 문장인지 판단	대화 흐름상 어떤 의미를 가지는지 찾음
1) 나는 - 나+는 - 날(다)+는 - 나(다)+는 2) 과학자들에게 - 과학자+들+에게		1) 사람이 사과를 먹는다 (O) 2) 사람이 비행기를 먹는다 (X) 3) 비행기가 사과를 먹는다 (X)	1) 철수는 주식으로 돈을 잃었다 그는 울고 말았다 2) 철수는 우승을 했다 그는 울고 말았다

언어 자원 분석

자연어 어휘, 문장, 문맥의 이해



[NLP]

① Natural Language (자연어) 정의 및 이해

② 한국어의 이해

③ NLU Processing(Natural Language Understanding)

④ Vectorization

Vectorization – OneHot Vector

One-Hot Vector

인간이 사용하는 자연어를 처리하기 위해서는 컴퓨터가 이해할 수 있는 형태(숫자나 벡터 등)로 변환 해야 함.

가장 간단한 방식은 단어 단위 **One-Hot 벡터**를 형성.

One-Hot 벡터는 표현하고자 하는 인덱스만 1의 값을 가지고, 나머지 모든 위치에 0이 존재하므로, 각각의 단어를 **상대적 비교 불가**.

Ex)One-Hot Vector

표현해야 할 단어가 3개라고 가정한다면, 아래와 같이 길이가 3인 벡터로 각 단어를 표현.

강아지 : [1, 0, 0]
고양이 : [0, 1, 0]
거북이 : [0, 0, 1]

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

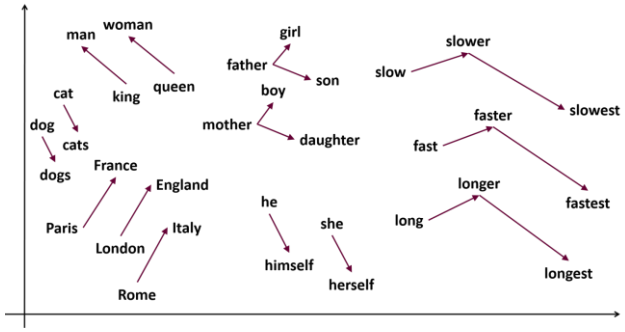
Vectorization – WordEmbedding

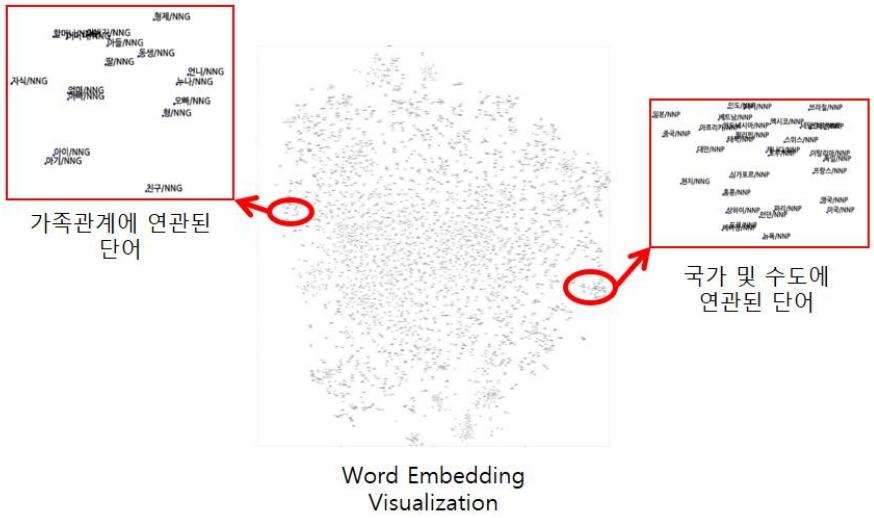
Word-Embedding

대량의 문서 집합을 이용하여 단어 하나 하나를 수십 혹은 수백차원의 벡터로 변환된 값.
서로 다른 단어가 갖는 '유사성' 및 '의미'를 표현하기 위해, 각 차원이 0이 아닌 **실수형 값을 갖는 벡터**로 표현.
각 단어들 사이의 유사도를 측정 할 수 있고, 단어에 대해 수치적으로 쉽게 다룰 수 있음.
유사한 단어들끼리는 **군집화(clustering)**된다는 특징이 있음.
워드 임베딩에는 Word2Vec, GloVe, FastText 등이 있으며, 단어 벡터들 간의 **상대 비교 가능**.

Ex)Word-Embedding
아래와 같이 실수형 값으로 벡터를 표현

강아지 = [0.1, 0.5, -0.3, 0.8, 0.2, 0.5]
고양이 = [0.4, -0.3, 0.2, 0.4, -0.9, 0.2]
거북이 = [0.2, -0.4, -0.7, 0.3, 0.2, -0.3]





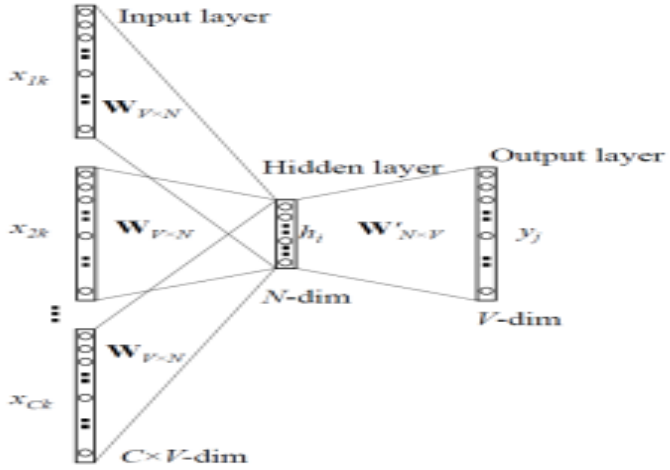
Vectorization – WordEmbedding

Word2Vec

2013년에 구글에서 발표한 Word-Embedding 방식
기존의 계산량을 획기적으로 감소시켜서 몇배 빠른 학습이 가능.
가장 많이 사용하고 있는 Word-Embedding 모델

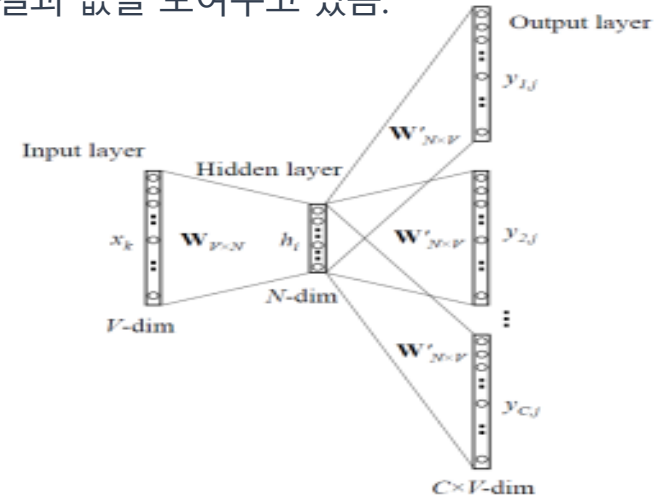
CBOW(Continuous Bag-of-Words) 모델

주어진 단어에 대해 앞/뒤로의 단어를 input으로 사용하여 주어진 단어를 맞추는 네트워크.
모델은 Input Layer, Projection Layer, Output Layer로 이루어짐.



Skip-gram 모델

주어진 단어 하나를 가지고 주변 여러 단어들에 대해서 확률적으로 유추.
CBOW가 더 빠른 연산량을 지니지만, 다소 더 나은 결과 값을 보여주고 있음.



Vectorization – WordEmbedding

GloVe(Global Vectors for Word Representation)

2014년 미국 스탠포드 대학 연구팀이 개발한 단어 임베딩 방법론

LSA(Latent Semantic Analysis)와 Word2Vec의 단점을 보완하고자 개발한 기술

등시등장확률 개념 등장(probability of co-occurrence)

- 학습말뭉치에서 동시 등장한 단어의 빈도를 각각 세어서 전체 말뭉치의 단어 개수로 나눈 결과 값.

특정 문맥 단어가 주어졌을때, 임베딩된 두 단어 벡터의 내적이 두 단어의 동시등장확률간 비율이 되게끔 임베딩.

LSA(Latent Semantic Analysis)

Global matrix factorization 기술을 활용하여 말뭉치 전체의 통계적 정보를 활용.

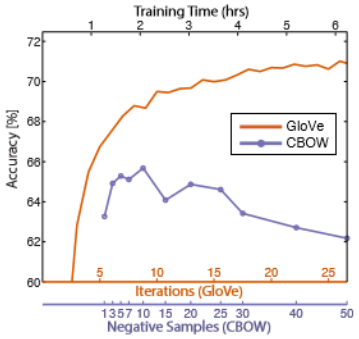
단어 및 문서간 유사도/관계를 측정하기 어려운 단점

+

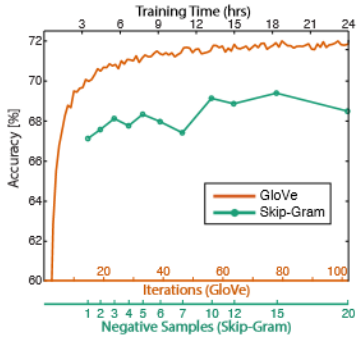
Word2Vec

임베딩된 단어속에서 유사도 측정에 높은 성능 보임.

전체 정보 이용 대신 일부 단어 정보만을 활용하여 추측.



(a) GloVe vs CBOW



(b) GloVe vs Skip-Gram

<출처 : GloVe, Stanford>

Vectorization – WordEmbedding

Word-Embedding 예시

각 단어의 벡터값에 따른 연관성이 높은 단어들을 나열.

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

서울시/nnp 에 따른 워드임베딩 결과

유사 키워드	품사	유사도
대전시	nnp	0.872288745481
대구시	nnp	0.863960947392
광주시	nnp	0.861110910976
서대문구청	nnp	0.854604645319
부산시	nnp	0.849407190551
서울특별시	nnp	0.84725031642
관악구	nnp	0.845485691971
인천시교육청	nnp	0.845327081104
울산시	nnp	0.84488487431
관악구청	nnp	0.839168328143
서귀포시청	nnp	0.834813048457
충남도	nnp	0.831908366916
광명시	nnp	0.828996607731
서초구청	nnp	0.82861986766
노원구	nnp	0.827777197305
아산시	nnp	0.827628197551
천안시	nnp	0.827323302249
충주시	nnp	0.824166605091
금천구	nnp	0.823213377638

출처 : Tomas Mikolov et al.
"Efficient Estimation of Word Representations in Vector Space"