

Research Topic

Jehyuk Lee

Department of AI, Big Data & Management

Kookmin University

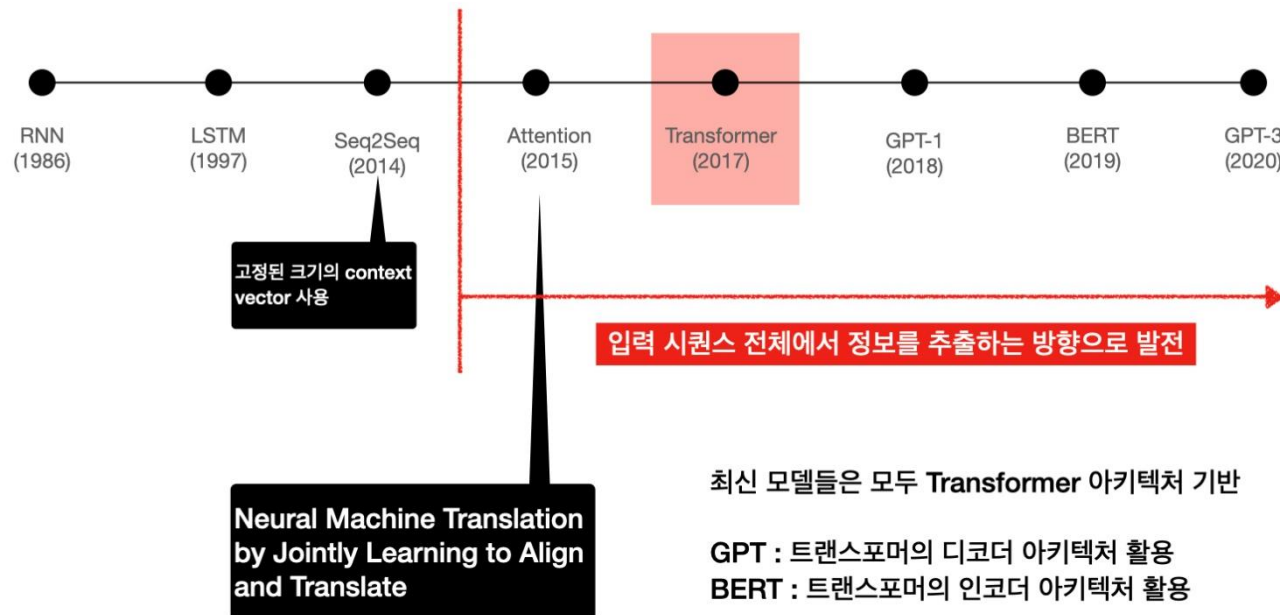
Contents

- 1. From RNN to GPT, BERT
- 2. Text analytics research review
- 3. How to find related papers?

1. From RNN to GPT, BERT

기계번역 발전 타임라인에서

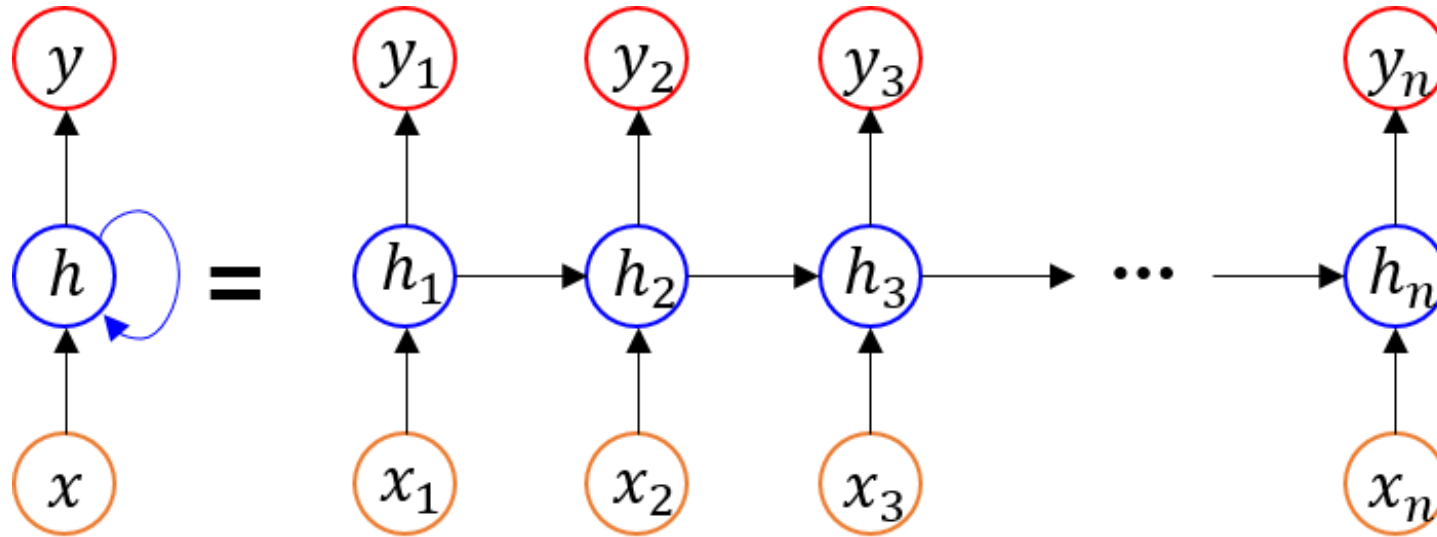
Attention is All You Need 위치는?



RNN family

- Recurrent Neural Network (RNN)

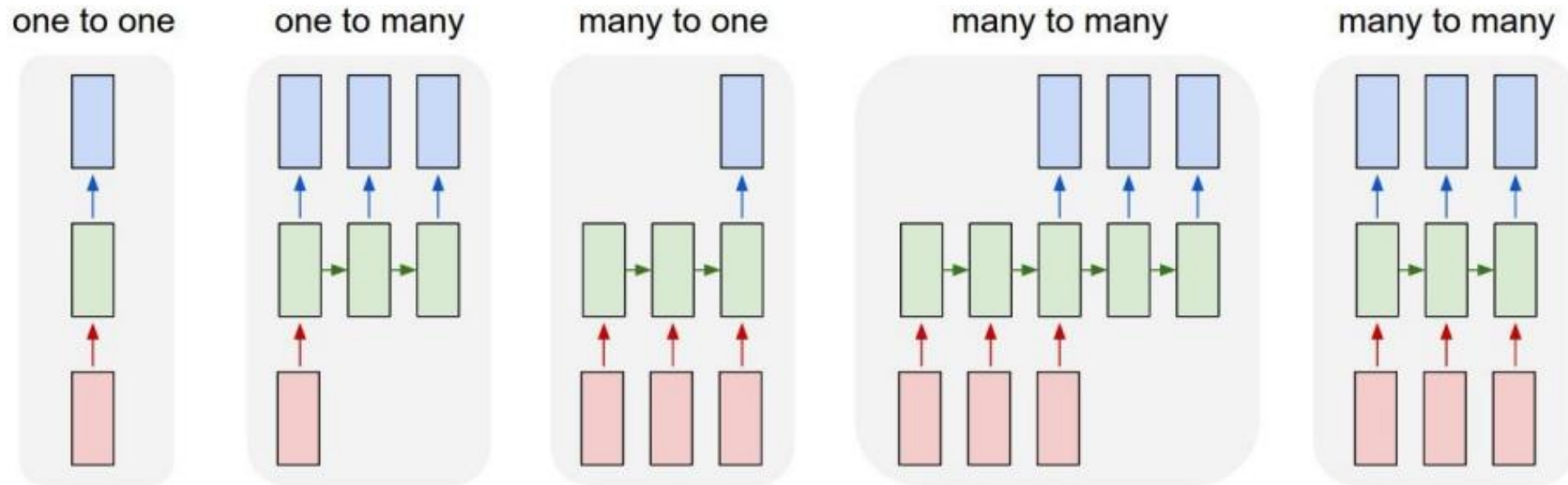
- 순서 정보가 중요한 data에 사용할 수 있는 neural network



RNN family

- Recurrent Neural Network (RNN)

- 순서 정보가 중요한 data에 사용할 수 있는 neural network

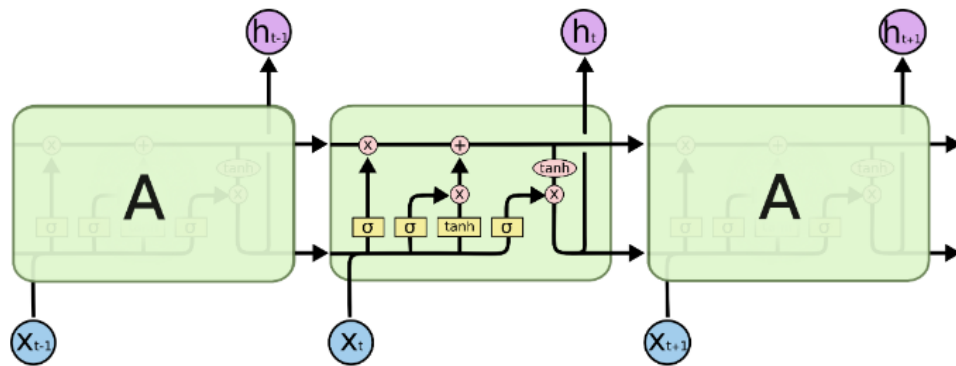


(그림 출처: http://cs231n.stanford.edu/slides/2022/lecture_10_ruohan.pdf)

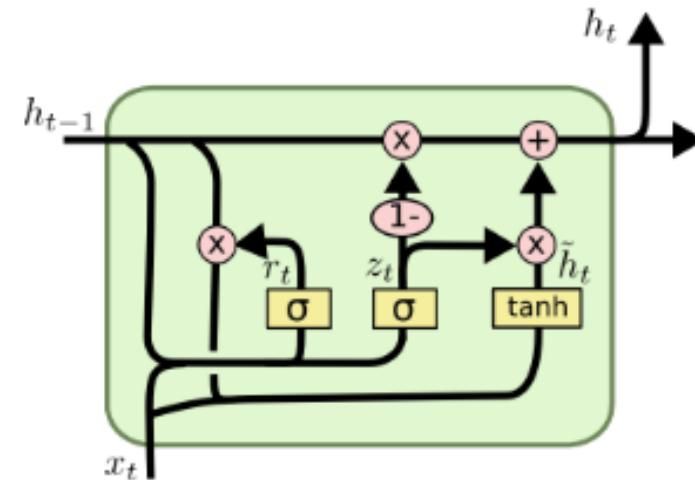
RNN family

- LSTM, GRU

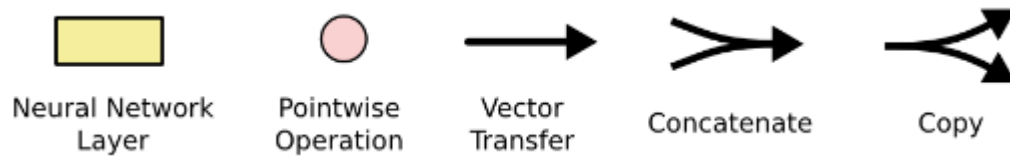
- Long-term dependency를 학습할 수 있도록 변형된 구조의 RNN



LSTM



GRU



(그림 출처: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

Machine Translation

- Translation?

- 특정 언어로 된 text(문장, 구, 절, 문서 등)를 다른 언어로 된 text로 번역하는 과정
 - 특정 언어로 된 text를 '이해'하고, 다른 언어로 된 text를 '생성' 하는 과정
 - Text Understanding + Text Generation
- 즉, 입력 문장(source)을 번역한 출력 문장(target)을 생성하는 task

- Machine Translation

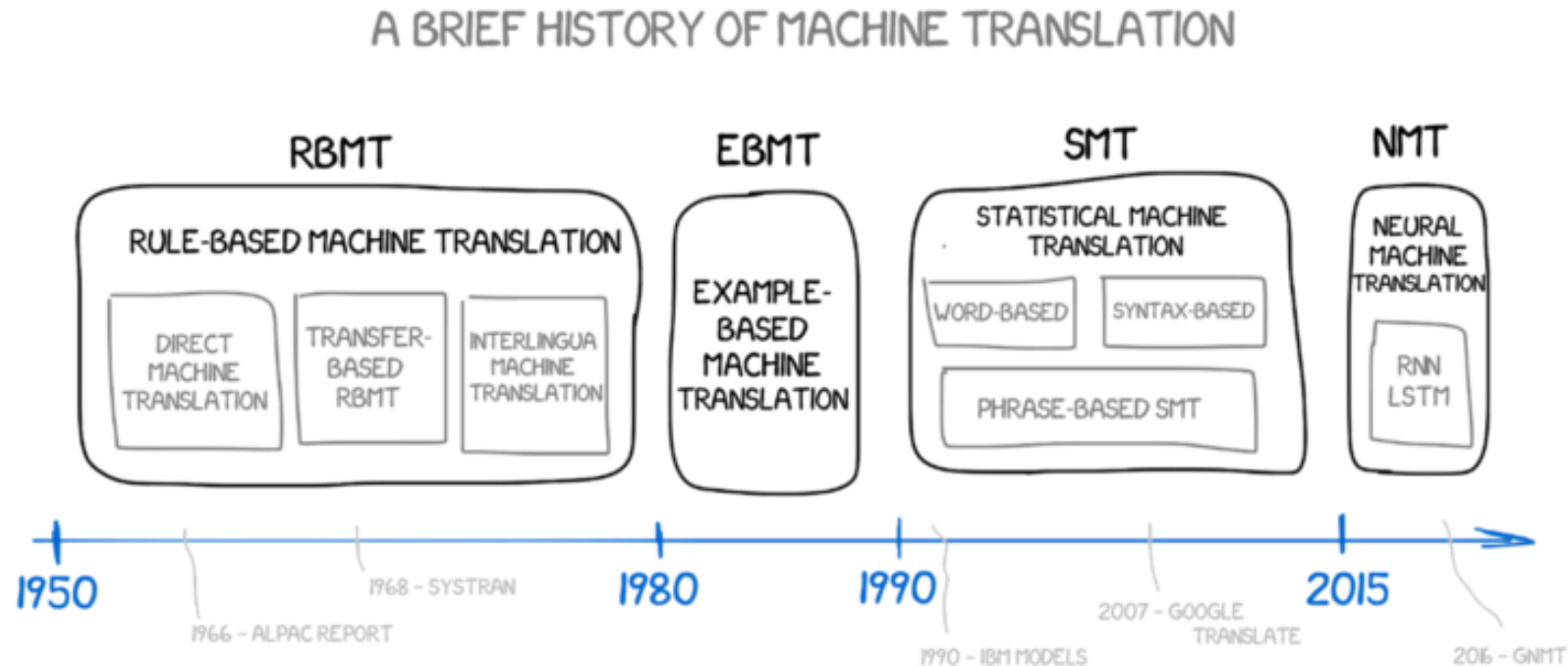
- Translation의 과정을 machine⁰ translation을 하는 과정
- Natural Language Understanding(NLU) + Natural Language Generation(NLG)

(그림 출처: https://www.researchgate.net/figure/An-example-of-a-basic-LSTM-cell-left-and-a-basic-RNN-cell-right-Figure-follows-a_fig2_306377072,
https://github.com/pilsung-kang/Machine-Learning-Basics-Bflysoft/blob/master/Lecture%20RNN_Auto%20Encoder.pdf, <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)

Machine Translation

• Brief History

- **NMT**: 신경망을 사용한 translation
 - 본격적인 성능 향상을 달성 → Neural Machine Translation(NMT)가 본격적으로 시작됨



(그림 출처: <https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/>)

Machine Translation

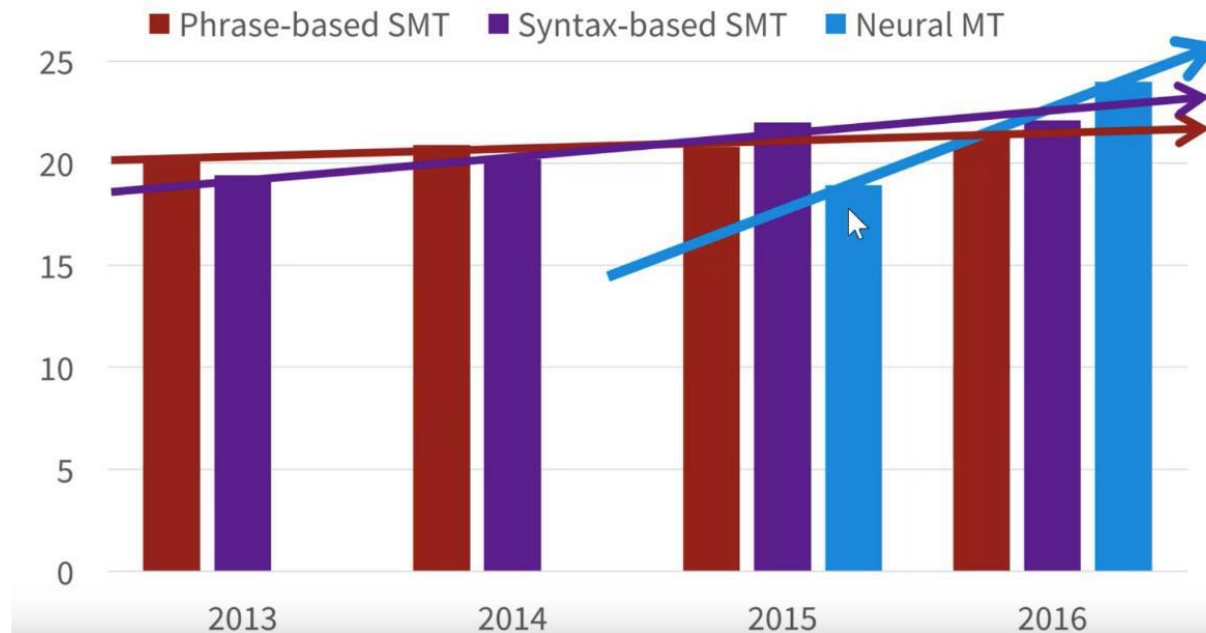
- Brief History

- NMT: 신경망을 사용한 translation

- 본격적인 성능 향상을 달성 → Neural Machine Translation(NMT)가 본격적으로 시작됨

Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



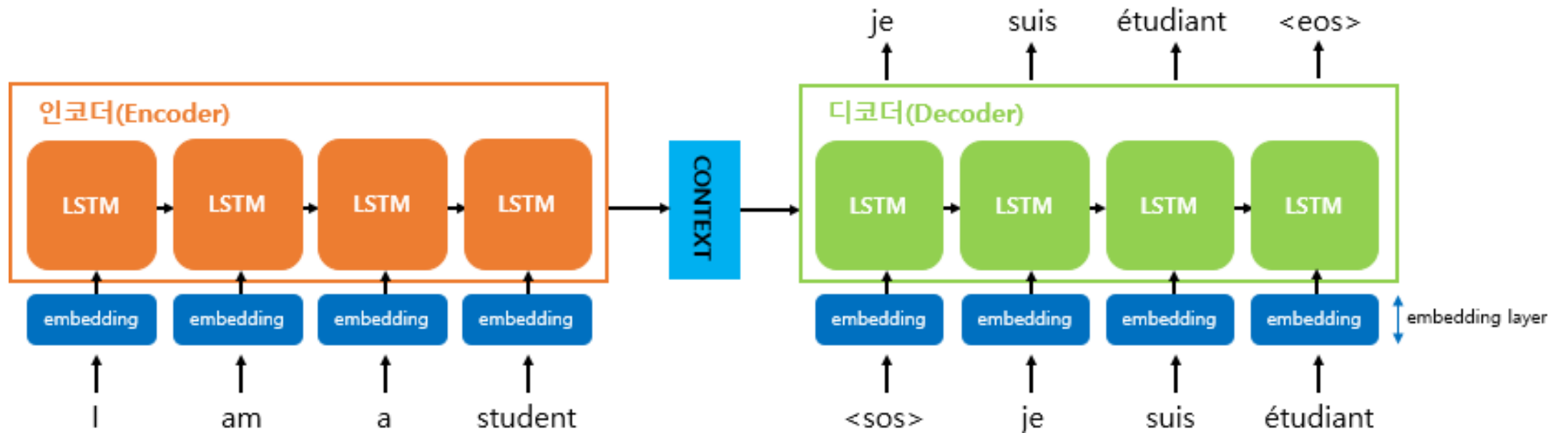
(그림 출처: <https://forums.fast.ai/t/deeplearning-lec11-notes/16407>)

Machine Translation

- Seq2Seq

- Encoder-Decoder구조를 사용한 Machine Translation 모델

- Encoder: Source 문장을 context vector에 압축
 - Decoder: Context vector로부터 문장을 생성

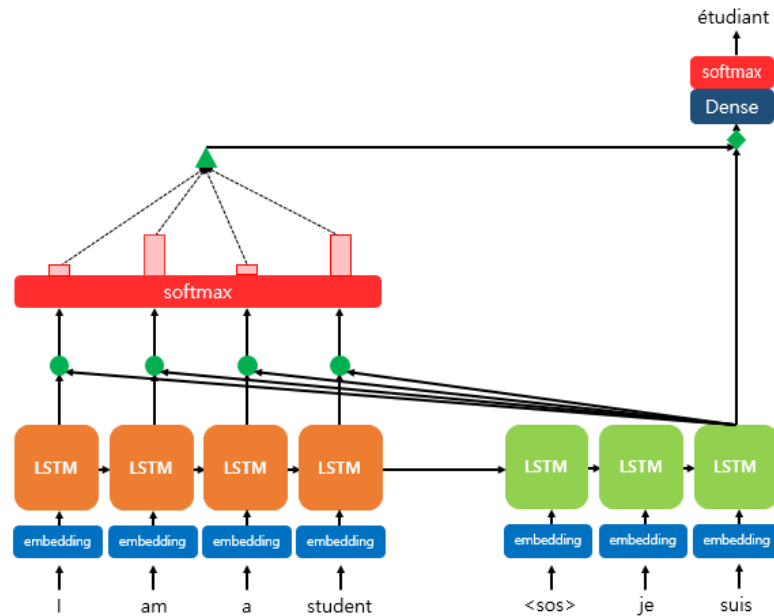


(그림 출처: <https://wikidocs.net/66108>)

Machine Translation

• Seq2Seq + Attention

- Encoder-Decoder구조를 사용한 Machine Translation 모델
- Attention: 어디에 집중할까?
 - Encoder에서 나오는 매 시점의 hidden vector를 decoder로 보냄
 - Decoder가 매 시점 token을 생성할 때, 위에서 나오는 정보를 활용
 - 즉, Decoder가 어떤 단어에 집중하여 출력해야 하는지 고민할 수 있음

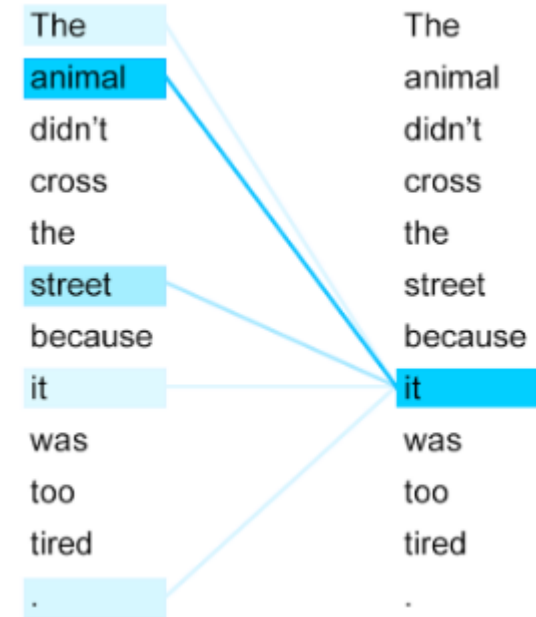
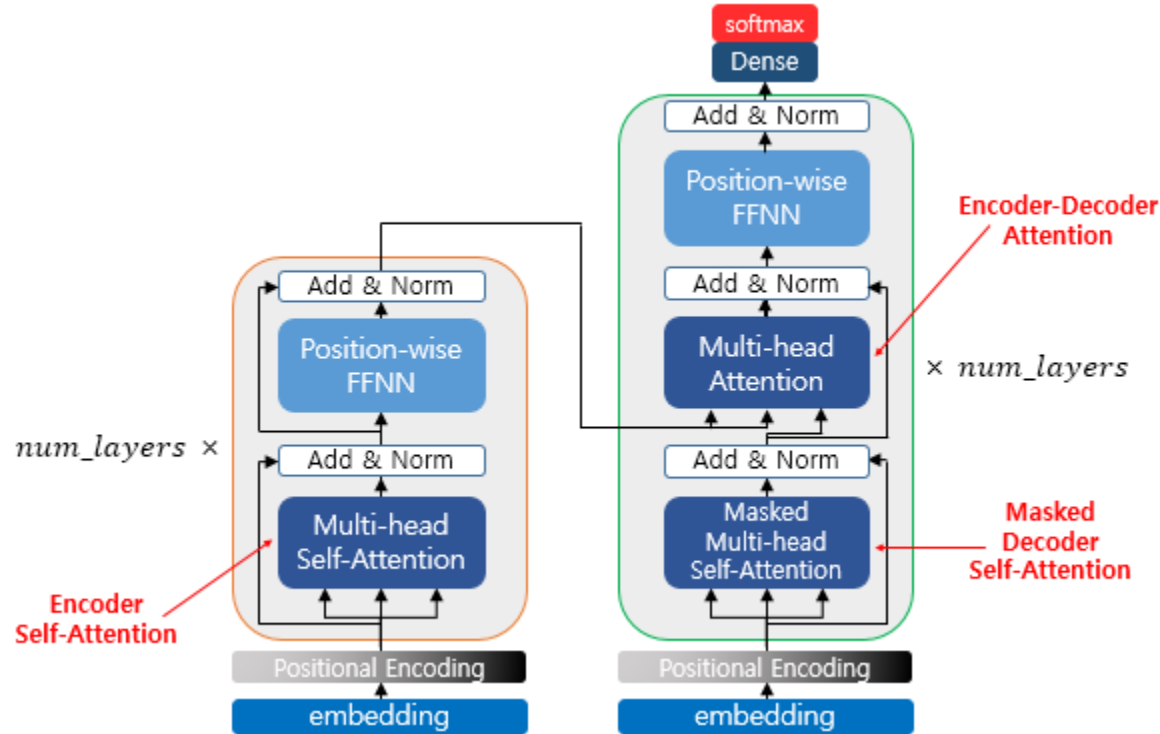


(그림 출처: <https://wikidocs.net/66108>)

Machine Translation

- Transformer

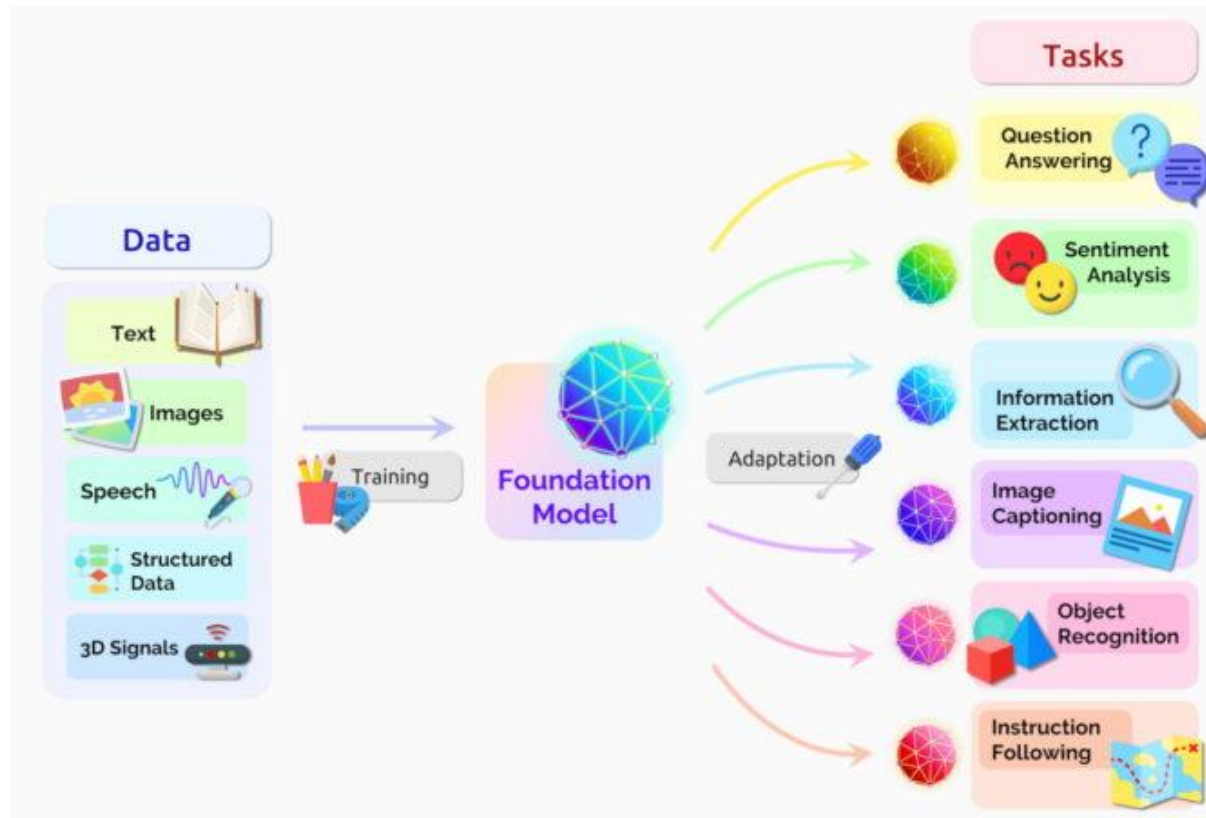
- RNN계열 모델을 사용하지 않고 Encoder-Decoder구조를 구축
 - Used Attention (Self-attention, Encoder-Decoder Attention)



(그림 출처: <https://wikidocs.net/31379>)

Machine Translation

- Transformer가 많은 분야에서 성능이 좋게 나옴
 - 그 중, 하나가 Language Model
 - 특히 Pretrained Language Model



(그림 출처: <https://blogs.nvidia.co.kr/2022/04/01/what-is-a-transformer-model/>)

Language Model (언어 모델)

- 단어 sequence(문장)에 확률을 할당한 모델

- 이를 통해, 가장 자연스러운 단어 sequence를 찾아낼 수 있음
 - 즉, 좀 더 그럴듯한 문장을 선택할 수 있음

아래에서 P는 확률(Probability)을 나타낸다.

기계 번역(Machine Translation)

- $P(\text{나는 버스를 탔다}) > P(\text{나는 버스를 태운다})$

음성 인식(Spell Correction)

- $P(\text{나는 메롱을 먹는다}) < P(\text{나는 메론을 먹는다})$

오타 교정(Spell Correction)

- $P(\text{선생님이 달려갔다}) > P(\text{선생님이 잘려갔다})$



딥 러닝을 이용한 |

딥 러닝을 이용한 부동산가격지수 예측
딥 러닝을 이용한 자연어 처리 입문
딥 러닝을 이용한 한국어 의존 구문 분석
딥 러닝을 이용한 개체명 인식
딥 러닝을 이용한 차량 번호판 검출
딥 러닝을 이용한 한국어 의미역 결정
딥 러닝을 이용한 한국어 형태소의 원형 복원 오류 수정
딥 러닝을 이용한
딥 러닝을 이용한 구문 분석

<검색 엔진에서의 언어 모델이 동작하는 예 : 다음 단어 예측>

(그림 출처: 딥 러닝을 이용한 자연어 처리 입문, 안상준, 유원준 저, Wikibooks)

Language Model (언어 모델)

- 단어 sequence(문장)에 확률을 할당한 모델

- 어떻게 문장에 확률을 할당할까?

- 전체 말뭉치에서 문장을 구성하는 단어들이 등장한 확률 구하기

$$P(\text{나는 버스를 탔다}) = P(\text{나는, 버스를, 탔다})$$

- 수식으로 표현하자면,

- 어떤 단어 sequence W 가 n 개의 단어($w_i, i = 1, 2, \dots, n$)들로 구성되어 있다.
 - 이 때, 단어 sequence W 가 등장할 확률은 다음과 같이 나타낼 수 있다.

$$P(W) = P(w_1, w_2, \dots, w_n)$$

Language Model (언어 모델)

- 단어 sequence(문장)에 확률을 할당한 모델

- Statistical Language Model(통계적 언어 모델, SLM)

- Language Model은 다음과 같이 조건부 확률을 활용하여 표현 가능
 - 어떤 단어 sequence W 가 n 개의 단어($w_i, i = 1, 2, \dots, n$)들로 구성되어 있다.
 - 이 때, 단어 sequence W 가 등장할 확률은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} P(W) &= P(w_1, w_2, \dots, w_n) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n|w_1, w_2, \dots, w_{n-1}) \\ &= \cancel{P(w_1)} \frac{\cancel{P(w_1, w_2)}}{\cancel{P(w_1)}} \frac{\cancel{P(w_1, w_2, w_3)}}{\cancel{P(w_1, w_2)}} \cdots \frac{P(w_1, w_2, \dots, w_n)}{\cancel{P(w_1, w_2, \dots, w_{n-1})}} \\ &= \prod_{i=1}^n P(w_i|w_1, w_2, \dots, w_{i-1}) \end{aligned}$$

Language Modeling (언어 모델링)

- 주어진 단어로부터 아직 모르는 단어를 예측하는 작업
 - Method: 이전 단어들이 주어졌을 때, 다음 단어를 예측
 - 즉, 언어 모델이 이전 단어들로부터 다음 단어를 예측하는 task
 - How?

나는 버스를 _____ ➡ 나는 버스를 탔다

찾다
택시
먹다
탔다
⋮

Language Modeling (언어 모델링)

• Language Modeling with SLM

- 주어진 단어로부터 아직 모르는 단어를 예측하는 작업
 - Method: 이전 단어들이 주어졌을 때, 다음 단어를 예측
- 즉, 언어 모델이 이전 단어들로부터 다음 단어를 예측하는 task 단어의 분산표현
 - 주어진 단어들로부터 그 다음에 각 단어가 올 확률을 계산 → 가장 확률이 높은 단어 선택!

나는 버스를 _____	➡	나는 버스를 <u>탔다</u>	←	예측
찾다		$P(\text{찾다} \text{나는, 버스를}) = 0.11$		
택시		$P(\text{택시} \text{나는, 버스를}) = 0.01$		
먹다		$P(\text{먹다} \text{나는, 버스를}) = 0.05$		
탔다		$P(\text{탔다} \text{나는, 버스를}) = 0.25$ ➡ Maximum!		
⋮		⋮		

→ 그럼 각 단어의 조건부 확률은 어떻게 구할까?

→ 즉, $P(\text{버스를}|\text{나는})$, $P(\text{탔다}|\text{나는, 버스를})$ 등의 값은 어떻게 구할까?

(그림 출처: 딥 러닝을 이용한 자연어 처리 입문, 안상준, 유원준 저, Wikibooks)

n-gram Language Model

- n-gram

- n개의 연속적인 단어 나열 (n은 사용자가 정하는 값)
- 이전 단어들이 주어졌을 때, 다음 단어의 등장 확률은 앞의 n-1개의 단어에만 dependent

오타니쇼헤이는 일본 국적의 투타 겸업을 하는 야구선수로, 2021년 아메리칸리그 MVP를 _____

- Bigram (n=2)

$$P(w|MVP를) = \frac{P(MVP를, w)}{P(MVP를)}$$

- Trigram (n=3)

$$P(w|아메리칸리그, MVP를) = \frac{P(아메리칸리그, MVP를, w)}{P(아메리칸리그, MVP를)}$$

- 4-gram (n=4)

$$P(w|2021년, 아메리칸리그, MVP를) = \frac{P(2021년, 아메리칸리그, MVP를, w)}{P(2021년, 아메리칸리그, MVP를)}$$

Neural Network-based Language Model (NNLM)

- FFNN으로 구현한 language model

- n 개의 이전 단어로부터 $n+1$ 번째의 단어를 예측
- 신경망에 입력하는 값과 정답(label) 값은 one-hot encoding vector
- 예시: $n=3$ 인 NNLM
 - 등장하는 단어: what, will, the, fat, cat, sit, on

단어	one-hot
what	[1,0,0,0,0,0,0]
will	[0,1,0,0,0,0,0]
the	[0,0,1,0,0,0,0]
fat	[0,0,0,1,0,0,0]
cat	[0,0,0,0,1,0,0]
sit	[0,0,0,0,0,1,0]
on	[0,0,0,0,0,0,1]

what will the fat cat _

애를 예측해봅시다.
 $n=3 \rightarrow$ the, fat, cat을 활용하여 예측

입력	
단어	one-hot
the	[0,0,1,0,0,0,0]
fat	[0,0,0,1,0,0,0]
cat	[0,0,0,0,1,0,0]

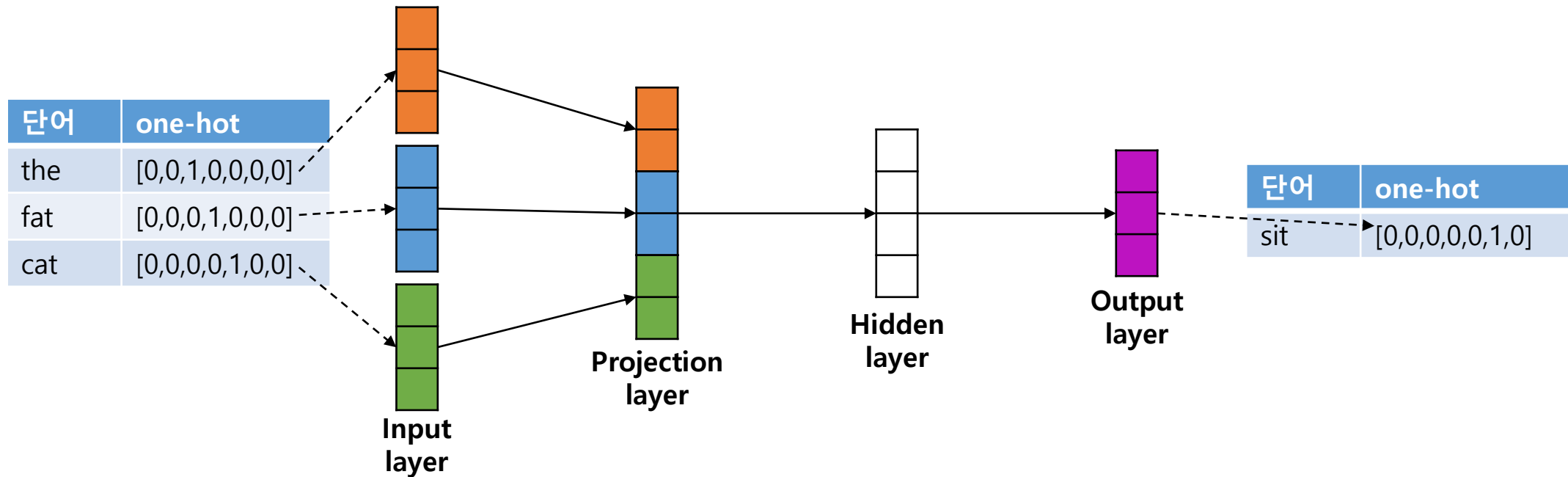
출력(예측 대상)

단어	one-hot
sit	[0,0,0,0,0,1,0]

Neural Network-based Language Model (NNLM)

- FFNN으로 구현한 language model

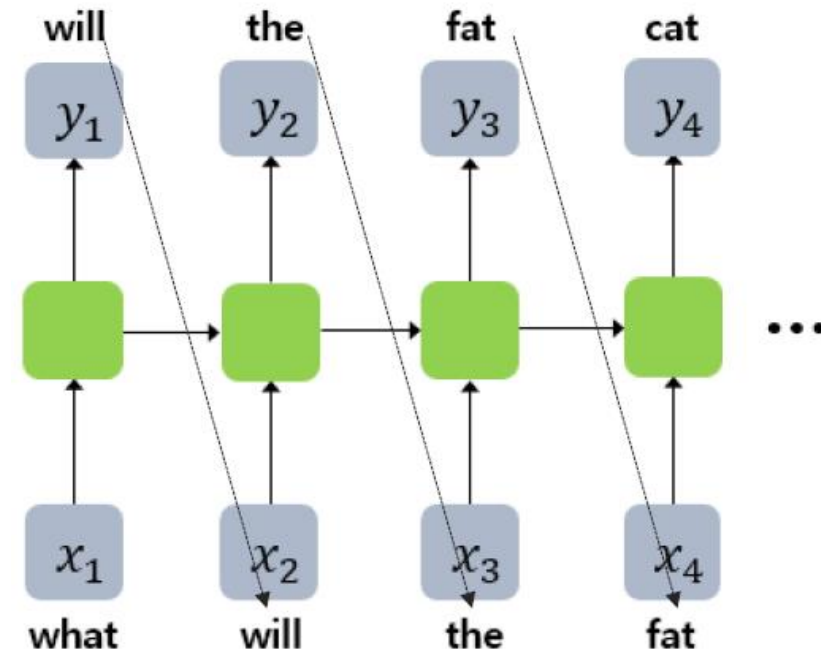
- n 개의 이전 단어로부터 $n+1$ 번째의 단어를 예측
- 신경망에 입력하는 값과 정답(label) 값은 one-hot encoding vector
- 예시: $n=3$ 인 NNLM
 - 등장하는 단어: what, will, the, fat, cat, sit, on



RNN based Language Model (RNNLM)

- RNN으로 구현된 Language Model
 - Language Model
 - 이전 단어들을 활용하여 다음 단어를 예측
 - Variable number of input

예시: What will the fat cat sit on

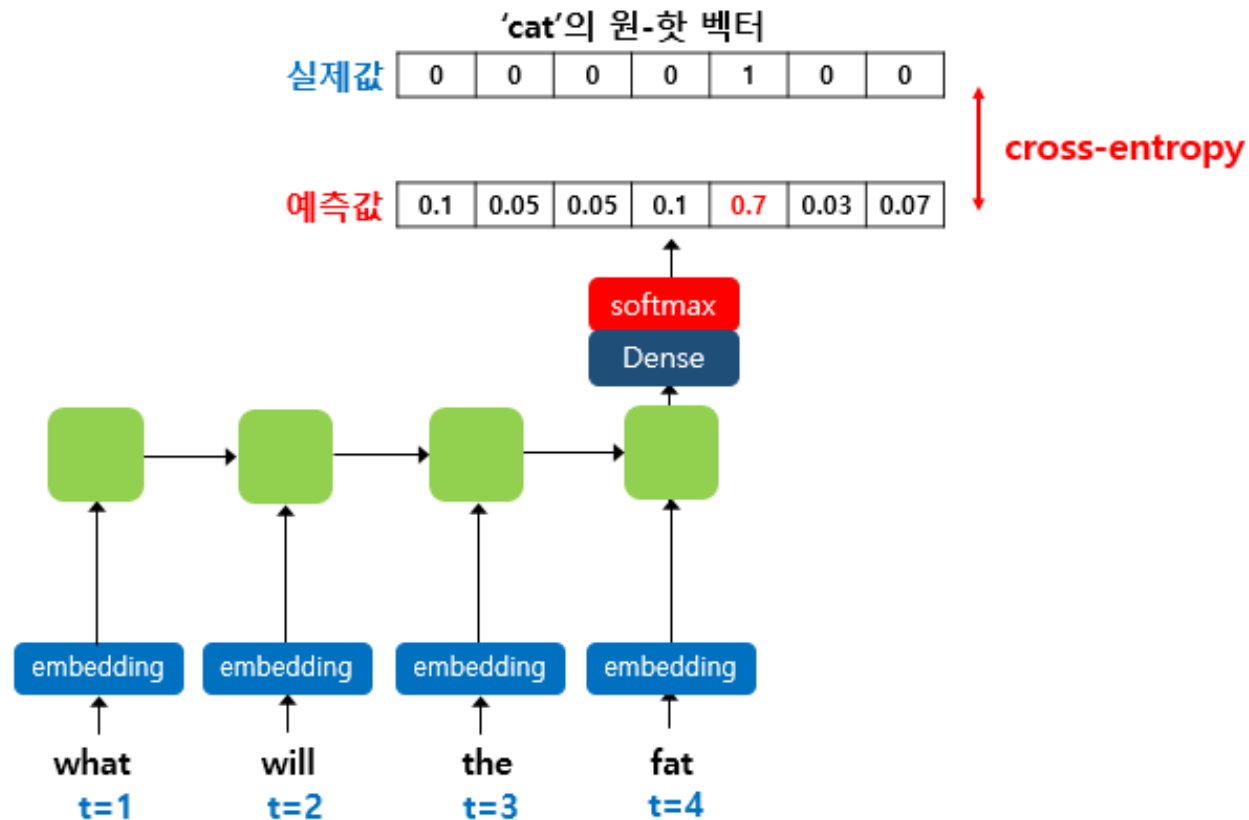


(그림 출처: 딥 러닝을 이용한 자연어 처리 입문, 안상준, 유원준 저, Wikibooks)

RNN based Language Model (RNNLM)

- Architecture

- Embedding layer, Hidden layer, Output layer로 구성
- Output layer에서는 Vocabulary set의 크기 만큼의 벡터를 생성 → Multi-class 분류



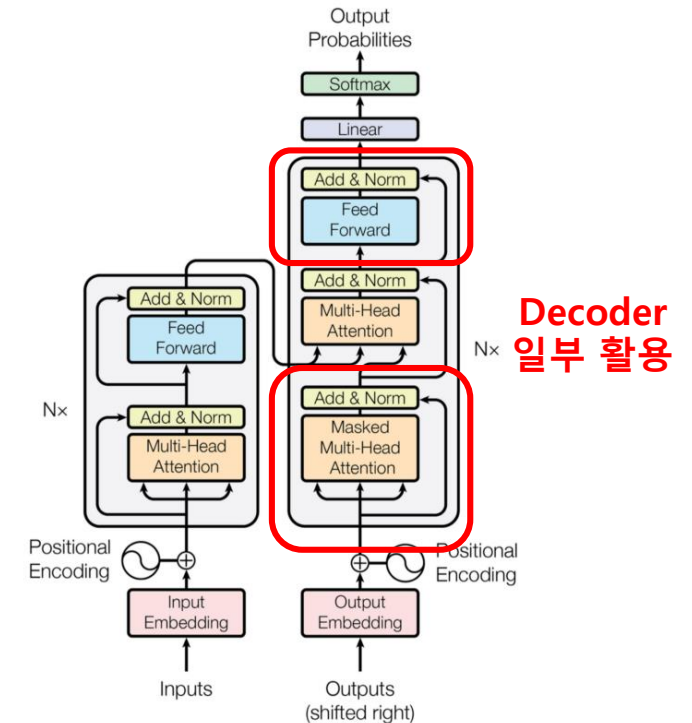
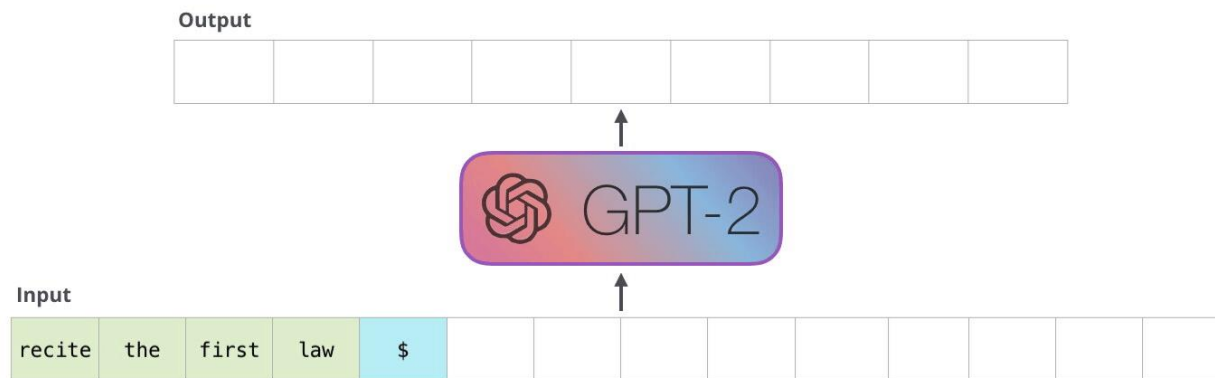
(그림 출처: 딥 러닝을 이용한 자연어 처리 입문, 안상준, 유원준 저, Wikibooks)

Pretrained Language Model

- Generative Pre-trained Transformer(GPT)

- OpenAI (2018)

- 생성하는(generative): 단어가 들어오면 다음에 올 단어를 '생성'하는 언어모델
 - 사전 학습된 (pre-trained): 대규모의 데이터로 미리 모델을 학습시킴
 - 트랜스포머(transformer): transformer의 Decoder구조를 활용



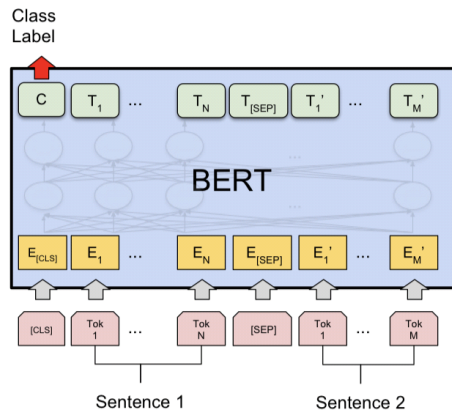
(그림 출처: <https://jalamar.github.io/illustrated-gpt2/>, Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.)

Pretrained Language Model

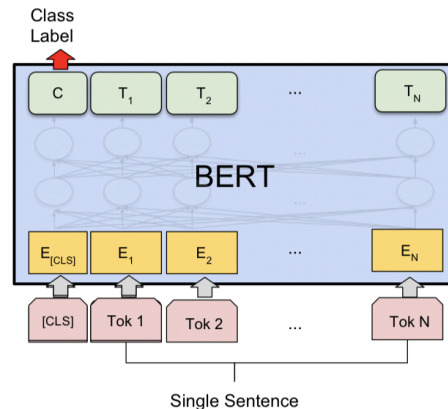
• Bidirectional Encoder Representations from Transformers (BERT)

– Google (2019)

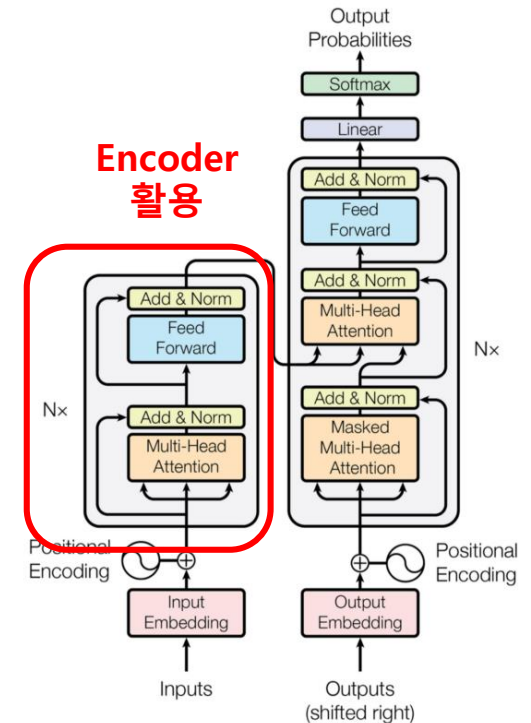
- 양방향의(Bidirectional): 양방향으로 구성된
- Encoder Representations: Encoder로 학습한 텍스트 표현 방법
- 트랜스포머(transformer): Transformer의 Encoder를 활용
- 학습 시 여러 개의 pretrained task를 활용하여 학습
 - BookCorpus(800M) + Wikipedia(2500M)



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA

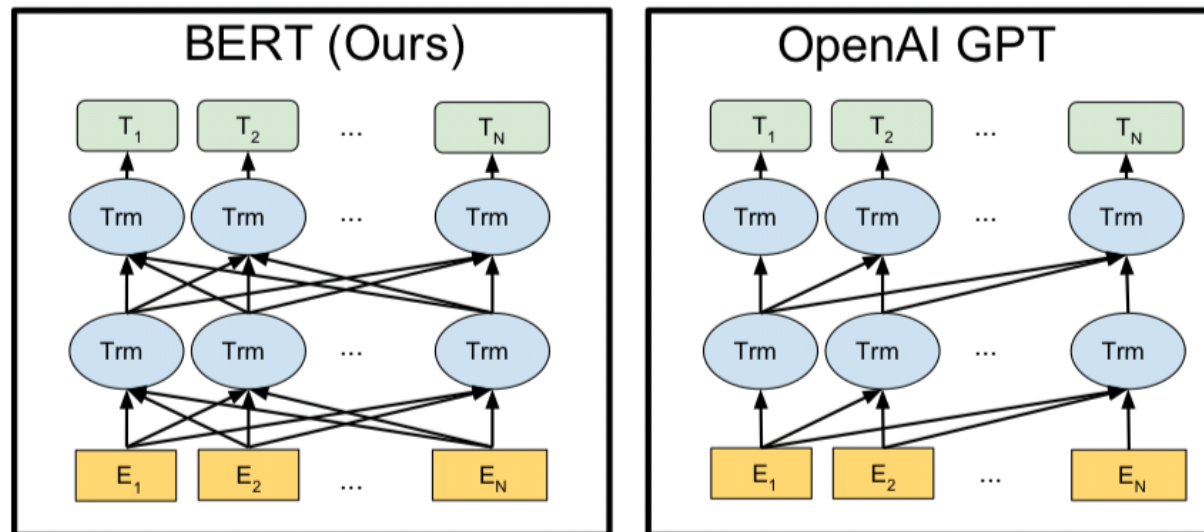


(그림 출처: <https://tmaxai.github.io/post/BERT/>)

Pretrained Language Model

- BERT vs GPT

BERT에서는 다양한 task를 다룰 수 있음



(그림 출처: <https://wikidocs.net/115055>)

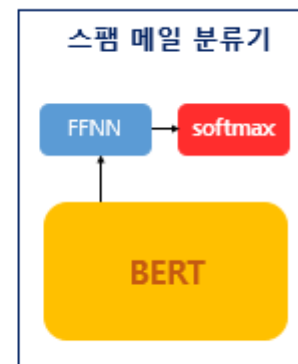
Pretrained Language Model

- Pretrained LM을 활용하여 다양한 task를 수행할 수 있음

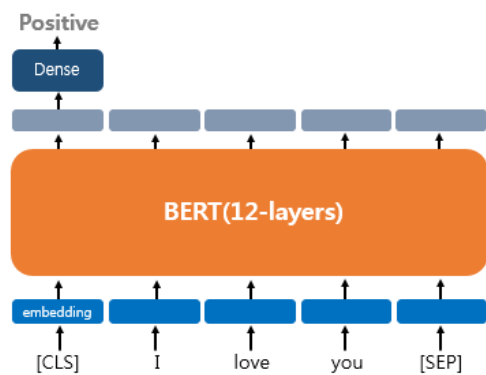


33억 단어에 대해서 4일간 학습시킨 언어 모델

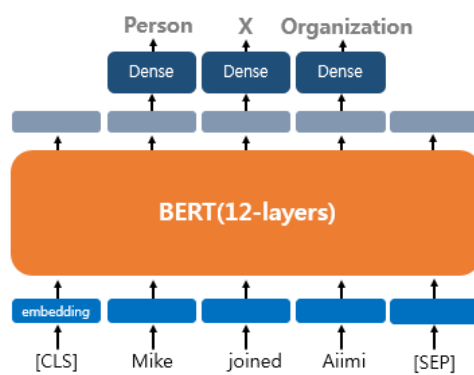
조금만 튜닝(Tuning)해서
다른 용도로 사용한다면?



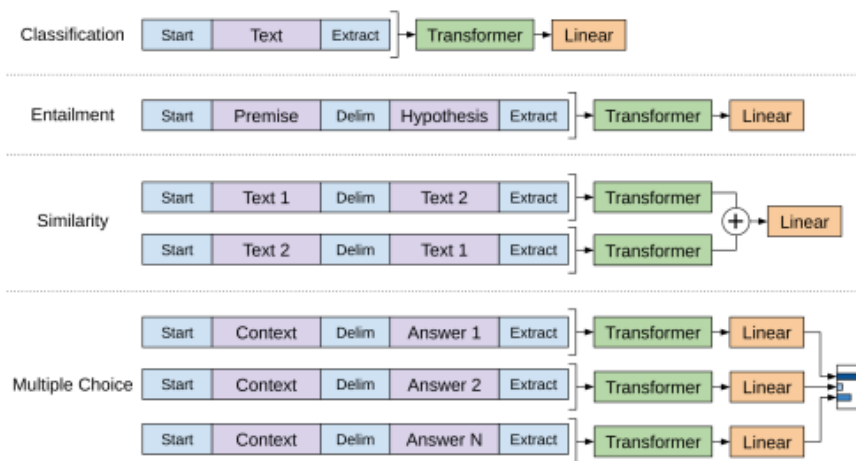
BERT의 지식을 이용한 스팸 메일 분류기



Text Classification



Tagging



(그림 출처: <https://wikidocs.net/115055>, Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.)

Pretrained Language Model

- 다양한 모델들이 파생되었습니다.
 - ROBERTa
 - KoBERT
 - BART
 - KeyBERT
 - ...

2. Text Analytics Research Review

Natural Language Processing (NLP) vs Text Analytics

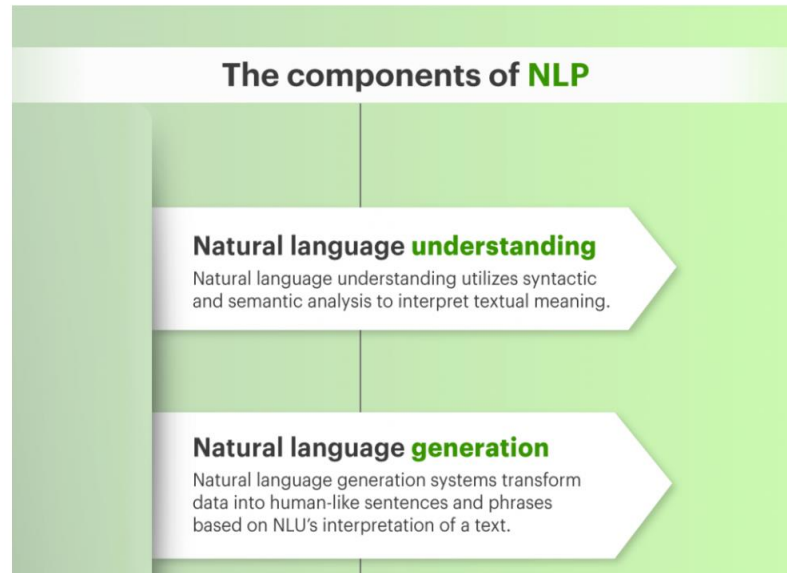
- 혼용해서 사용하는 경우가 많습니다

- 다만, 개인적으로는 다른 영역이라고 생각합니다 NLP를 잘해야 Text Analytics를 잘 할 수 있다.

- **NLP:** Machine이 사람의 언어(Natural Language)를 이해하도록 하는 과정
 - **Text Analytics:** NLP, ML을 활용하여 다양한 Text를 분석, insight를 도출하는 과정

As Ryan explains, "when it comes to NLP, Google has a definition that I think summarizes it well. NLP is 'the application of computational techniques to the analysis and synthesis of natural language and speech.'"

Generally, natural language processing consists of two components:



Then comes text mining...

Text analysis – or text mining – can be hard to understand, so we asked Ryan how he would define it in a sentence or two.

In his words, **text analytics** is "extracting information and insight from text using AI and NLP techniques. These techniques turn unstructured data into structured data to make it easier for data scientists and analysts to actually do their jobs."

Wait, so are NLP and text mining the same?

Related? Yes. The same? No.

In simple terms, NLP is a technique that is used to *prepare* data for analysis. As humans, it can be difficult for us to understand the need for NLP, because our brains do it automatically (we understand the meaning, sentiment, and structure of text without processing it). But because computers are (thankfully) not humans, they need NLP to make sense of things.

As Ryan explains, "language is full of different layers which all work together. Humans combine all these layers with ease, but it is much more labor intensive for computers." And now we're talking about layers, it naturally makes sense to look to our favorite Scottish ogre for inspiration. "To take an analogy from Shrek", observes Ryan, "language is truly like an onion. NLP picks apart that onion, identifying each layer."

A little unconventional as an analogy, but we'll bite. Ryan continues: "Now, you generally wouldn't just eat all those layers of onion on their own, you would do something with them. Analysts can swoop in and make a nice French onion soup with all those layers. The recipes they use may be varied, but ultimately, their goal is the same: to make something beautiful – aka, extract meaning and insight."

The difference between text mining vs. NLP? They're not the same but closely connected. In short, you can have NLP without text analytics, but it would be difficult to do text analytics without NLP.

(Source: <https://relativeinsight.com/text-mining-vs-nlp/>)

NLP를 활용한 (혹은 관련된) 연구 분야

- Named Entity Recognition (NER)

The screenshot displays a text processing interface for Named Entity Recognition. At the top, a legend bar identifies six entity types with their corresponding labels: Person (p), Loc (l), Org (o), Event (e), Date (d), and Other (z). Below this, a paragraph of text is shown with various entities highlighted in colored boxes and marked with a small 'x' icon. The entities and their labels are: 'Barack Hussein Obama II' (Person, p), 'August 4, 1961' (Date, d), 'American' (Other, z), 'the United States' (Loc, l), 'January 20, 2009' (Date, d), 'January 20, 2017' (Date, d), 'Democratic Party' (Org, o), 'African American' (Other, z), 'United States Senator' (Other, z), 'Illinois' (Loc, l), and 'Illinois State Senate' (Org, o).

Person p Loc l Org o Event e Date d Other z

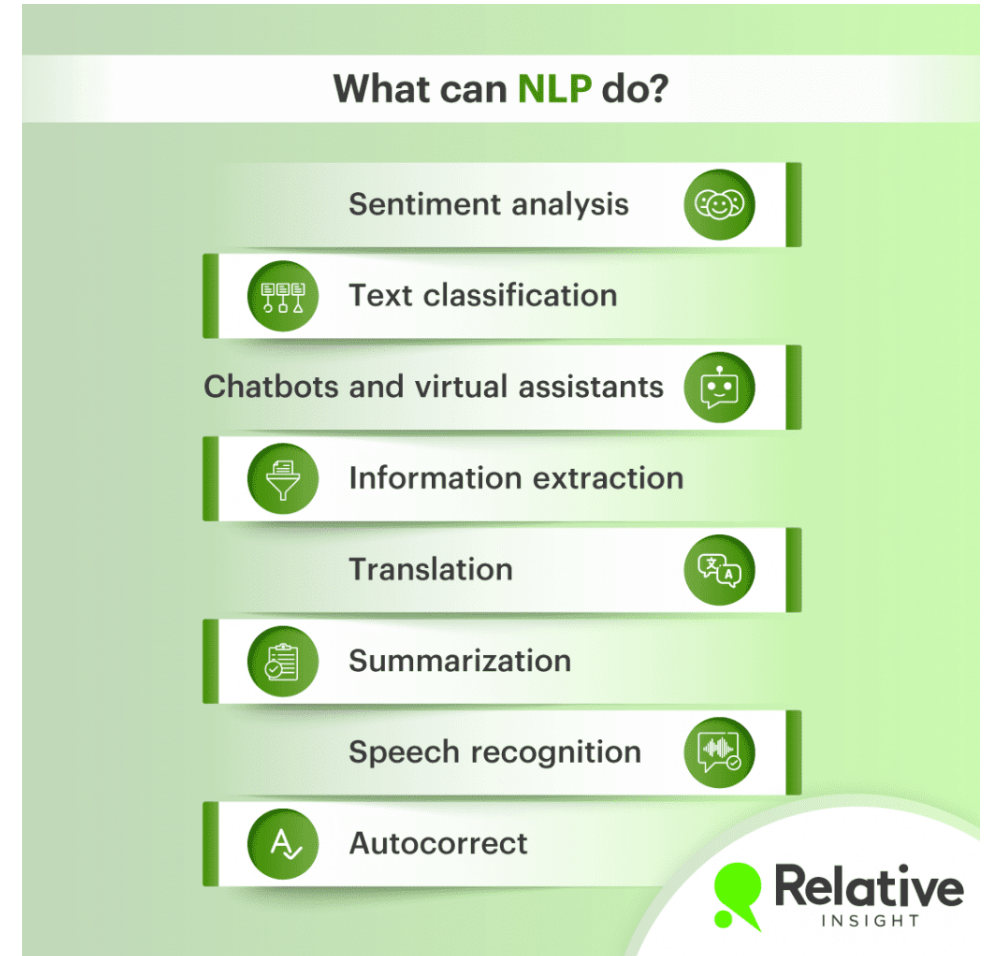
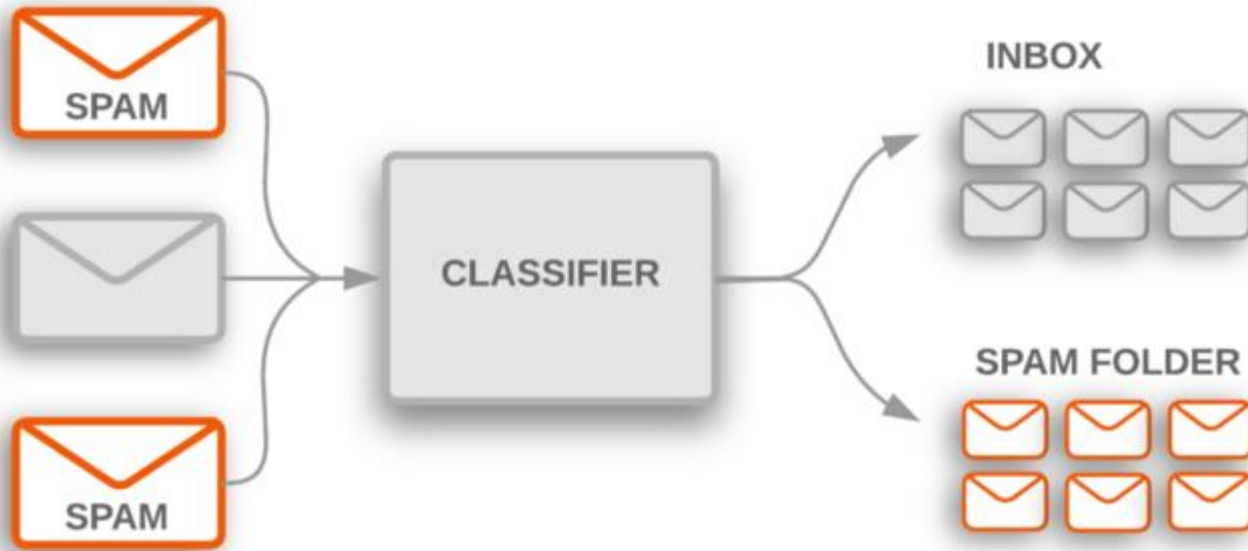
Barack Hussein Obama II x (born August 4, 1961 x) is an American x attorney and politician who served as the 44th President of the United States x from January 20, 2009 x, to January 20, 2017 x. A member of the Democratic Party x, he was the first African American x to serve as president. He was previously a United States Senator x from Illinois x and a member of the Illinois State Senate x.

(Source: <https://nlpcloud.com/>)

NLP를 활용한 (혹은 관련된) 연구 분야

- Text Classification

- 예시: Spam Filtering

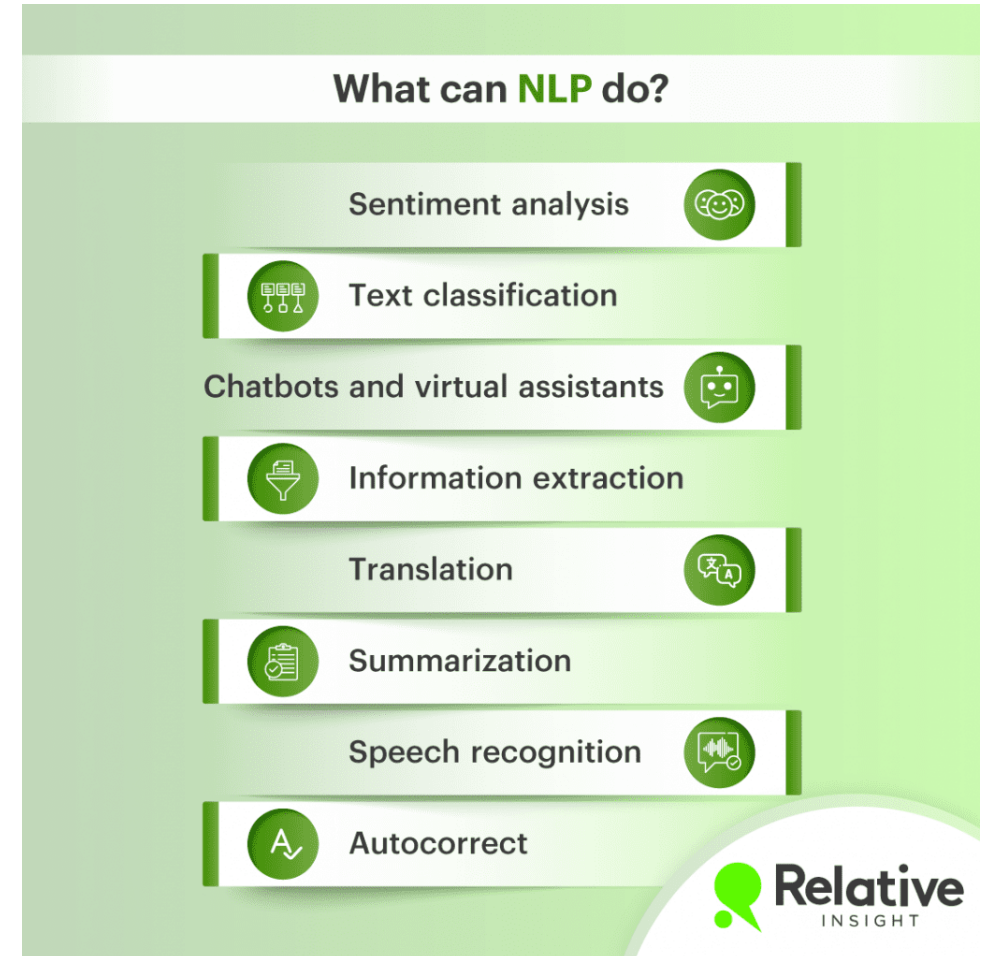
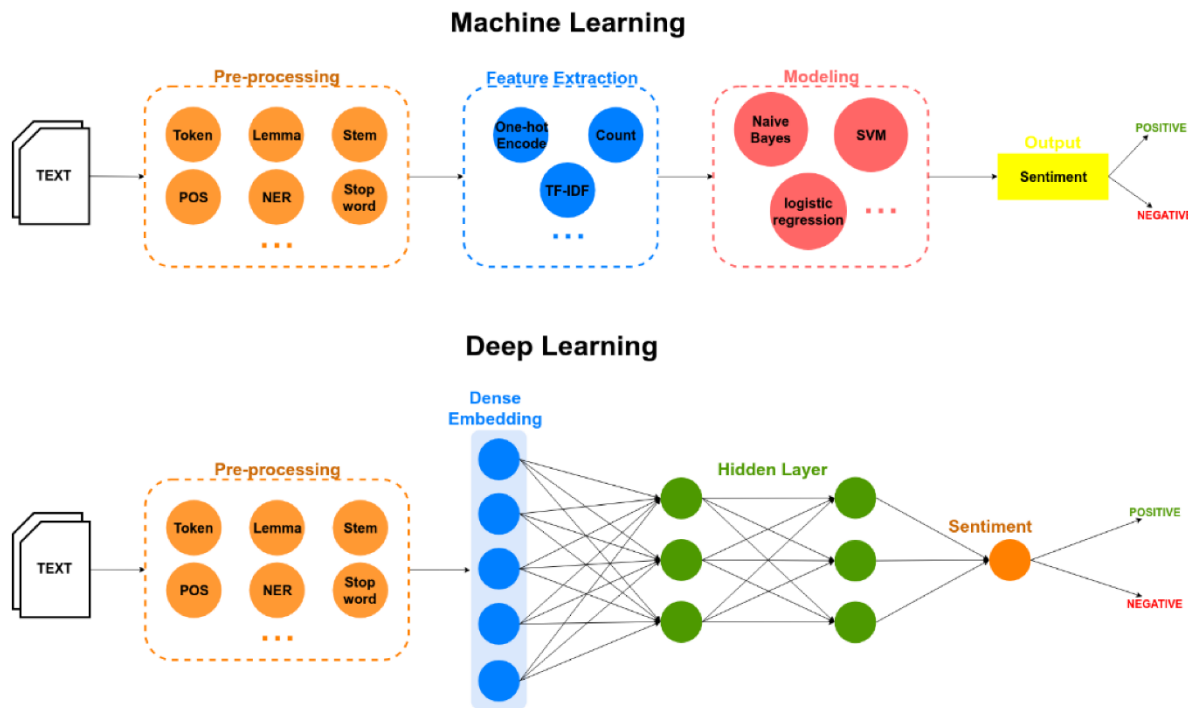


(Source: <https://medium.com/@naveeen.kumar.k/naive-bayes-spam-detection-7d087cc96d9d>, <https://relativeinsight.com/text-mining-vs-nlp/>)

NLP를 활용한 (혹은 관련된) 연구 분야

• Text Classification

– 예시: Sentiment Analysis



(Source: <https://medium.com/@jaylikesmessi/sentiment-analysis-using-nlp-libraries-dfb8219e0b35>, <https://relativeinsight.com/text-mining-vs-nlp/>)

NLP를 활용한 (혹은 관련된) 연구 분야

- NLU

- 예시: Question Answering

Airport

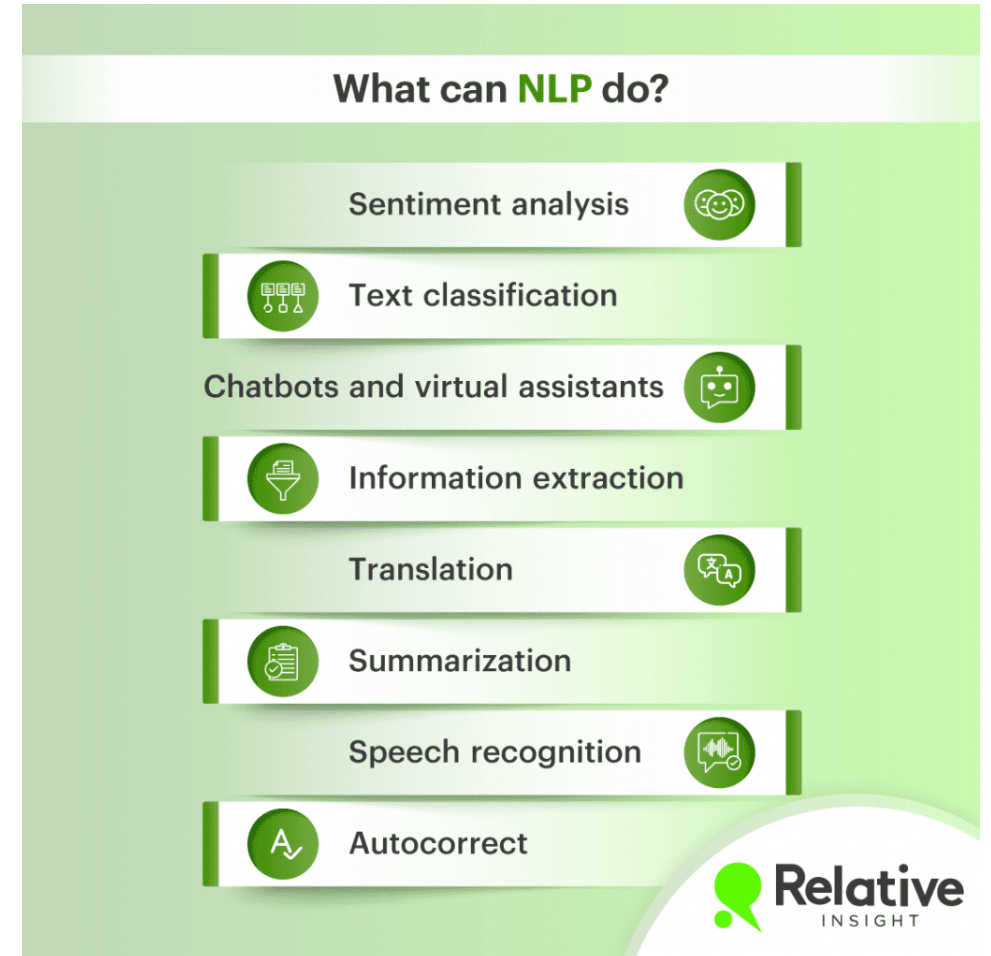
The Stanford Question Answering Dataset

An **airport** is an aerodrome with **facilities** for **flights** to take off and **land**. Airports often have **facilities** to store and maintain aircraft, and a control tower. An **airport** consists of a **landing** area, which comprises an aerially accessible open space including at least one operationally active surface such as a runway for a plane to take off or a helipad, and often includes adjacent utility buildings such as control towers, hangars and terminals. Larger airports may have fixed base operator services, **airport** aprons, air traffic control centres, passenger **facilities** such as restaurants and lounges, and emergency services.

What is an aerodrome with facilities for flights to take off and land?
airport

What is an aerially accessible open space that includes at least one active surface such as a runway or a helipad?
landing area

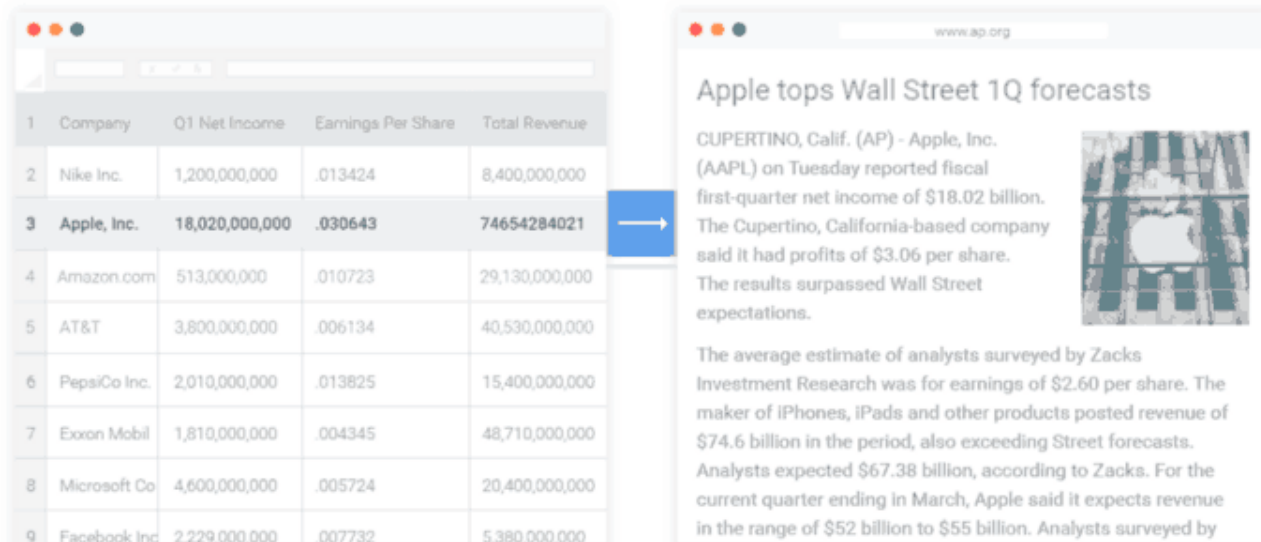
What is an airport?
aerodrome with facilities for flights to take off and land



(Source: <https://tensorflow.blog/2016/06/20/squad-stanford-question-answering-dataset/>, <https://relativeinsight.com/text-mining-vs-nlp/>)

NLP를 활용한 (혹은 관련된) 연구 분야

- NLG



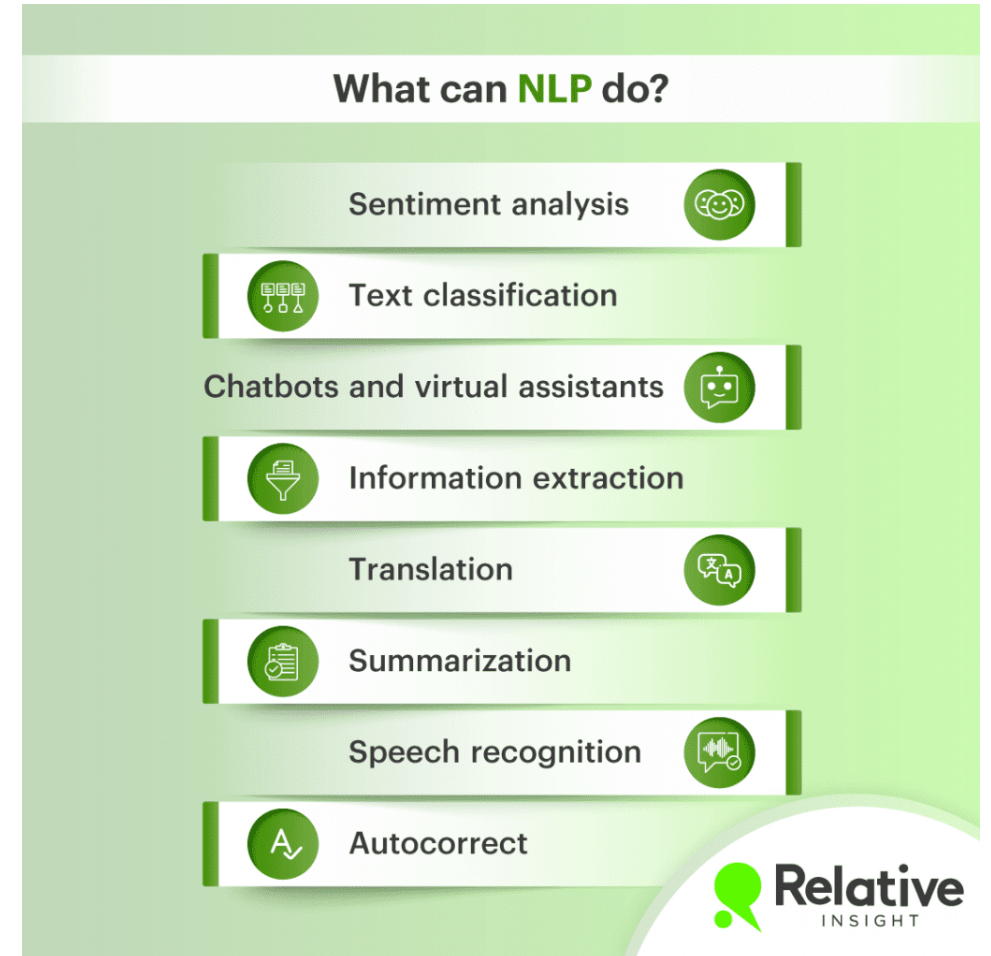
The image shows two side-by-side browser windows. The left window displays a financial table with columns for Company, Q1 Net Income, Earnings Per Share, and Total Revenue. The right window shows a news article titled 'Apple tops Wall Street 1Q forecasts' from AP.org, with an arrow pointing from the 'Apple, Inc.' row in the table to the article.

1	Company	Q1 Net Income	Earnings Per Share	Total Revenue
2	Nike Inc.	1,200,000,000	.013424	8,400,000,000
3	Apple, Inc.	18,020,000,000	.030643	74654284021
4	Amazon.com	513,000,000	.010723	29,130,000,000
5	AT&T	3,800,000,000	.006134	40,530,000,000
6	PepsiCo Inc.	2,010,000,000	.013825	15,400,000,000
7	Exxon Mobil	1,810,000,000	.004345	48,710,000,000
8	Microsoft Co	4,600,000,000	.005724	20,400,000,000
9	Facebook Inc	2,229,000,000	.007732	5,380,000,000

Apple tops Wall Street 1Q forecasts

CUPERTINO, Calif. (AP) - Apple, Inc. (AAPL) on Tuesday reported fiscal first-quarter net income of \$18.02 billion. The Cupertino, California-based company said it had profits of \$3.06 per share. The results surpassed Wall Street expectations.

The average estimate of analysts surveyed by Zacks Investment Research was for earnings of \$2.60 per share. The maker of iPhones, iPads and other products posted revenue of \$74.6 billion in the period, also exceeding Street forecasts. Analysts expected \$67.38 billion, according to Zacks. For the current quarter ending in March, Apple said it expects revenue in the range of \$52 billion to \$55 billion. Analysts surveyed by

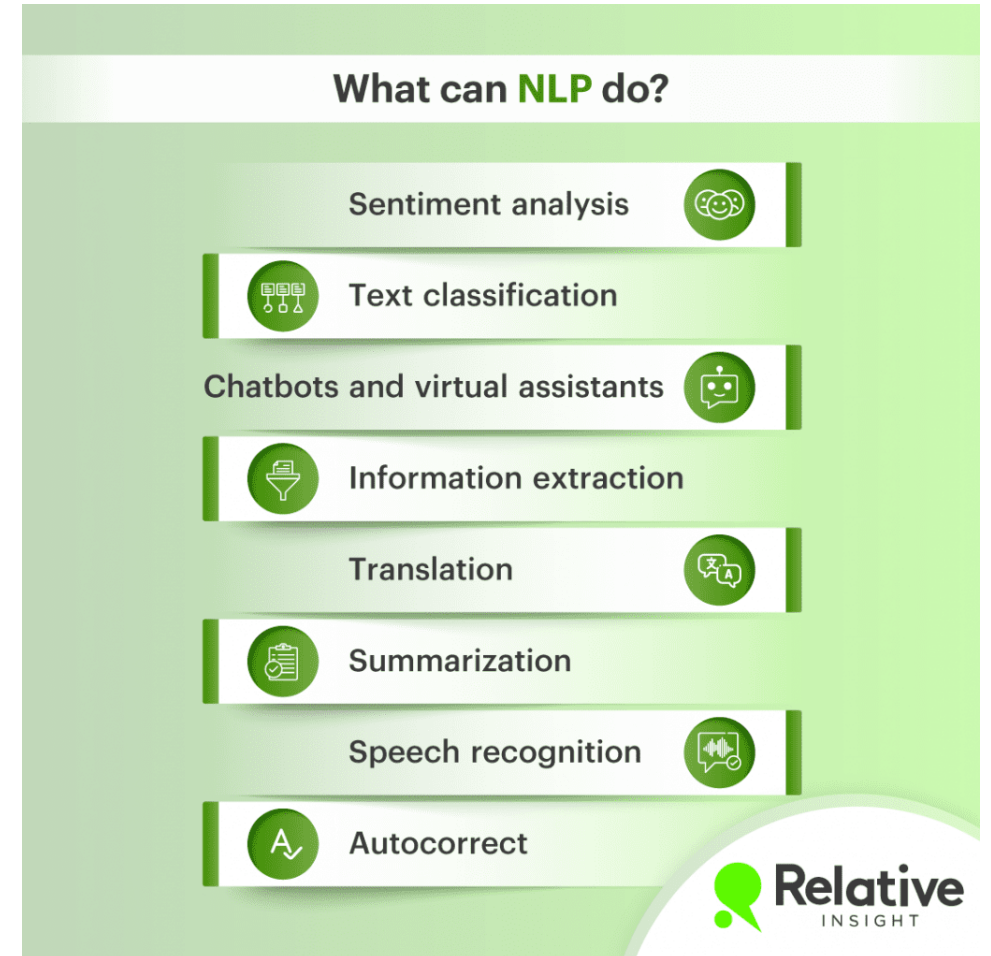


(Source: <https://medium.com/@jaylikesmessi/sentiment-analysis-using-nlp-libraries-dfb8219e0b35>, <https://relativeinsight.com/text-mining-vs-nlp/>)

NLP를 활용한 (혹은 관련된) 연구 분야

• NLU & NLG

– 예시: Chatbot



(Source: <https://blog.happytalk.io/insight/?idx=13017508&bmode=view>, https://luda.ai/luda_thumbnail.jpg, <https://relativeinsight.com/text-mining-vs-nlp/>)

Text Analytics 관련 연구분야

- Opinion mining

텍스트 마이닝 기법을 활용한 게임소비자 인식에 관한 연구:
온라인커뮤니티 리뷰(Reddit)를 중심으로*

한국컴퓨터정보학회 하계 학술대회 논문집 제28권 제2호 (2020. 7)

텍스트 마이닝 기법을 활용한 웹툰 댓글 분석 :
네이버 베스트 도전 웹툰을 중심으로



디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 6, pp. 1209-1220, Jun. 2023

챗GPT 관련 사회적 이슈에 대한 탐색적 연구: 뉴스 빅데이터 기반 토픽 모델링 분석을
중심으로



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine
learning on COVID-19 vaccination Twitter dataset

Text Analytics 관련 연구분야

• Industry-related Text analytics

Year	Common Concern	Federal Reserve System	European Central Bank	Deutsche Bundesbank	Bank of England	Bank of Japan
2004	Sustainability Credibility	Expansion Imports	Parliaments Cooperation	Retirement Ages Working Hours	Household-Spending	QE Deflation
2005	China Inflation	Deficits Competitive	Financial Integration	Global Imbalance	Households Future Inflation	QE Recession
2006	Competitive Global Imbalance	Incentives Risk Taking	Administered Price Indirect Taxes	Inflation	China / India Commodities	Domestic and-External Demand
2007	Subprime-Mortgage	Subprime-Mortgage	Price Stability Turmoil	Banking Supervision Disclosure	Credit	Subprime-Mortgage
2008	Financial Turmoil Commodity Prices	Financial Turmoil Funding Markets	Financial Turmoil Liquidity	Financial Turmoil Subprime	Commodity Prices Housing Market	Securitized Product
2009	Financial Crisis Lehman Brothers	Financial Crisis ABS	Non-standard-Measure	Financial Crisis Rescue	Asset Purchase Recovery	Credit Bubble Financial Crisis
2010	Recovery Reform	Recovery Recession	ESRB/ FSB Deficits	Microprudential Macprudential	Recovery/ QE VAT/ TAX	Deflation
2011	Sovereign Debt Basel III	Dodd-Franc Act Recovery	Sovereign Debt EFSF	Debt Crisis Basel III	Commodity Prices Basel III	Asset Purchase ETFs / REITs
2012	Europe Deleveraging	Recovery (has been) Labor Market	OMT/ ESM/ SSM Fragmentation	Banking Union Taxpayer	Investment-Banking	European Debt Deleveraging
2013	Real Economy Price Stability	(At least as long as) Unemployment	SRM/ SSM/ OMT	SRM / SSM Banking Union	Prudential-Regulation	Price Stability QE



Hot topic detection in central bankers' speeches

Jihye Park ^a, Hye Jin Lee ^a, Sungzoon Cho ^{a,b,*}

^a Department of Industrial Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea

^b Institute for Industrial Systems Innovation, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Hot topic detection
Central bankers' speech
Early warning system
Keyword extraction
Emerging term

ABSTRACT

Analysis of central bankers' statements is essential to understand current economic conditions and predict future market trends. The most popular method used for this purpose – qualitative monitoring and evaluation (QME) – is based on the input of domain experts, which is expensive and subjective by nature. Several researchers have attempted to use text mining techniques as alternatives. However, they have primarily focused on aligning textual information with macroeconomic indicators, rather than directly extracting keywords. The primary aim of this study is to identify the possibility of automatically detecting potential financial risk factors by applying text mining techniques to central bankers' speeches. We propose a text mining framework to extract risk factors *ex ante* by detecting hot topics in speeches made by chairs of the Federal Reserve System. In the framework, we use a simple and effective unsupervised keyword-scoring method,¹ which treats bigrams as keywords and incorporates the temporal importance of keywords by estimating the "emergence" of a term at a certain time period relative to previous time periods. In-depth analysis through extensive experiments was conducted to compare risk factor detection performance using eight existing methods including statistical approaches and recent pretrained language model-based approaches. Experimental results demonstrate that our proposed method adopting a statistical approach effectively captures potential risk factors in central

Text Analytics 관련 연구분야

• Industry-related Text analytics

FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

Dogu Tan Araci
dogu.araci@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Financial sentiment analysis is a challenging task due to the specialized language and lack of labeled data in that domain. General-purpose models are not effective enough because of specialized language used in financial context. We hypothesize that pre-trained language models can help with this problem because they require fewer labeled examples and they can be further trained on domain-specific corpora. We introduce FinBERT, a language model based on BERT, to tackle NLP tasks in financial domain. Our results show improvement in every measured metric on current state-of-the-art results for two financial sentiment analysis datasets. We find that even with a smaller training set and fine-tuning only a part of the model, FinBERT outperforms state-of-the-art machine learning methods.

NLP transfer learning methods look like a promising solution to both of the challenges mentioned above, and are the focus of this thesis. The core idea behind these models is that by training language models on very large corpora and then initializing down-stream models with the weights learned from the language modeling task, a much better performance can be achieved. The initialized layers can range from the single word embedding layer [23] to the whole model [5]. This approach should, in theory, be an answer to the scarcity of labeled data problem. Language models don't require any labels, since the task is predicting the next word. They can learn how to represent the semantic information. That leaves the fine-tuning on labeled data only the task of learning how to use this semantic information to predict the labels.

One particular component of the transfer learning methods is the ability to further pre-train the language models on domain specific

2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)

Stock Price Prediction Using News Sentiment Analysis

Saloni Mohan¹, Sahitya Mullapudi¹, Sudheer Sammeta¹, Parag Vijayvergia¹ and David C. Anastasiu^{1,*}

Abstract—Predicting stock market prices has been a topic of interest among both analysts and researchers for a long time. Stock prices are hard to predict because of their high volatile nature which depends on diverse political and economic factors, change of leadership, investor sentiment, and many other factors. Predicting stock prices based on either historical data or textual information alone has proven to be insufficient.

Existing studies in sentiment analysis have found that there is a strong correlation between the movement of stock prices and the publication of news articles. Several sentiment analysis studies have been attempted at various levels using algorithms such as support vector machines, naive Bayes regression, and deep learning. The accuracy of deep learning algorithms depends upon the amount of training data provided. However, the amount of textual data collected and analyzed during the past studies has been insufficient and thus has resulted in predictions with low

financial news of a company instead of only considering the past stock prices can lead to better prediction results.

II. RELATED WORKS

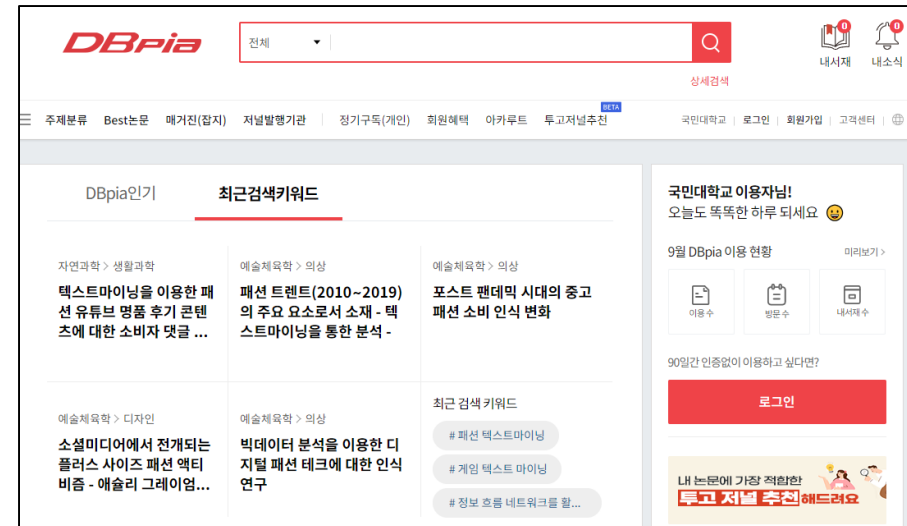
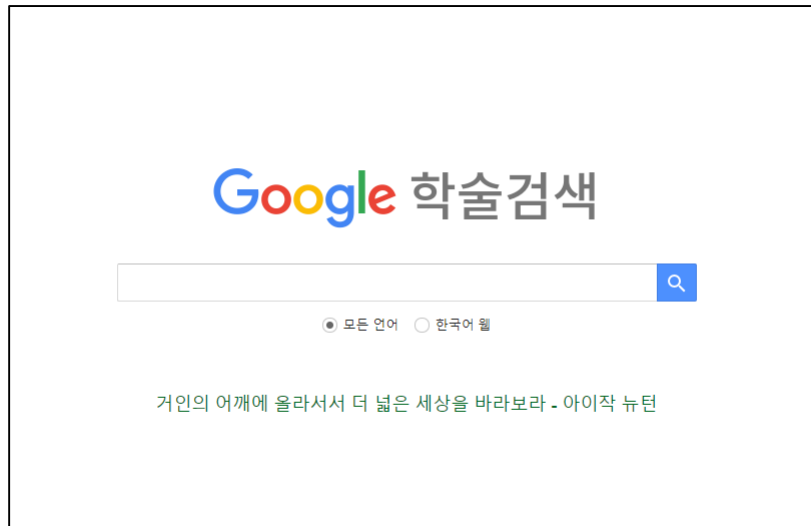
Most of the stock prediction approaches have been built on technical and fundamental analyses of stocks. In recent studies, it has been evident that there is a strong correlation between news articles related to a company and its stock price movements.

Alostad and Davulcu [1] used hourly stock prices of 30 stocks and online stock news articles from the NASDAQ website. They collected tweets related to those 30 stocks for a period of six months. Li et al. [2] collected five years of Hong Kong stock exchange data. They gathered financial

3. How to find related papers?

How to find research paper?

- 저는 논문을 찾을 때 크게 3개의 webpage를 활용합니다.
 - Google
 - research topic에 대한 keyword를 탐색
 - Google scholar
 - 논문 탐색(English)
 - DBpia
 - 논문 탐색(Korean)



How to find research paper?

- 저는 논문을 찾을 때 크게 3개의 webpage를 활용합니다.
 - 집에서 논문을 찾아 읽으려고 할 때, 돈을 지불하라는 경우가 존재
 - 학교에서 논문을 찾으면 돈을 지불하지 않아도 됨
 - 학교 차원에서 구독을 하기 때문에 별도 금액 지불 불필요
 - 즉, 굳이 무료 논문 (arXiv, Open Access)만 찾아 읽을 필요가 없음



How to find research paper?

- 이 세상에는 무수히 많은 논문들이 있습니다.
- 어떤 논문을 찾아서 읽어야 할까요?

How to find research paper?

- 일단 하고 싶은 주제를 생각합니다.
- 그리고, 이를 학계에서 어떤 식으로 부르는지 알아야 합니다.



How to find research paper?

- 이제 앞의 keyword로 Review Paper(=Survey Paper)를 찾아봅니다.
 - 해당 주제에 대한 연구 동향
 - Paper list 및 대략적인 논문 설명

연구 세부 분류

Google 학술검색

fake news detection survey

🔍

📖 학술자료

검색결과 약 8,590개 (0.09초)

모든 날짜

2023년부터

2022년부터

2019년부터

기간 설정...

관련도별 정렬

날짜별 정렬

모든 언어

한국어 웹

모든 유형

검토 자료

☐ 특허 포함

☒ 서지정보 포함

☒ 알림 만들기

[PDF] Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task

A Bondielli, P Dell'Oglio, A Lenci, F Marcelloni... - Proceedings of the ..., 2023 - ceur-ws.org

... Fake-DetectIVE shared task for the EVALITA 2023 campaign. The task was aimed at exploring multimodality within the realm of fake news ... of multimodal fake news detection systems. In ...

☆ 저장 99 인용 5회 인용 🔗

Fake News Detection through Graph-based Neural Networks: A Survey

S Gong, RQ Sinnott, J Qi, C Paris - arXiv preprint arXiv:2307.12639, 2023 - arxiv.org

... to fake news detection and ... Fake news detection is a process used to detect fake news items, eg, Twitter posts. As noted, in this survey we explore graph-based fake news detection ...

☆ 저장 99 인용 1회 인용 전체 4개의 버전 🔗

[HTML] Content Based Fake News Detection with machine and deep learning: a systematic review

N Capuano, G Fenza, V Loia, F D Nota - Neurocomputing, 2023 - Elsevier

... fake news has on our society it is critical to provide an indication to researchers to improve the performance of Automatic Fake News Detectors... helping build better fake news detectors. ...

☆ 저장 99 인용 9회 인용 관련 학술자료 전체 2개의 버전 Web of Science: 4 🔗

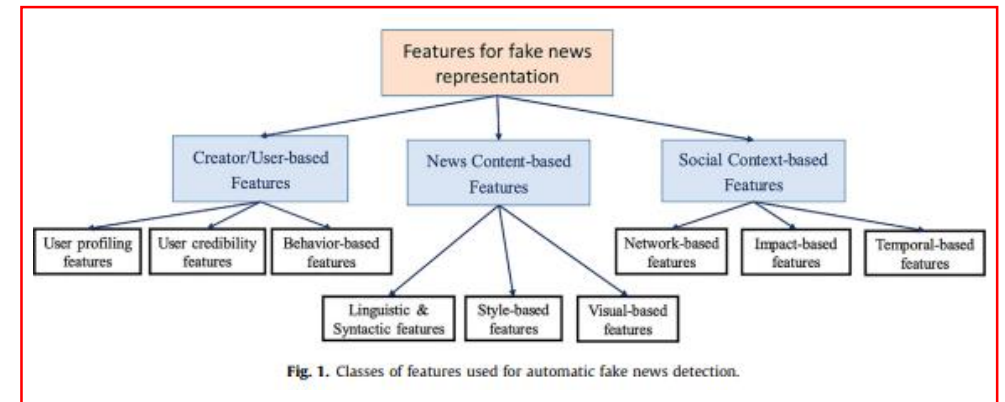


Table 9
pre-trained Deep learning algorithms.

Identifier	Description	Studies
BERT	BERT is a Word Embedding model that, in comparison with other word embedding models (such as Word2Vec), is able to distinguish the meaning of the same or similar words being used in different contexts. As with other embedding models it must be used in conjunction with a classifier. In literature, the use of the term BERT alone is associated with a model made of BERT followed by a dense neural network. Robustly optimized BERT approach is a retraining of BERT with improved training methodology. RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs.	[3,47,50,52]
RoBERTa	It differs from BERT because instead of masking words, it uses a small BERT-like network as a generator to replace some words with its predictions. Then, the main discriminator network is used to determine which of the words have been replaced.	[47]
DistilBERT	It is a character-based model using character convolutions and can handle out-of-vocabulary words for this reason. However, the learned representations are words. Both ELMo and BERT can generate different word embeddings for a word that captures the context of a word. However, ELMo uses LSTM internally while BERT uses transformers.	[47,52]
ELECTRA	It is a lighter version of BERT	[47]
ELMo	It is a large bidirectional transformer that uses improved training methodology, larger data, and more computational power to achieve better than BERT prediction metrics	[47]
AI-BERT		[50]
XLNet		[50]

Paper
List

(Source: Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content Based Fake News Detection with machine and deep learning: a systematic review. Neurocomputing.)

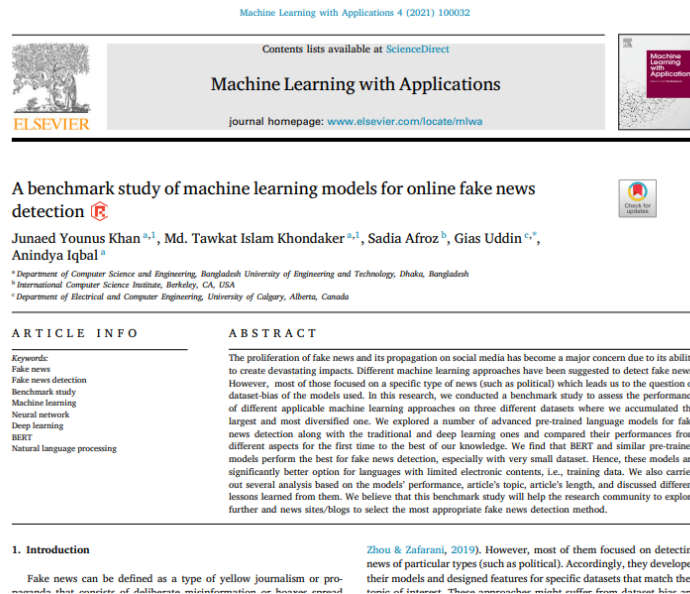
How to find research paper?

- 관심있는 몇가지 논문을 찾아봅니다.

- 먼저 이 논문이 여러분이 생각한 것과 관계 있는지 알아야합니다.

- Introduction, Conclusion을 먼저 잘 읽어보고 내가 생각했던 바와 일치하는지 보세요.

- 생각했던 것과 다른 논문들도 많습니다.



- [42] S. Vijayaraghavan, Y. Wang, Z. Guo, J. Voong, W. Xu, A. Nasser, J. Cai, L. Li, K. Vuong, E. Wadhwa, Fake news detection with different models, arXiv preprint arXiv:2003.04978 (2020).
- [43] A. Agarwal, A. Dixit, Fake news detection: an ensemble learning approach, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020, pp. 1178–1183.
- [44] X. Zhou, A. Jain, V.V. Phoha, R. Zafarani, Fake news early detection: A theory-driven model, Digital Threats: Res. Practice 1 (2020) 1–25.
- [45] Z. Khanam, B. Alwasel, H. Sirafi, M. Rashid, Fake news detection using machine learning approaches, IOP Conference Series: Materials Science and Engineering, vol. 1099, IOP Publishing, 2021, p. 012040.
- [46] A. Choudhary, A. Arora, Linguistic feature based learning model for fake news detection and classification, Expert Syst. Appl. 169 (2021).
- [47] J.Y. Khan, M.T.I. Khondaker, S. Afroz, G. Uddin, A. Iqbal, A benchmark study of machine learning models for online fake news detection, Mach. Learn. Appl. 4 (2021).
- [48] T. Reuber, Constraint 2021: Machine learning models for covid-19 fake news detection shared task, arXiv preprint arXiv:2101.03717 (2021).
- [49] P.K. Verma, P. Agrawal, I. Amorim, R. Prodán, Wefake: word embedding over linguistic features for fake news detection, IEEE Trans. Comput. Soc. Syst. 8 (2021) 881–893.
- [50] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).

Knowledge Extraction and Management, Situation and Context Awareness, Semantic Information Retrieval, Service Oriented Architecture, and Ontology Learning. More recently, he has worked in Automating Open Source Intelligence and Big Data Analytics for countering extremism and supporting information disorder awareness. He is currently an Associate Professor in Computer Science at the University of Salerno.

Vincenzo Loia graduated in Computer Science at the University of Salerno, Italy, in 1985 and received his Ph. D. in Computer Science in 1989 at the Université Pierre & Marie Curie Paris VI, France. He is currently a Full Professor in Computer Science at the University of Salerno, where he served as a researcher from 1989 to 2000 and as an associate professor from 2000 to 2004. He is the Co-Editor-in-Chief of Soft Computing and the Editor-in-Chief of Ambient Intelligence and Humanized Computing. He serves as an Editor for 14 other international journals.

Francesco David Nota earned a Master's Degree in Business Innovation and Informatics at the University of Salerno, Italy, in 2020. In 2021 he started his Ph. D. Program in Innovation Sciences for Defence and Security-Digital Transformation and CyberSecurity at CASD (Center for Higher Defence Studies), and he is focusing his research interests on the field of Cognitive Warfare.

Table 9
pre-trained Deep learning algorithms.

Identifier	Description	Studies
BERT	BERT is a Word Embedding model that, in comparison with other word embedding models (such as Word2Vec), is able to distinguish the meaning of the same or similar words being used in different contexts. As with other embedding models it must be used in conjunction with a classifier. In literature, the use of the term BERT alone is associated with a model made of BERT followed by a dense neural network. Robustly optimized BERT approach is a retraining of BERT with improved training methodology. RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs.	[3,47,50,52]
RoBERTa	It learns an approximate version of BERT, retaining 97% performance but using only half the number of parameters	[47]
DistilBERT	It differs from BERT because instead of masking words, it uses a small BERT-like network as a generator to replace some words with its predictions. Then, the main discriminator network is used to determine which of the words have been replaced.	[47,52]
ELECTRA	It is a character-based model using character convolutions and can handle out-of-vocabulary words for this reason. However, the learned representations are words. Both ELMo and BERT can generate different word embeddings for a word that captures the context of a word. However, ELMo uses LSTM internally while BERT uses transformers.	[47]
ELMo	It is a lighter version of BERT	[50]
ALBERT	It is a large bidirectional transformer that uses improved training methodology, larger data, and more computational power to achieve better than BERT prediction metrics	[50]
XLNet		[50]

(Source: Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content Based Fake News Detection with machine and deep learning: a systematic review. Neurocomputing. Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.)

How to find research paper?

- 관심있는 몇가지 논문을 찾아봅니다.
 - 이제 중요한 부분을 잘 읽어보세요.
 - 제안 방법론, 활용 데이터, 평가 방법 등

3.3.3. Advanced language models

Here, we first discuss the advanced language models that we used in this study and then describe their experimental setup.

- (1) **BERT**: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model which was designed to learn contextual word representations of unlabeled texts (Devlin et al., 2018). Among the two versions of BERT (i.e., BERT-Base and BERT-Large) proposed originally, we used BERT-Base for this study considering the huge time and memory requirements of the BERT-Large model. The BERT-Base model has 12 layers (transformer blocks) with 12 attention heads and 110 million parameters.
- (2) **RoBERTa**: RoBERTa (Robustly optimized BERT approach), originally suggested in (Liu et al., 2019), is the second pre-trained model that we experimented. It achieves better performance than original BERT models by using larger mini-batch sizes to train the model for a longer time over more data. It also removes the NSP loss in BERT and trains on longer sequences. Moreover, it dynamically changes the masking pattern applied to the training data.
- (3) **DistilBERT**: DistilBERT (Sanh et al., 2019) is a smaller, faster, cheaper, and lighter version of original BERT which has 40% fewer parameters than the BERT-Base model. Though original BERT models perform better, DistilBERT is more appropriate for production-level usage due to its low resource requirements. Considering potential users of non-profit blogs and online media, we think low-resource models have a good appeal. Hence, this is worth investigating.
- (4) **ELECTRA**: ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020) is a transformer model for self-supervised language representation learning. This model pre-trained with the use of another (small) masked language model. First, a language model takes an input text and randomly masked the text with generated input token. Then, ELECTRA models are trained to distinguish "real" input tokens vs "fake" input tokens generated by the former language model. At small scale, ELECTRA can achieve strong results even when trained on a single GPU.

3.4. Evaluation metrics

We created a standard training and test set for each of the three datasets by splitting it in an 80:20 ratio so that different models can be evaluated on the same ground. For the first two datasets (i.e., Liar, Fake or Real), we did the split randomly as they only contain one type of news. On the other hand, as the Combined Corpus covers a wide variety of topics, we took 80% (20%) data from each topic and include them in train (test) set to maintain a balanced distribution of every topic in training and test data.

We report the performance of each model in terms of accuracy, precision, recall, and F1-score. For precision, recall, and F1-score, we considered the macro-average of both class.

In our experiment, we considered real news as 'positive class', and fake news as 'negative class'. Hence, True Positive (TP) means the news is actually real, and also predicted as real while False Positive (FP) indicates that the news is actually false, but predicted as real. True Negative (TN) and False Negative (FN) imply accordingly. Accuracy is the number of correctly predicted instances out of all instances.

$$\text{Accuracy (A)} = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

Precision is the ratio between the number of correctly predicted instances and all the predicted instances for a given class. For real and fake classes, we presented this metric as P(R) and P(F) respectively. Hence, the macro-average precision, P will be the average of P(R) and P(F).

$$P(R) = \frac{TP}{TP + FP}, P(F) = \frac{TN}{TN + FN}, P = \frac{P(R) + P(F)}{2} \quad (2)$$

Recall represents the ratio of the number of correctly predicted instances and all instances belonging to a given class. For real and fake classes, we presented this metric as R(R) and R(F) respectively. Hence, the macro-average recall, R will be the average of R(R) and R(F).

$$R(R) = \frac{TP}{TP + FN}, R(F) = \frac{TN}{TN + FP}, R = \frac{R(R) + R(F)}{2} \quad (3)$$

F1-score is the harmonic mean of the precision and recall.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (4)$$



- Knowledge Extraction and Management, Situation and Context Awareness, Semantic Information Retrieval, Service Oriented Architecture, and Ontology Learning. More recently, he has worked in Automating Open Source Intelligence and Big Data Analytics for countering extremism and supporting information disorder awareness. He is currently an Associate Professor in Computer Science at the University of Salerno.
- Vincenzo Loia graduated in Computer Science at the University of Salerno, Italy, in 1985 and received his Ph. D. in Computer Science in 1989 at the Université Pierre & Marie Curie Paris VI, France. He is currently a Full Professor in Computer Science at the University of Salerno, where he served as a researcher from 1989 to 2000 and as an associate professor from 2000 to 2004. He is the Co-Editor-in-Chief of Soft Computing and the Editor-in-Chief of Ambient Intelligence and Humanized Computing. He serves as an Editor for 14 other international journals.
- Francesco David Nota earned a Master's Degree in Business Innovation and Informatics at the University of Salerno, Italy, in 2020. In 2021 he started his Ph. D. Program in Innovation Sciences for Defence and Security-Digital Transformation and CyberSecurity at CASD (Center for Higher Defence Studies), and he is focusing his research interests on the field of Cognitive Warfare.
- [42] S. Vijayaraghavan, Y. Wang, Z. Guo, J. Voong, W. Xu, A. Nasser, J. Cai, L. Li, K. Vuong, E. Wadhwa, Fake news detection with different models, arXiv preprint arXiv:2003.04978 (2020).
- [43] A. Agarwal, A. Dixit, Fake news detection: an ensemble learning approach, in: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020, pp. 1178–1183.
- [44] X. Zhou, A. Jain, V.V. Phoha, R. Zafarani, Fake news early detection: A theory-driven model, Digital Threats: Res. Practice 1 (2020) 1–25.
- [45] Z. Khanam, B. Alwasel, H. Sirafi, M. Rashid, Fake news detection using machine learning approaches, IOP Conference Series: Materials Science and Engineering, vol. 1099, IOP Publishing, 2021, p. 012040.
- [46] A. Choudhary, A. Arora, Linguistic feature based learning model for fake news detection and classification, Expert Syst. Appl. 169 (2021).
- [47] J.Y. Khan, M.T.I. Khondaker, S. Afroz, G. Uddin, A. Iqbal, A benchmark study of machine learning models for online fake news detection, Mach. Learn. Appl. 4 (2021).
- [48] T. Reuber, Constraint 2021: Machine learning models for covid-19 fake news detection shared task, arXiv preprint arXiv:2101.03717 (2021).
- [49] P.K. Verma, P. Agrawal, I. Amorim, R. Prodán, Welfake: word embedding over linguistic features for fake news detection, IEEE Trans. Comput. Soc. Syst. 8 (2021) 881–893.
- [50] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).

Table 9
pre-trained Deep learning algorithms.

Identifier	Description	Studies
BERT	BERT is a Word Embedding model that, in comparison with other word embedding models (such as Word2Vec), is able to distinguish the meaning of the same or similar words being used in different contexts. As with other embedding models it must be used in conjunction with a classifier. In literature, the use of the term BERT alone is associated with a model made of BERT followed by a dense neural network. Robustly optimized BERT approach is a retraining of BERT with improved training methodology. RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs.	[3,47,50,52]
RoBERTa	It learns an approximate version of BERT, retaining 97% performance but using only half the number of parameters	[47]
DistilBERT	It differs from BERT because instead of masking words, it uses a small BERT-like network as a generator to replace some words with its predictions. Then, the main discriminator network is used to determine which of the words have been replaced.	[47,52]
ELECTRA	It is a character-based model using character convolutions and can handle out-of-vocabulary words for this reason. However, the learned representations are words. Both ELMo and BERT can generate different word embeddings for a word that captures the context of a word. However, ELMo uses LSTM internally while BERT uses transformers.	[47]
ELMo	It is a lighter version of BERT	[50]
AlBERT	It is a large bidirectional transformer that uses improved training methodology, larger data, and more computational power to achieve better than BERT prediction metrics	[50]
XLNet		

(Source: Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content Based Fake News Detection with machine and deep learning: a systematic review. Neurocomputing. Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032.)

How to find research paper?

- 한국어로 된 논문도 마찬가지로 있습니다.

- 단, 한국어로 된 review paper는 상대적으로 많지 않습니다.
- keyword로 찾아보세요

J Intell Inform Syst 2023 June; 29(2): 267-283
http://dx.doi.org/10.13088/jis.2023.29.2.267

품사별 출현 빈도를 활용한 코로나19 관련 한국어 가짜뉴스 탐지*

김지혁
국민대학교 비즈니스IT전문대학원
(kjh9654@kookmin.ac.kr)

안현철
국민대학교 비즈니스IT전문대학원
(hcahn@kookmin.ac.kr)

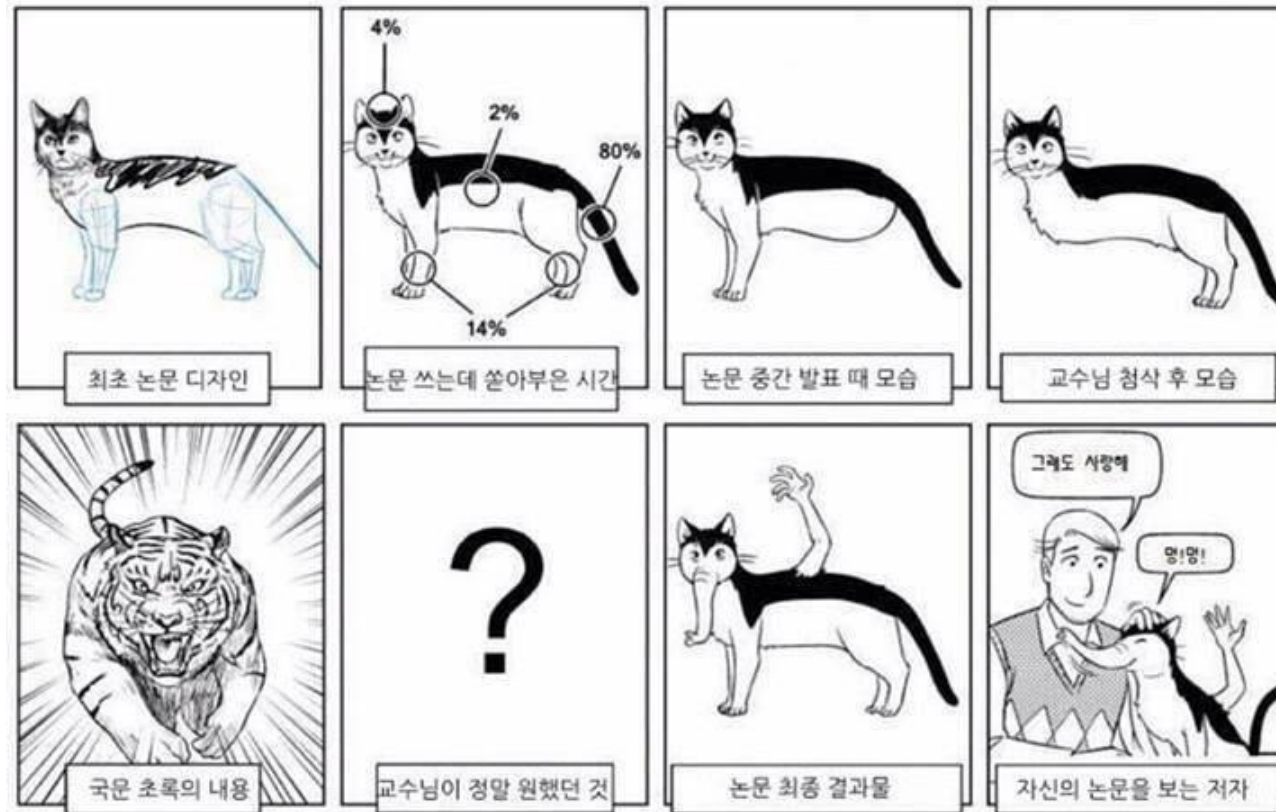
2019년 12월부터 현재까지 지속되고 있는 코로나19 팬데믹으로 인해 대중들은 감염병 대응을 위한 정보를 필요로 하게 되었다. 하지만 소셜미디어에서 유포되는 코로나19 관련 가짜뉴스로 인해 대중들의 건강이 심각하게 위협받고 있다. 특히 코로나19와 관련된 가짜뉴스가 유사한 내용으로 대량 유포될 경우 사실인지 거짓인지 진위를 가리기 위한 검증에 소요되는 시간이 길어지게 되어 우리 사회의 전반에 심각한 위협이 될 수 있다. 이에 학계에서는 신속하게 코로나19 관련 가짜뉴스를 탐지할 수 있는 지능형 모델에 대한 연구를 활발하게 수행해 오고 있으나, 대부분의 기존 연구에 사용된 데이터는 영문으로 구성되어 있어 한국어 가짜뉴스 탐지에 대한 연구는 매우 드문 실정이다. 이에 본 연구에서는 소셜 미디어 상에서 유포되는 한국어로 작성된 코로나19 관련 가짜뉴스 데이터를 직접 수집하고, 이를 기반으로 한 지능형 가짜뉴스 탐지 모델을 제안한다. 본 연구의 제안모델은 언어학적 특성 중 하나인 품사별 빈도 정보를 추가적으로 활용하여, 기존 연구에서 주로 사용되어 온 문서 임베딩 기반인 Doc2Vec 기반 가짜뉴스 탐지 모델의 예측 성능을 제고하고자 하였다. 실증분석 결과, 제안 모델이 비교 모델에 비해 Recall 및 F1 점수가 높아져 코로나19 관련 한국어 가짜뉴스를 보다 정확하게 판별함을 확인하였다.

주제어 : 코로나19 관련 가짜뉴스, 한국어 가짜뉴스, 소셜 미디어, Doc2Vec, 품사 구분

(Source: 김지혁, 안현철. (2023). 품사별 출현 빈도를 활용한 코로나19 관련 한국어 가짜뉴스 탐지. 지능정보연구, 29(2), 267-283.)

우리 꼭 결과물을 다 만들어 봅시다.

논문의 완성 과정



Sandra and Woo by Oliver Knörzer (writer) and Powree (artist) - www.sandraandwoo.com

(그림 원본: Sandra and Woo by Oliver Knörzer [출처] 논문의 완성 과정[작성자 의학논문 치트키.]

End of the documents
