BSC LAB
Business & Service Computing

PNU Industrial Data Science

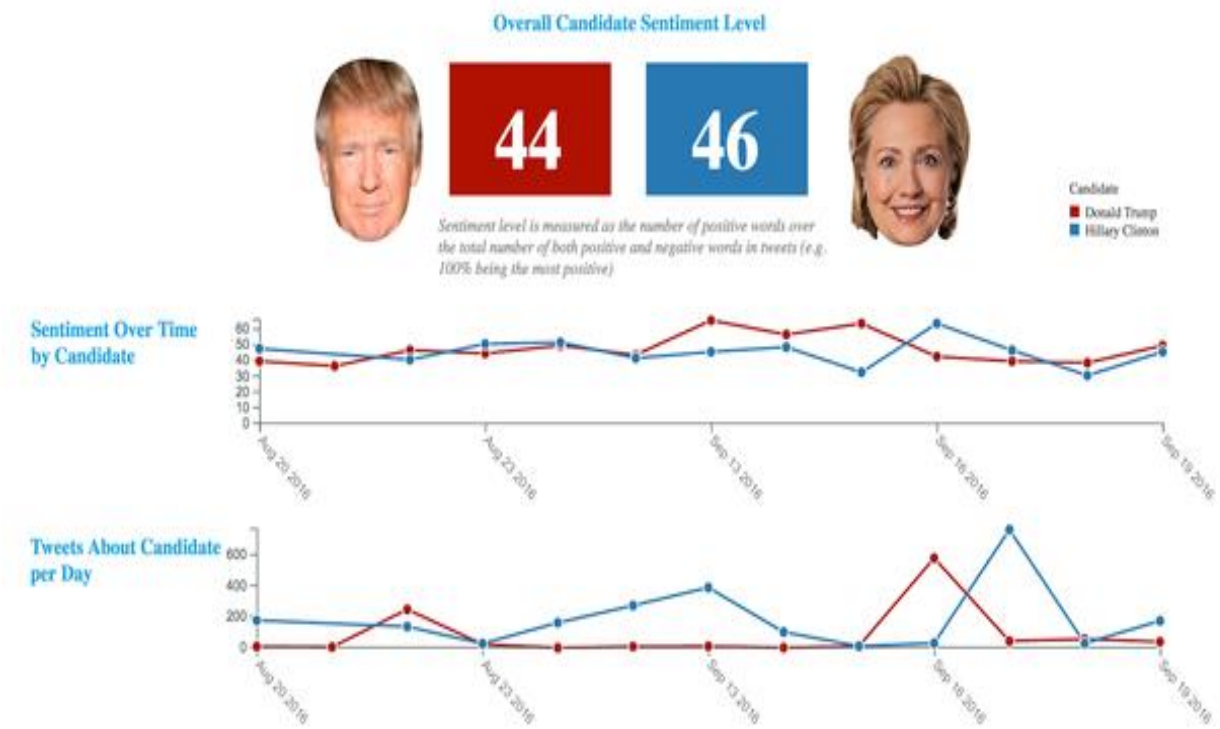
Text mining and NLP

비정형 데이터의 분석

Prof. Hyerim Bae

What is big data?

- Why does everybody want to know about big-data?

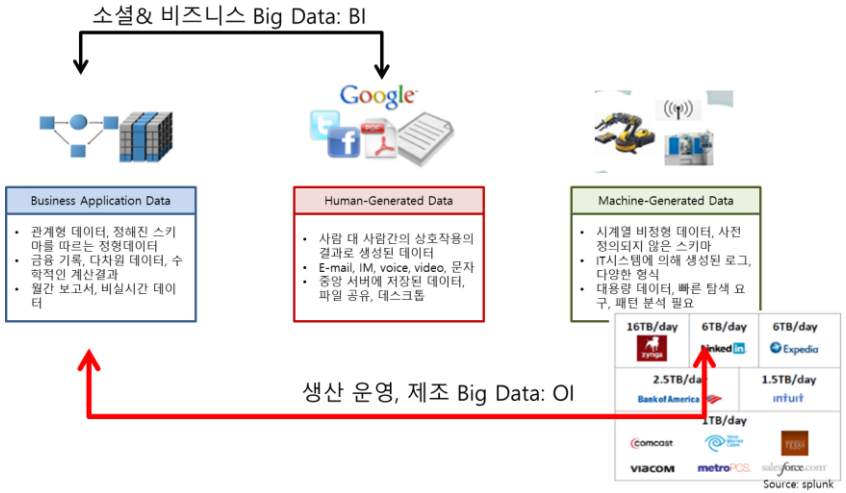


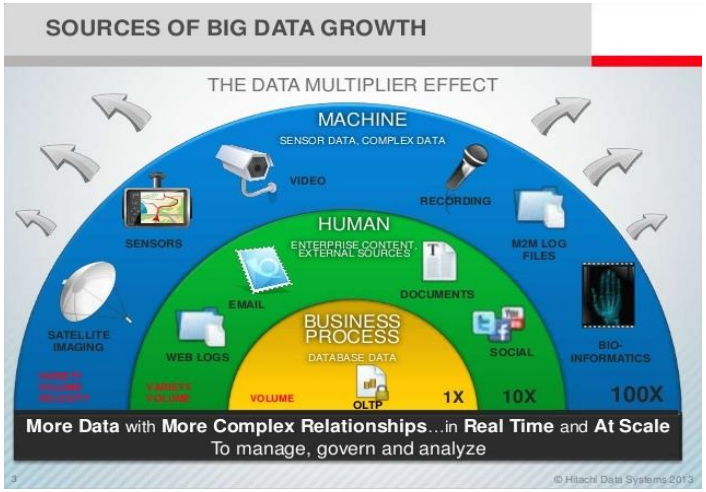
The Trump campaign paid Cambridge Analytica more than $6 million to help it target voters through ads on Facebook.

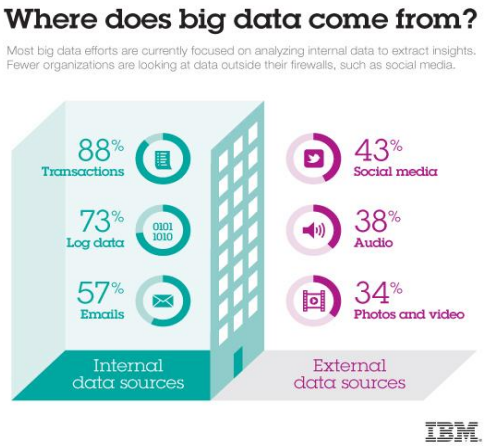
Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

빅데이터의 종류







Source: "Capitalize on Big data through Hitachi Innovation", 2013

데이터의 종류에 따른 분석

데이터 마이닝
머신러닝

프로세스 마이닝

텍스트 마이닝
이미지분석

- 정형화되지 않은 데이터
- 미리 정의된 데이터 모델(구조)을 가지고 있지 않은 데이터

예 아주 많은 양의 데이터를 가지고, 구조와 형태가 다르고 정형화되지 않은 문서, 영상, 음성 등

» 책, 저널, 문서, 메타데이터, 건강 기록, 오디오, 비디오, 아날로그 데이터, 이미지, 파일, 이메일 메시지, 웹페이지, 워드프로세스 문서 등

비정형 데이터

- 텍스트
  - 텍스트 마이닝, 자연어 처리
- 이미지
  - 이미지 분석
- 음성과 영상
  - 영상 처리
- 로그파일
  - 프로세스 마이닝

다음은 누구의 사진일가요?

아리엘공주 in 인어공주

티아나공주 in 공주와 개구리

엘사 공주 in 겨울왕국

벨 공주 in 미녀와 야수

이두나 왕비 in 겨울왕국

모아나 in 모아나

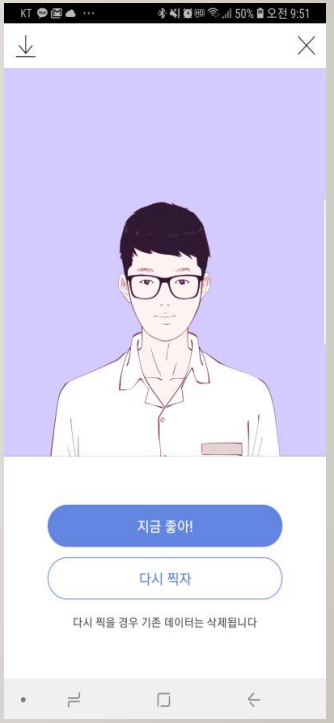
안나 in 겨울왕국

This clip of President Obama talking is fake

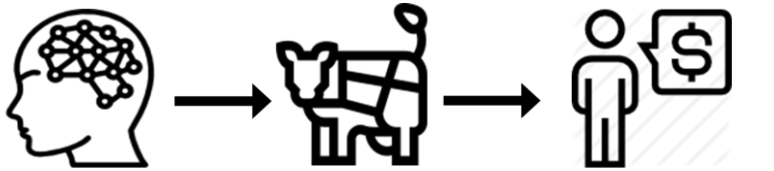
LYREBIRD

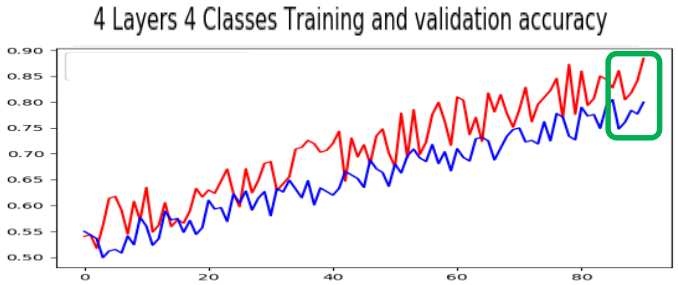
Haeundae Beach is the most famous beach in Busan. The white sand beach is roughly 1.5 kilo-meter long, over a 30~50 meter wide area, creating a beautiful coastline before a shallow bay, making Haeundae Beach perfect for swimming.

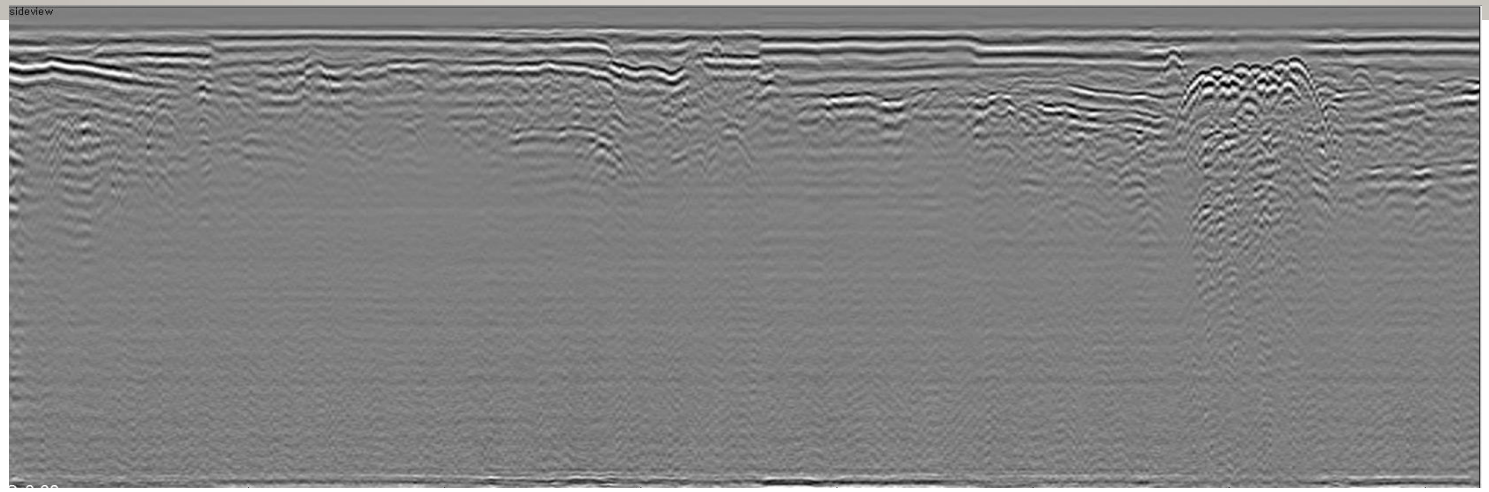
- GNN의 활용 예시
  - N사의 Webtoon에 활용

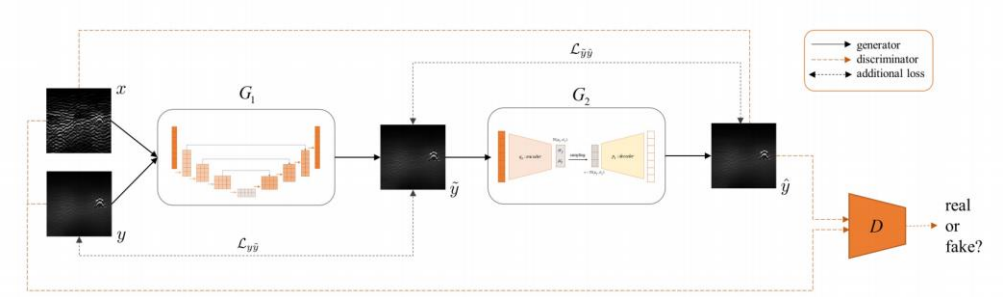


MBC News '한우 등급 속여 판 정육점 적발'









Without noise

Model	MSE
DAE	0.0464
DVAE	0.0442
GAFN	0.0048

Noise = 0.5

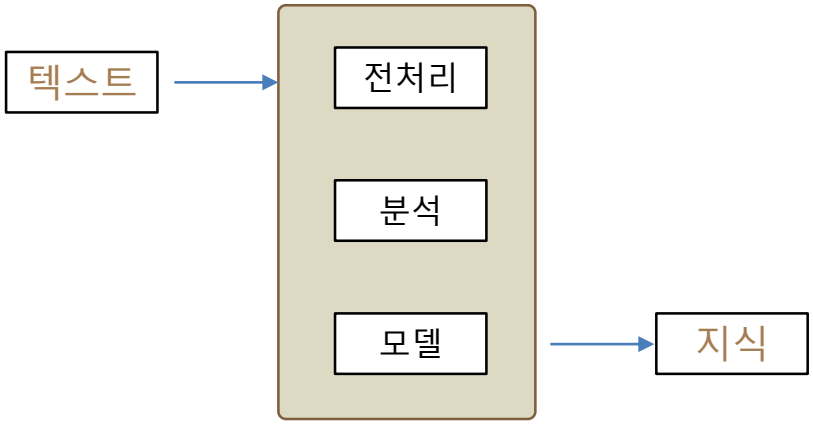
Model	MSE
DAE	0.0473
DVAE	0.0457
GAFN	0.0155

비정형 데이터의 마이닝

- 비정형 데이터의 정형화
  - 텍스트 마이닝
    - 데이터 구조로 변화: Meta data의 활용

텍스트 마이닝

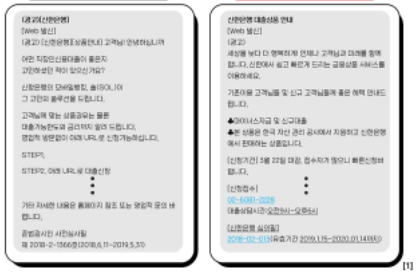
- 대규모 문서에서
  - 정보추출, 연계성 파악, 분류, 군집화, 요약 등을 시행
  - 텍스트에서 지식을 발견



사례

01 연구배경 및 필요성

대표적 전자금융사기 – 스미싱, Smishing



[1]

기존, 스미싱 판별 방식

메시지 발신번호 확인
내용에 첨부된 URL 확인

But,
- 전화번호가 새롭거나 도용되었다면 ?
- 메시지 안에 URL이 없는 경우라면 ?

[1] 지디넷 코리아
[2] WhoWho&Company

01 연구배경 및 필요성

[20.04.24 실제 피해사례]

- 새로운 전자금융사기 – 메신저 피싱
  - 가족, 주변 지인의 프로필 사진과 이름을 도용
  - 전자기기 사용이 미숙한 중·장년층을 타겟 설정
  - 다급한 상황 연출 + 타인 계좌로의 송금 요청
  - 반복적인 보이스톡 연결을 통한 외부로의 연락 방해

“ 해결방법은 단지 스스로 조심하기 ”

02 주제 소개

Point : 접근유형이 다양하게 변하여도 내용은 변하지 않는다.

“ 문제에 대한 해결책을 Text 그 자체에서 찾아보자 ! ”

04 최종 결과물 – 모델 성능 평가, 메신저 피싱

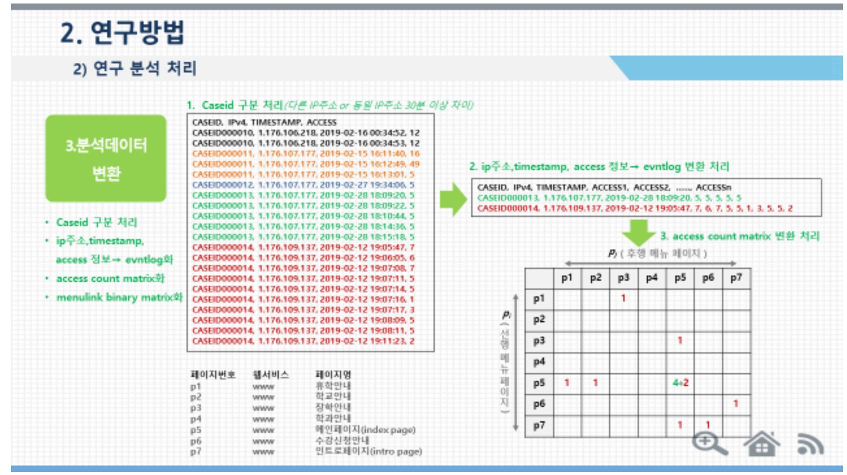
3개의 모델 모두, 약 83%의 특이도과 약 76%의 정확도의 성능을 보여준다.
이것은 어느 수준 이상이 되면 모델의 선택이 아닌 데이터 전처리를 재고민해야 함을 알 수 있다.

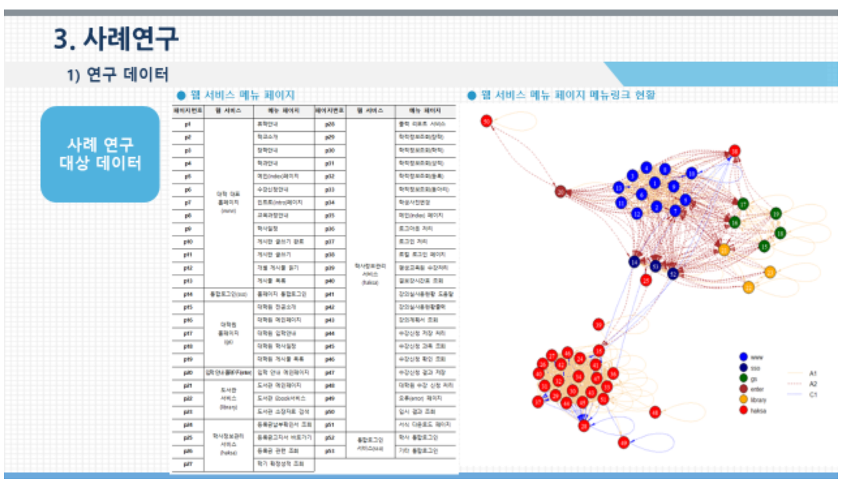
웹마이닝

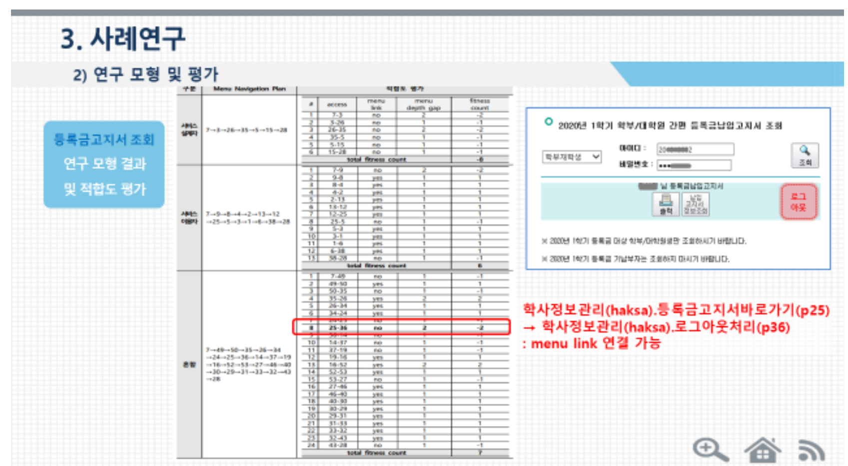
- 웹 콘텐츠 분석
  - 웹페이지내의 컨텐트로 부터 데이터, 정보, 지식을 추출함
- 웹 행위 분석
  - 웨사이트의 페이지간의 연결구조를 분석

사례


2. 연구방법
1) 연구 분석 절차
1. 수집
2. 데이터 전처리
3.분석데이터 변환
4.모델링 (정수계획법)
5. 적합도 평가
분석 기간 설정
웹정보시스템법
web access log 수집
log 중복 제거
log 형식 통일
RDBMS 변환 입력
Caseid 구분 처리
ip주소,timestamp, access 정보→ evntlog화
access count matrix화
menulink binary matrix화
LP 공통 제약조건 수립
설계자/이용자/혼합 LP 비용계수 산출
LP 목적함수 최소화 Solve
Menu Navigation Plan Modeling
Path 적합도 점수 도출
개선 가능 Path 발견 및 개선 제안

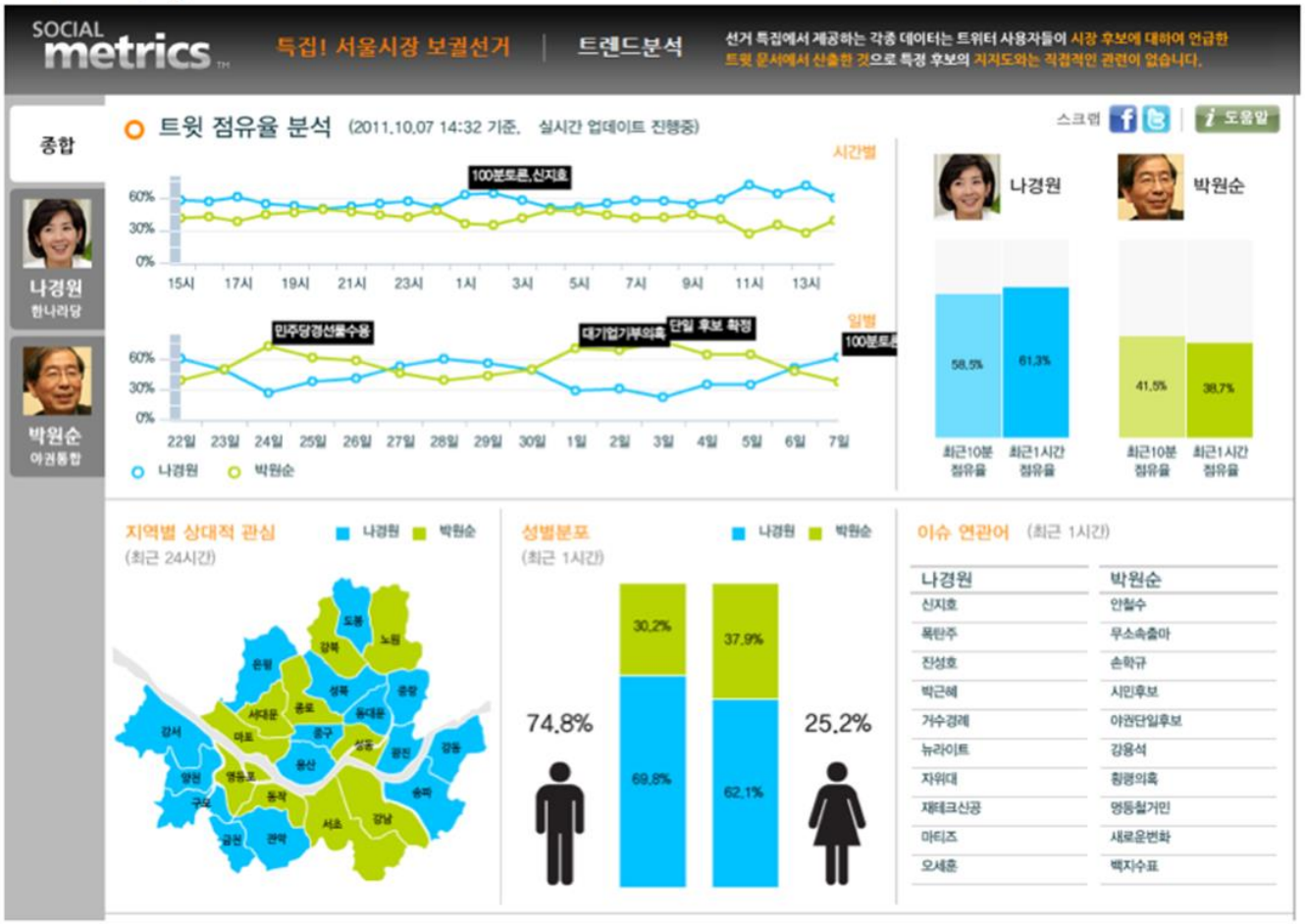

2. 연구방법
2) 연구 분석 처리
3.분석데이터 변환
Caseid 구분 처리
ip주소,timestamp, access 정보→ evntlog화
access count matrix화
menulink binary matrix화
1. Caseid 구분 처리
2. ip주소,timestamp, access 정보→ evntlog 변환 처리
3. access count matrix 변환 처리
CASEID, IPv4, TIMESTAMP, ACCESS
페이지번호 웹서비스 페이지명
p1 p2 p3 p4 p5 p6 p7


3. 사례연구
1) 연구 데이터
사례 연구 대상 데이터
웹 서비스 메뉴 페이지
웹 서비스 메뉴 페이지 메뉴링크 현황

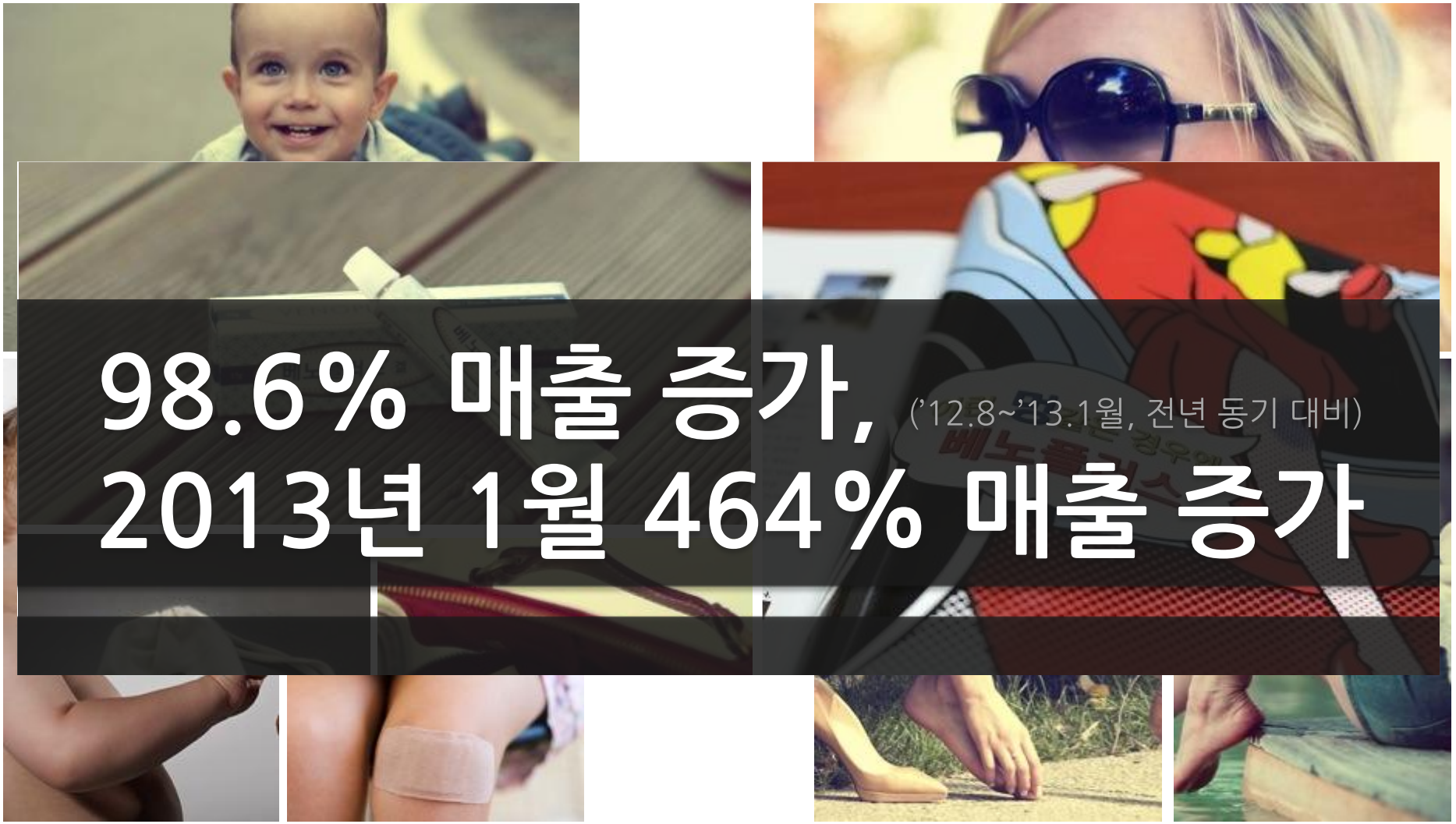

3. 사례연구
2) 연구 모형 및 평가
등록금고지서 조회 연구 모형 결과 및 적합도 평가
Menu Navigation Plan
total fitness count
2020년 1학기 학부/대학원 간편 등록금납입고지서 조회
학사정보관리(haksa).등록금고지서바로가기(p25)
→ 학사정보관리(haksa).로그아웃처리(p36)
: menu link 연결 가능

Opinion Mining

http://campaign.socialmetrics.co.kr/



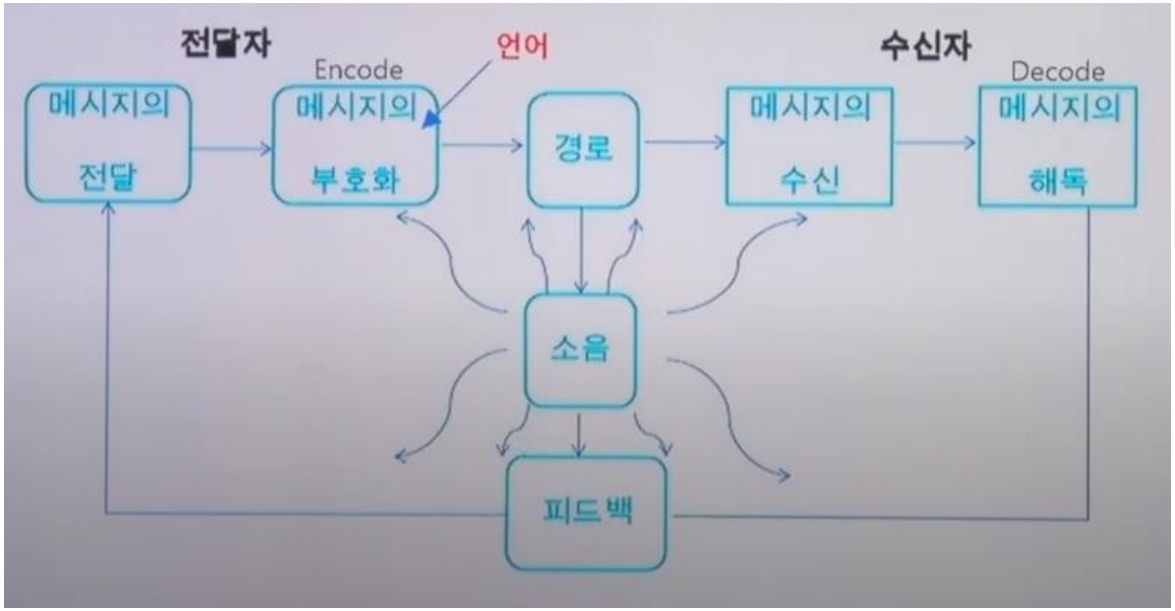
약 장수



98.6% 매출 증가, ('12.8~'13.1월, 전년 동기 대비)
2013년 1월 464% 매출 증가

hrbae@pusan.ac.kr

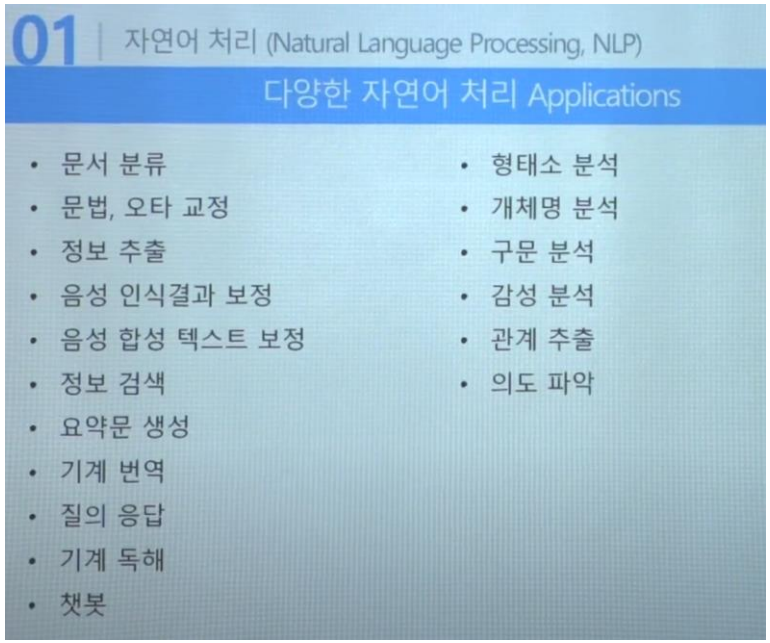
NLP



출처:[토크ON세미나] 자연어 언어모델 'BERT' 1강 - 자연어 처리 (NLP) | T아카데미
https://www.youtube.com/watch?v=qlxrXX5uBoU&list=PLCTJCvcltJv0mSXDydqtOLB2InrtksJEQ&index=26

NLP Problem

- Doc. Classification
- Grammar correction
- Information Extraction
- Voice recognition
- Information search
- Abstracting
- Translation
- Q&A
- Machine interpretation
- Chatbot
- Morpheme analysis
- Sentiment analysis
- Intention analysis



출처:[토크ON세미나] 자연어 언어모델 'BERT' 1강 - 자연어 처리 (NLP) | T아카데미
https://www.youtube.com/watch?v=qlxrXX5uBoU&list=PLCTJCvcltJv0mSXDydqtOLB2InrtksJEQ&index=26

Classification and NLP

- Many NLP problems belong to classification

Question?
Request?
Rejection?
Approval?

Predicate?
Adjective?
Noun?

Positive?
Neutral?
Negative?

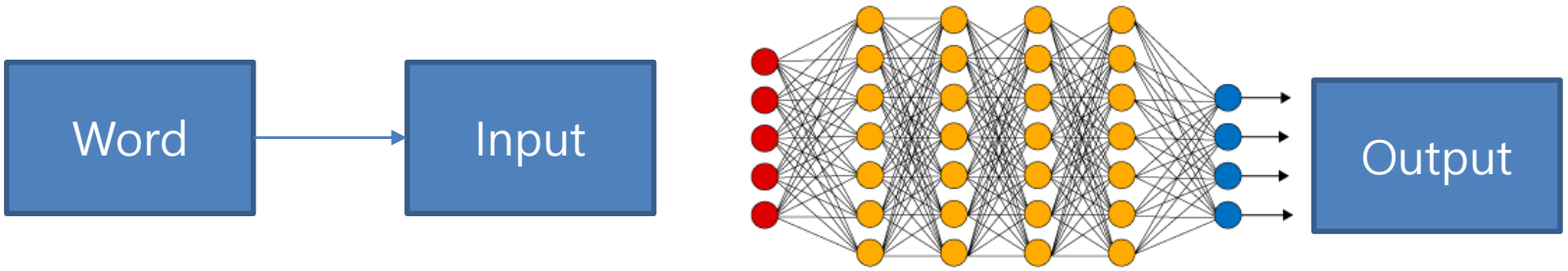
Noun?
Adverb?
Proper noun?
Verb?

Organization?
Person?
Position?

출처:[토크ON세미나] 자연어 언어모델 'BERT' 1강 - 자연어 처리 (NLP) | T아카데미
https://www.youtube.com/watch?v=qlxrXX5uBoU&list=PLCTJCvcltJv0mSXDydqtOLB2InrtksJEQ&index=26

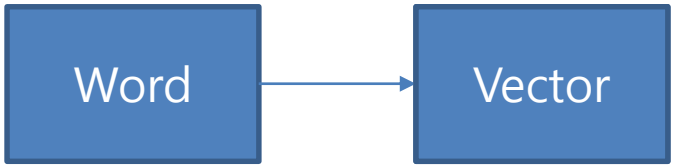
In order to analyze text…

- Can words be input to deep learning?
- We need to convert words to input of deep learning



What is word2Vec?

- Converting words to input of deep learning
- Encoding
  - Converting text to number
  - Number is represented as vector



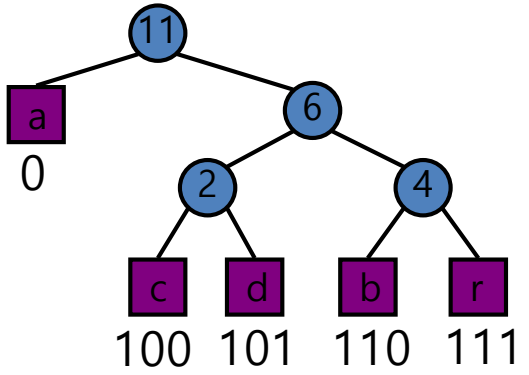
Encoding

Encoding

- How to convert text into number
  - "Hyerim loves Wonkyung"
  - Hyerim: 0, loves: 1, Wonkyung: 2
- Frequency based encoding: Huffman Encoding

X = abracadabra

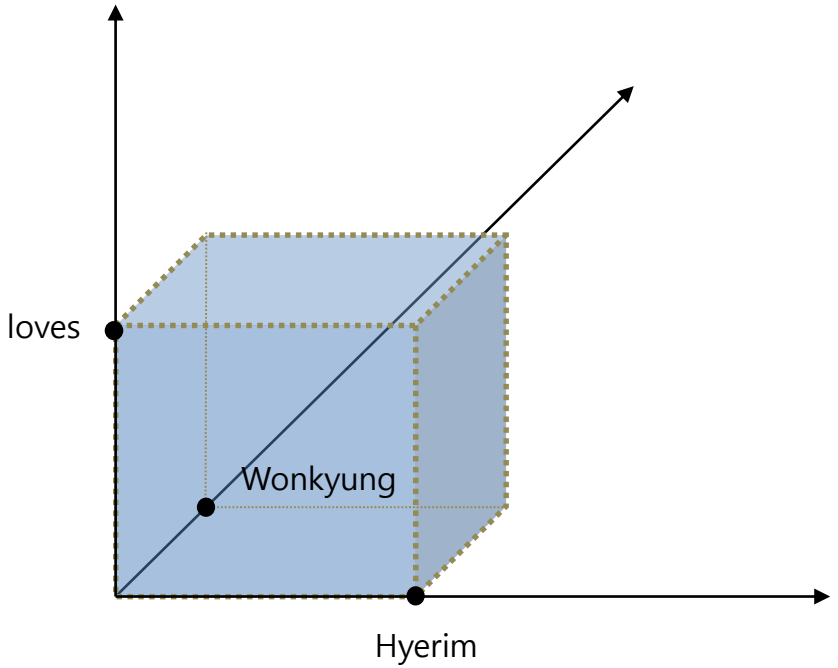
01101110100010 1...



- Label encoding
  - Hyerim: 1, loves: 2, Wonkyung: 3…
- One-hot-ecoding
  - Hyerim: [1, 0, 0], loves: [0, 1, 0], Wonkyung: [0, 0, 1]

Problems of OHE

- No similarity
  - Cannot represent similarity b/w words
  - Cosine similarity is always 0

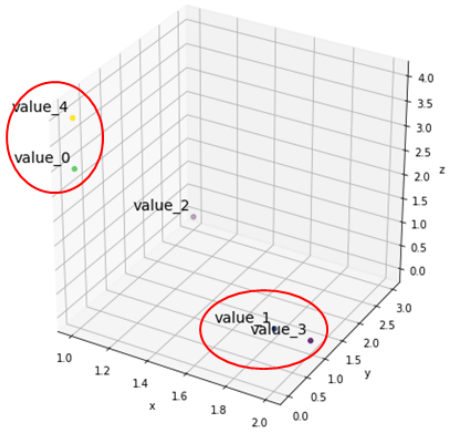


Embedding

- Embedding is dense vector (One category is not one dimension)
  - With similarity

Value of Categorical Feature	Embedding
value_0	[1.0, 0.0, 3.0]
value_1	[2.0, 0.0, 1.0]
value_2	[1.0, 3.0, 0.0]
value_3	[2.0, 1.0, 0.0]
value_4	[1.0, 0.0, 4.0]

$$p(w_c|w) = \frac{\exp(\mathbf{w} \cdot \mathbf{w_c})}{\sum_i \exp(\mathbf{w} \cdot \mathbf{w_i})}$$



Word2Vec is word embedding

- Similarity comes from neighboring relations
  - "king brave man", "queen beautiful woman"

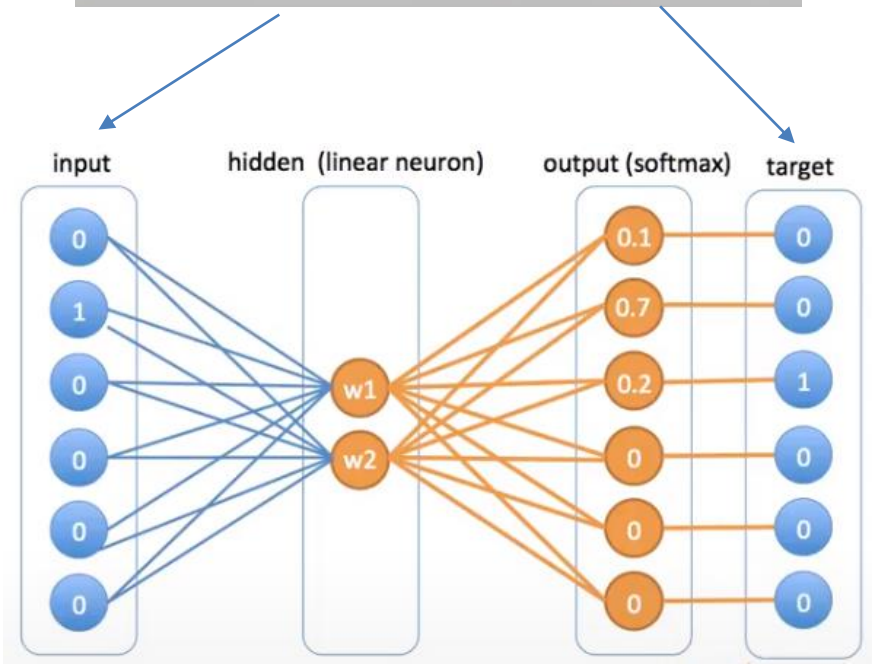
word	neighbor
king	brave
brave	king
brave	man
man	brave
queen	beautiful
beautiful	queen
beautiful	woman
woman	beautiful

Windows size =1

Word2Vec using NN

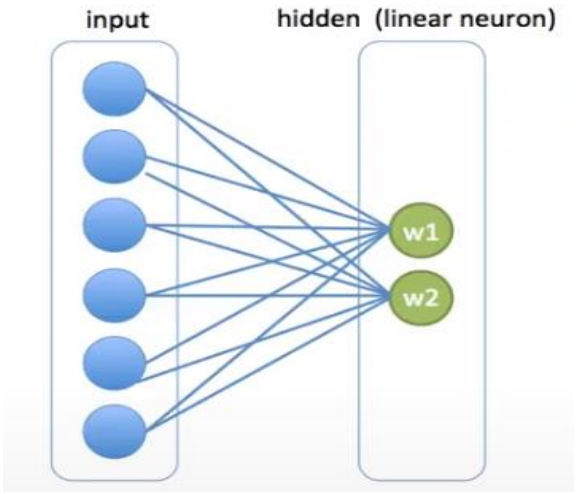
- Training data (Time window = 2)

word	word one hot encoding	neighbor	neighbor one hot encoding
king	[1, 0, 0, 0, 0, 0]	brave	[0, 1, 0, 0, 0, 0]
king	[1, 0, 0, 0, 0, 0]	man	[0, 0, 1, 0, 0, 0]
brave	[0, 1, 0, 0, 0, 0]	king	[1, 0, 0, 0, 0, 0]
brave	[0, 1, 0, 0, 0, 0]	man	[0, 0, 1, 0, 0, 0]
man	[0, 0, 1, 0, 0, 0]	king	[1, 0, 0, 0, 0, 0]
man	[0, 0, 1, 0, 0, 0]	brave	[0, 1, 0, 0, 0, 0]
queen	[0, 0, 0, 1, 0, 0]	beautiful	[0, 0, 0, 0, 1, 0]
queen	[0, 0, 0, 1, 0, 0]	woman	[0, 0, 0, 0, 0, 1]
beautiful	[0, 0, 0, 0, 1, 0]	queen	[0, 0, 0, 1, 0, 0]
beautiful	[0, 0, 0, 0, 1, 0]	woman	[0, 0, 0, 0, 0, 1]
woman	[0, 0, 0, 0, 0, 1]	queen	[0, 0, 0, 1, 0, 0]
woman	[0, 0, 0, 0, 0, 1]	beautiful	[0, 0, 0, 0, 1, 0]



Word2Vec

- Word2Vec is hidden layer



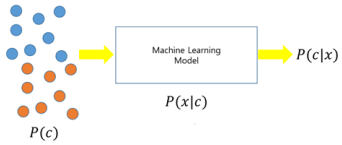
unique word	embedding
king	[1, 1]
brave	[1, 2]
man	[1, 3]
queen	[5, 5]
beautiful	[5, 6]
woman	[5, 7]

http://github.com/minsuk-heo/python_tutorial/blob/master/data_science/nlp/word2vec_tensorflow.ipynb

Bayesian NLP

- Recall ML classification and Bayesian
- From text we want to classify
  - Is the mail spam or not?
  - Use Naïve Bayes
- Problem of NB
  - Cannot consider the order in the text

- Given data, decide a class
- Machine Learning은 data set이 주어졌을 때, 특정 사건 or 가설의 확률을 높여줄 수 있는 최적의 모델을 찾는 것을 목적으로 함
  - 베이지안 모델이 주어진 정보를 update 해 가며 최적의 사후확률을 계산하는 방식과 일맥상통



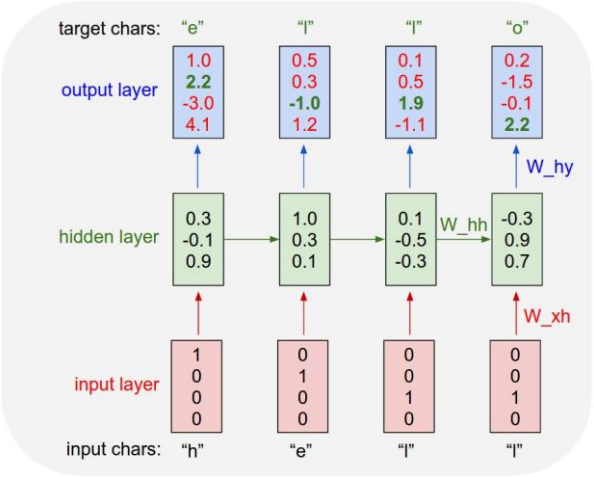
$$P_{nb}(C_1|, \cdots, x_p)$$

$$= \frac{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)]}{P(C_1)[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)] + \cdots + P(C_m)\,[P(x_1|C_m)P(x_2|C_m)\cdots P(x_p|C_m)]}$$

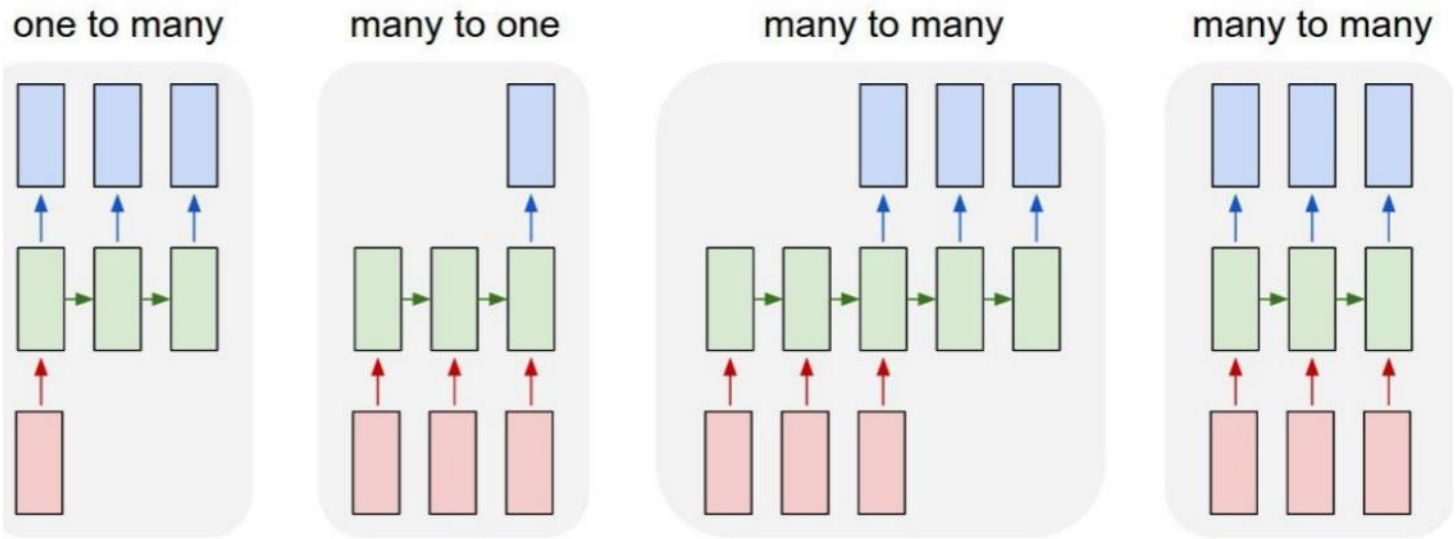
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

NLP and time series

- Previous character influences next character

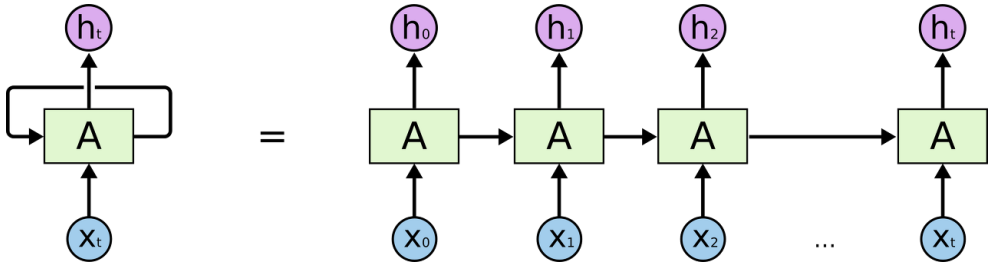


- RNN can solve various types of time series problems



RNN and LSTM

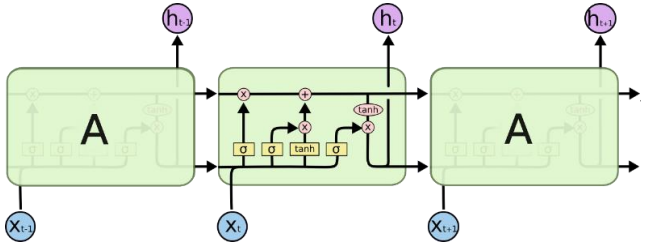
- RNN



- LSTM

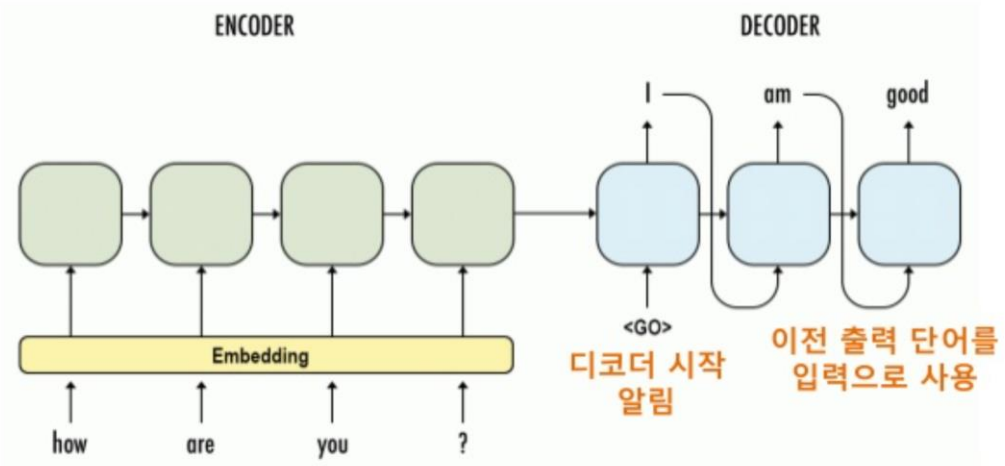
"the clouds are in the *sky*"

"I grew up in France... I speak fluent *French*"



Sequence to Sequence

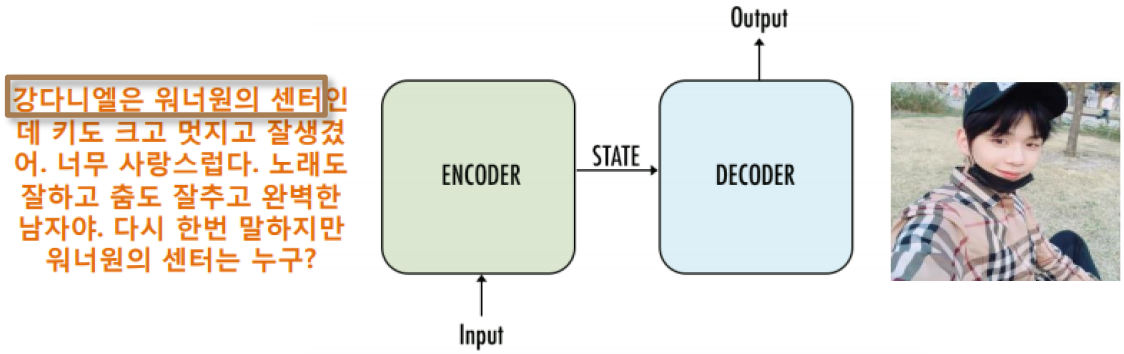
- RNN base approach
  - Cannot process whole text input

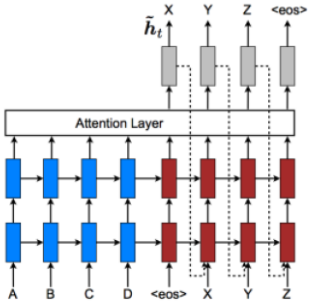


출처: 우종하, "딥러닝 자연어처리:RNN에서 BERT까지"
https://www.slideshare.net/deepseaswjh/rnn-bert?from_action=save

Attention

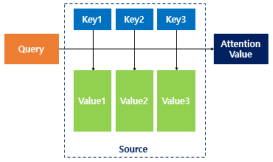
- Seq2Seq
  - In Encoder, there is loss of information





Transformer

- By google, 2017
- LSTM is not needed
- Self attention

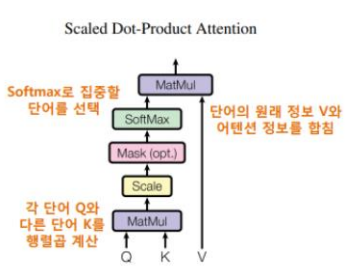


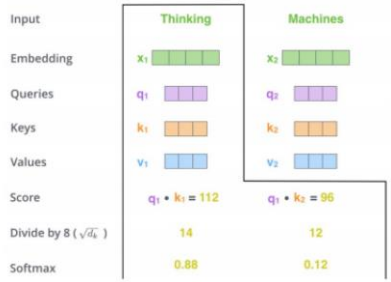
Attention Is All You Need

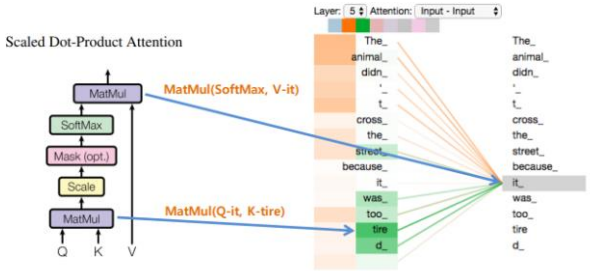
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

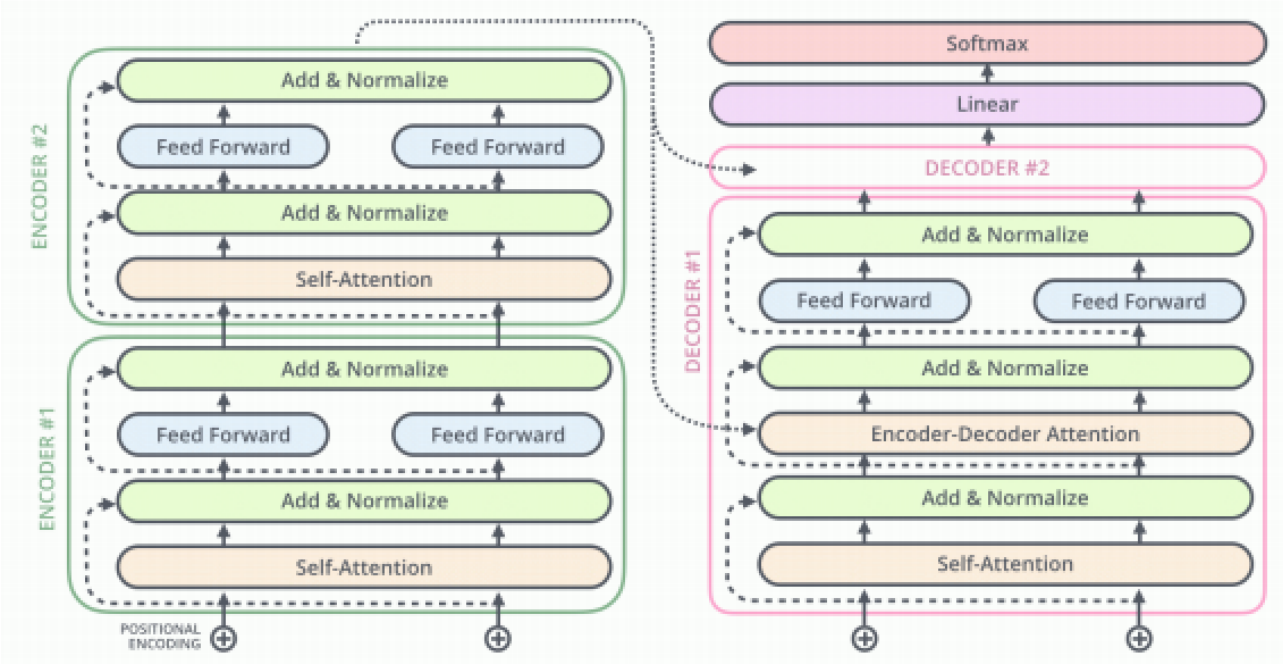






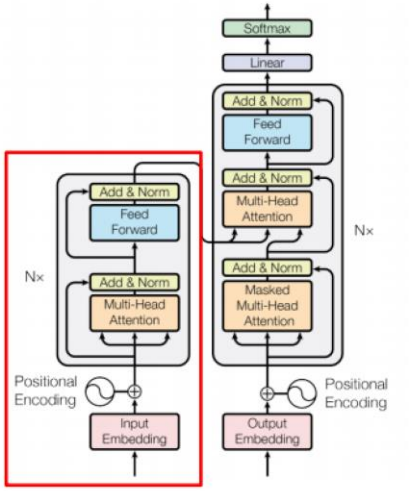
Structure of ‘Transformer’

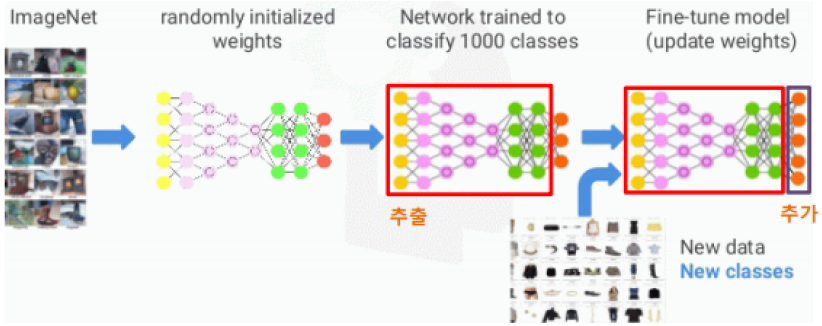
- Nested structure of encoder and decoder



BERT

- By google, 2018
- Pretrained model for transfer learning
- Using transformer
  - BERT Base: 12 transformers
  - BERT Large: 24 transformers





SQuAD competition ranking

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	**83.536**	**86.096**
2 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
3 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
4 Dec 16, 2018	PAML+BERT (single model) *PINGAN GammaLab*	82.577	85.603
5 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.374	85.310