A Shallow Dive into NLP

이상아
서울대학교 언어학과

2023-1 인지과학제문제
2023. 04. 26.

Computational Linguistics

- A subfield of linguistics and computer science
  - concerned with the interaction of human language and computers
- Includes:
  - the analysis of written texts and spoken discourse
  - the translation of text and speech from one language into another
  - the use of human languages for communication between computers and people
  - the modeling and testing of linguistic theories
- Statistical analysis of written texts and spoken discourse
  - analysis on corpus: relative frequencies or collocation

    of letters, sounds, morphemes, words, …

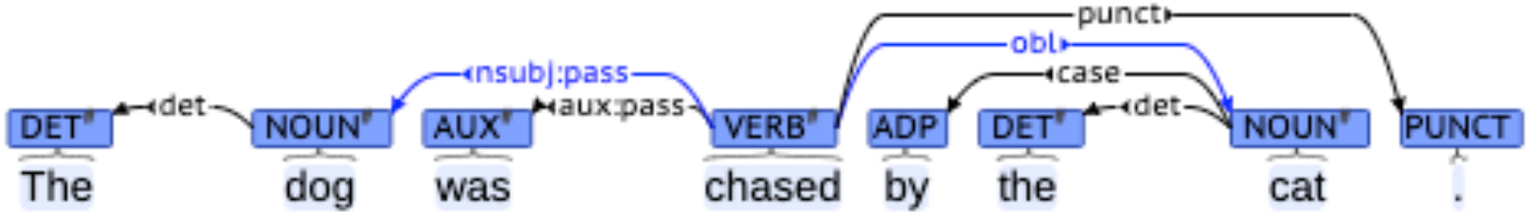
Fromkin et al. (2018). "An Introduction to Language". Cengage Learning Australia.

https://medium.com/dair-ai/building-your-first-wordcloud-with-google-colaboratory-and-python-b3c8c1be5883

Computational Linguistics

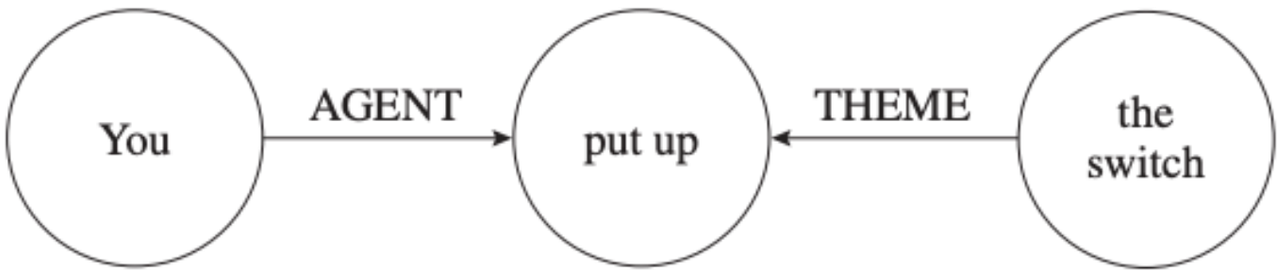
- The interaction between language and computers in all dimensions
  - Computational Phonetics and Phonology

    - Speech Recognition, Speech Synthesis (Text-to-Speech)
  - Computational Morphology: processing of word structures
  - Computational Syntax



https://universaldependencies.org/

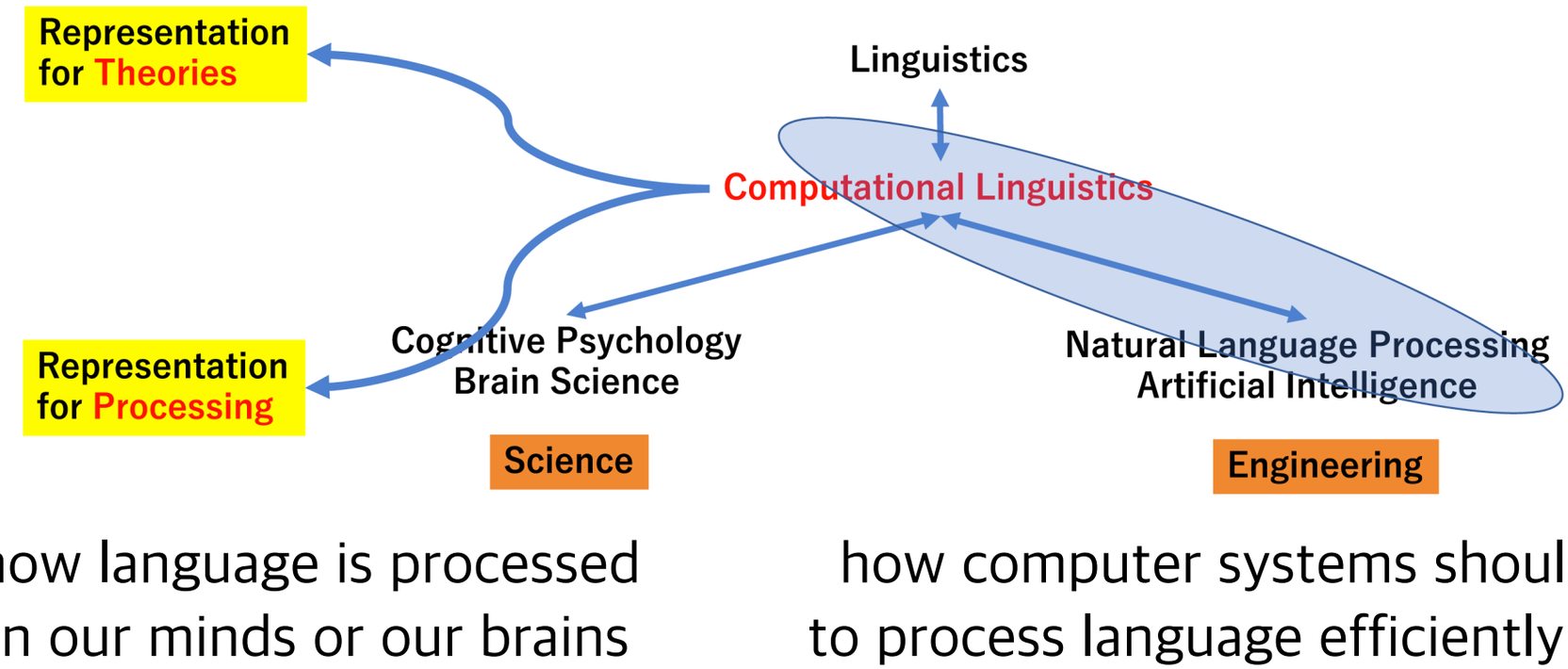
- Computational Semantics: speech understanding and generation



- Computational Pragmatics: sentence disambiguation, coreference resolution, …

Natural Language Processing

- A schematic view of research disciplines: CL and NLP



- NLP may be included either in CL and other fields as their subfield.
  (Any other fields can deal with and utilize language data.)
- And this kind of research can be performed through…

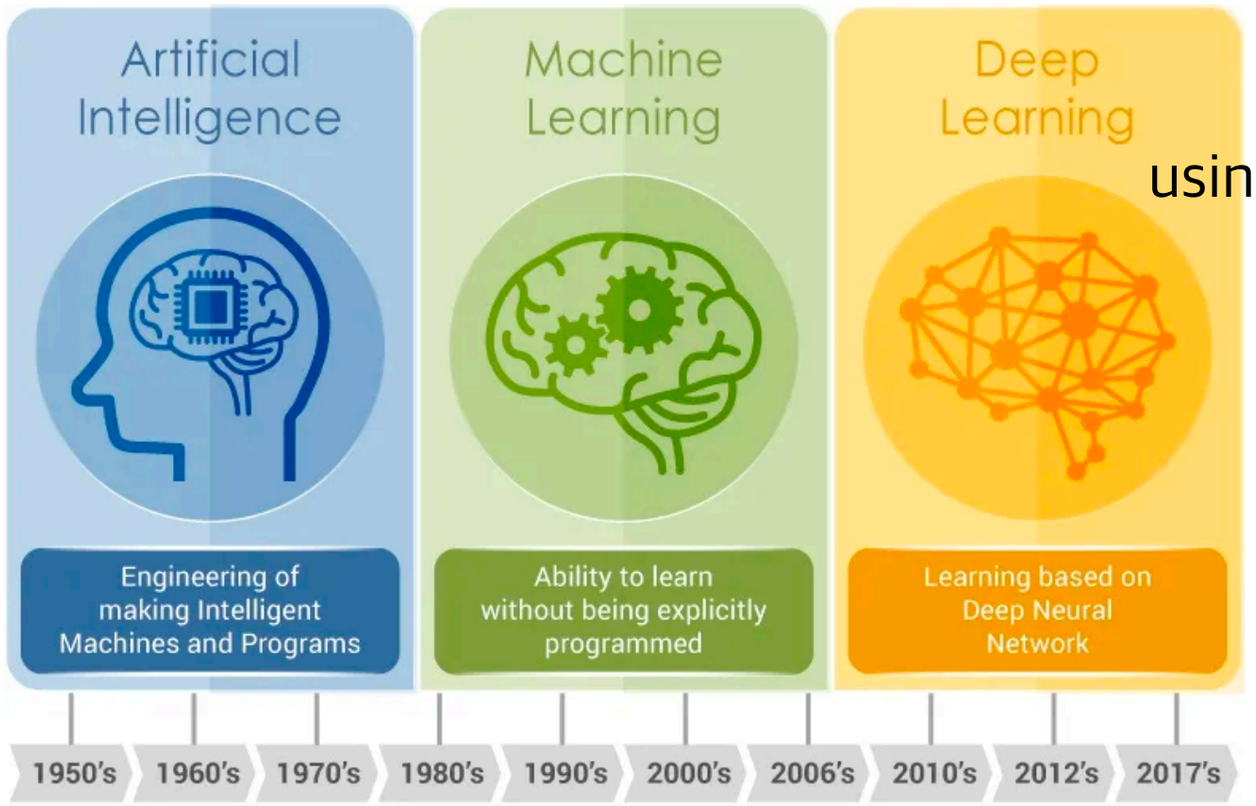
Natural Language Processing

- To deal with data consisting of natural language (text and speech)
- Extracting or understanding, processing information included in language
  - Text Classification: sentiment analysis, topic classification, spam detection, …
    - Aspect-Based Sentiment Analysis: dealing with aspects and opinions
  - Natural Language Inference
  - Named Entity Recognition
  - Relation Extraction
  - Text Summarization
  - Question Answering
  - Generation, Translation
  - Text-to-Speech, Speech-to-Text
  - …

https://medium.com/nlplanet/two-minutes-nlp-33-important-nlp-tasks-explained-31e2caad2b1b

Where are we now?

- Deep neural networks have been widely applied for various tasks in recent years.



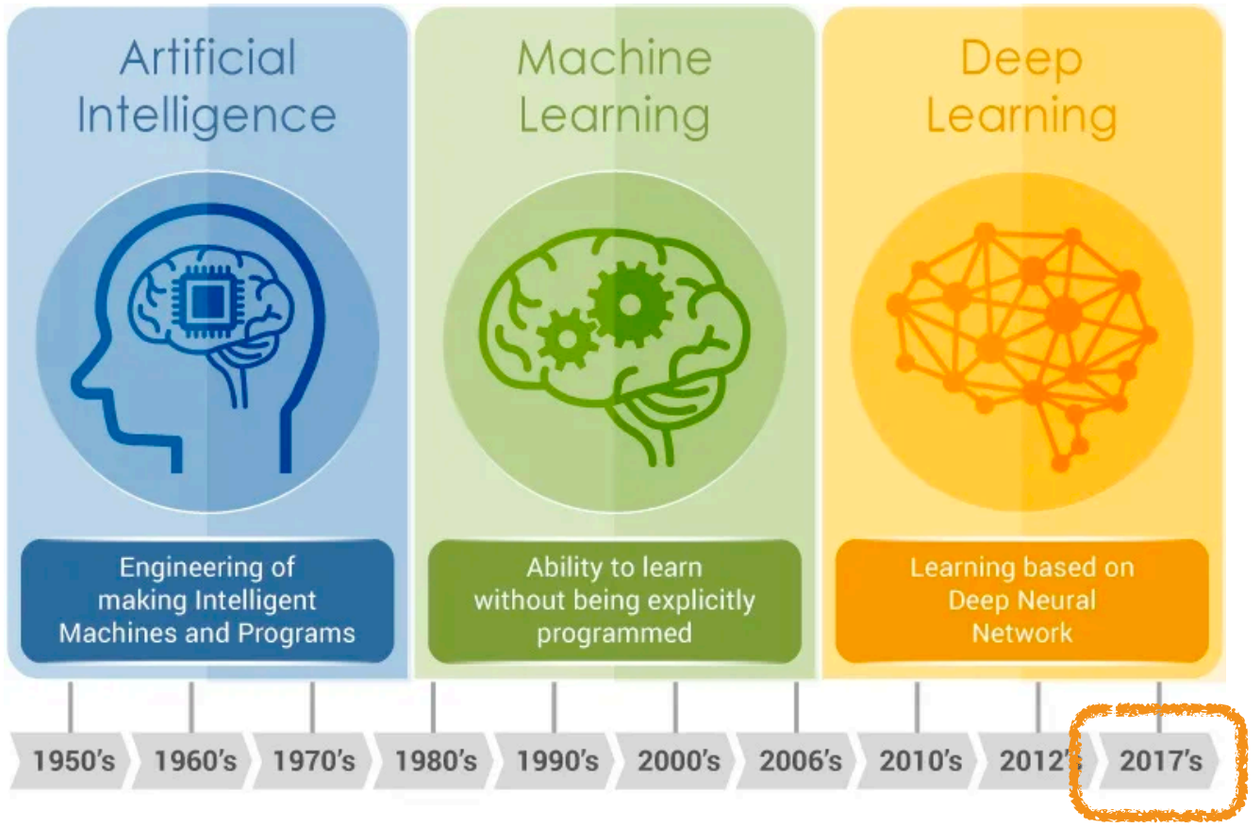
using RNN, CNN, LSTM, …

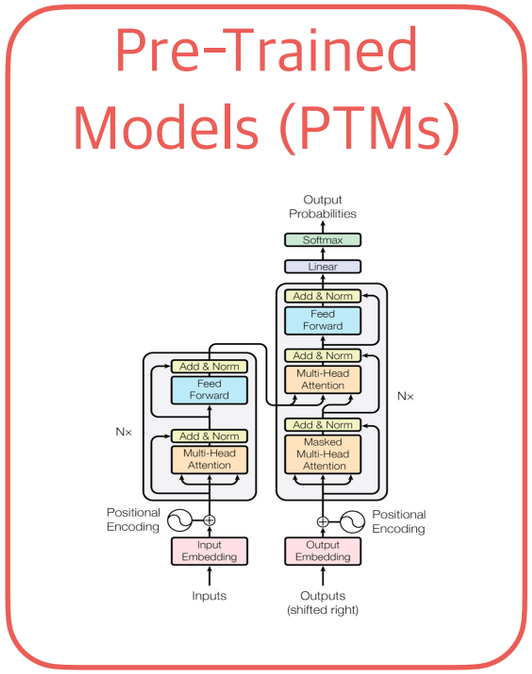
- Different from previous non-neural models (hand-crafted features and statistical methods)
  Neural networks can automatically learn low-dimensional continuous vectors from data as task-specific features (no need for complex feature engineering)

https://medium.com/@mail2princeyadav/machine-learning-vs-deep-learning-b5c5a4fc5c

Where are we now?

- The introduction of Transformers





Attention Mechanism (Google, 2017)

- Very deep neural models for NLP tasks
- Large-scale PTMs with hundreds of millions of parameters e.g. BERT, GPT
  - Can capture polysemous disambiguation, lexical and syntactic structures, factual knowledge, … from the text

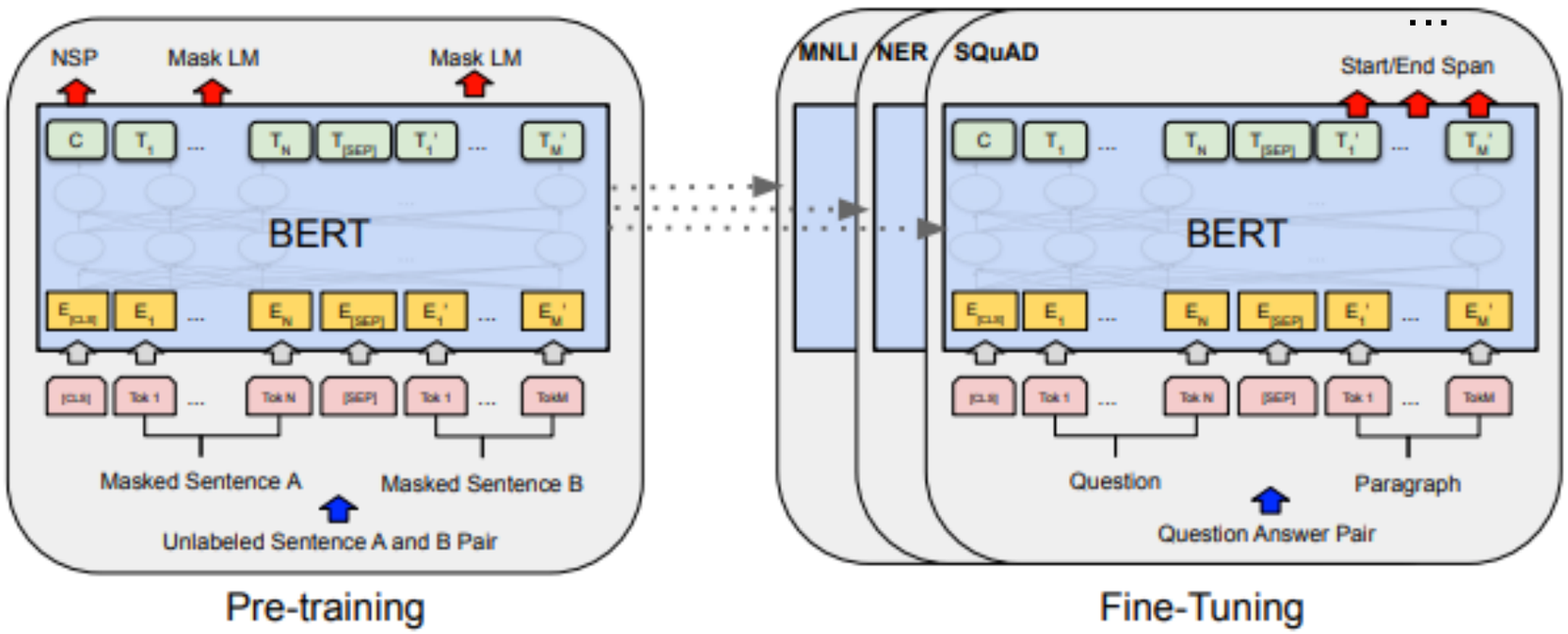
Transfer Learning

- How to train effective deep neural models for specific tasks
  with limited human-annotated data
- People can rely on previously learned knowledge to solve new problems
  and even achieve better results.
- Capturing important knowledge from multiple source tasks
  and then apply the knowledge to a target task
- Source and target tasks may have completely different data domains and task settings,
  yet the knowledge required to handle these task is consistent.

Transfer Learning

- Usually performed with pre-trained language models

tasks:
Natural Language Inference
Named Entity Recognition
Question Answering
…



import and load a given BERT model

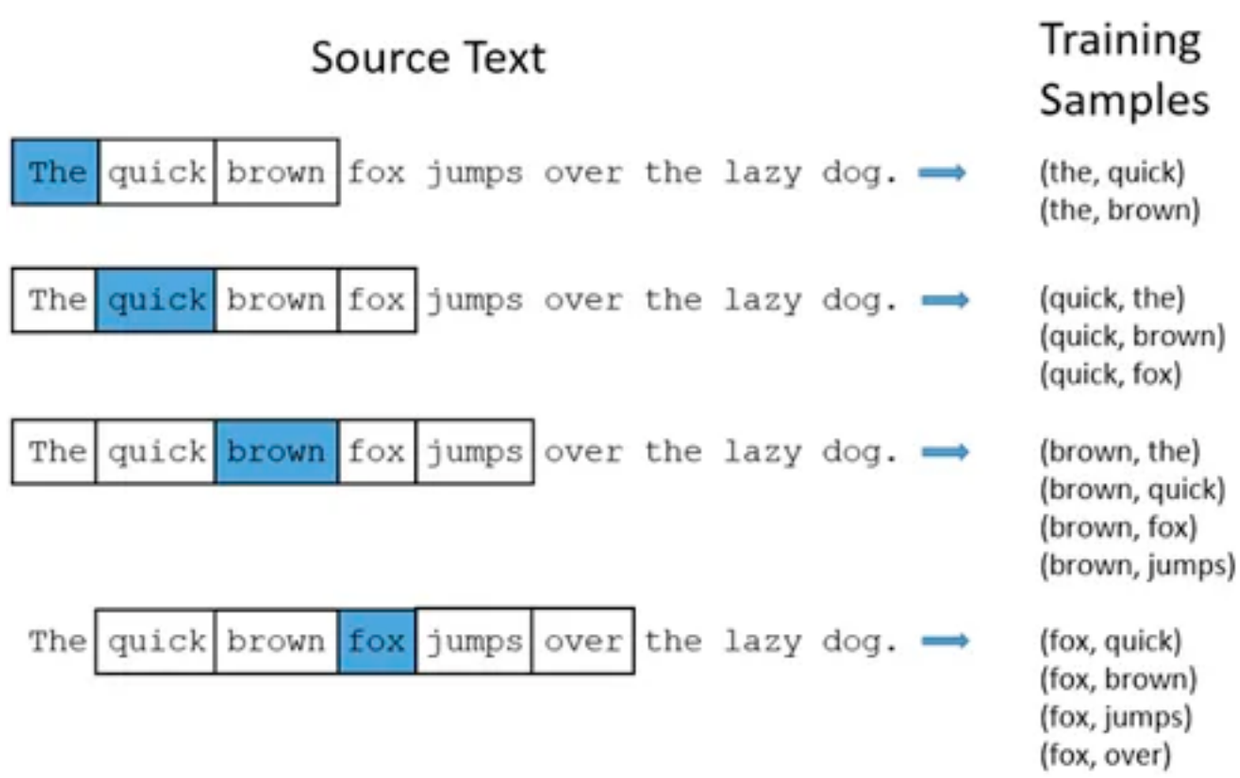
let the model perform specific task

- Such PTMs are trained via self-supervised learning (self-supervised pre-training).

Devlin et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171-4186.

Self-Supervision

- Self-supervised learning: pre-training on large-scale unsupervised data
  - to leverage intrinsic correlations in the text as supervision signals
- Early PTMs: Word Embeddings (e.g. Word2Vec, GloVe)
  - Self-supervised methods to transform words into distributed representations
  - Capturing syntactic and semantic information in the text
  - …by constructing samples semi-automatically from the unsupervised texts



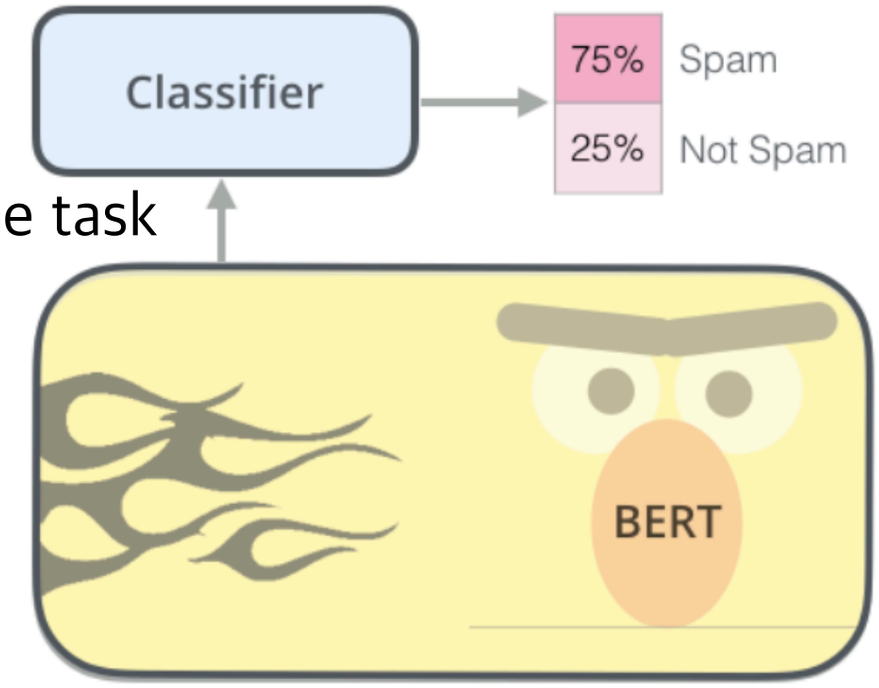
https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1

Self-Supervision

- Transformer-based PTMs: GPT, BERT
  - After pre-training on large-scale textual corpora
  - Both the architecture and parameters of PTMs can serve as a starting point for specific NLP tasks

Add some fine-tuned information about the task

A pre-trained BERT model
(Including well-trained
general linguistic knowledge)

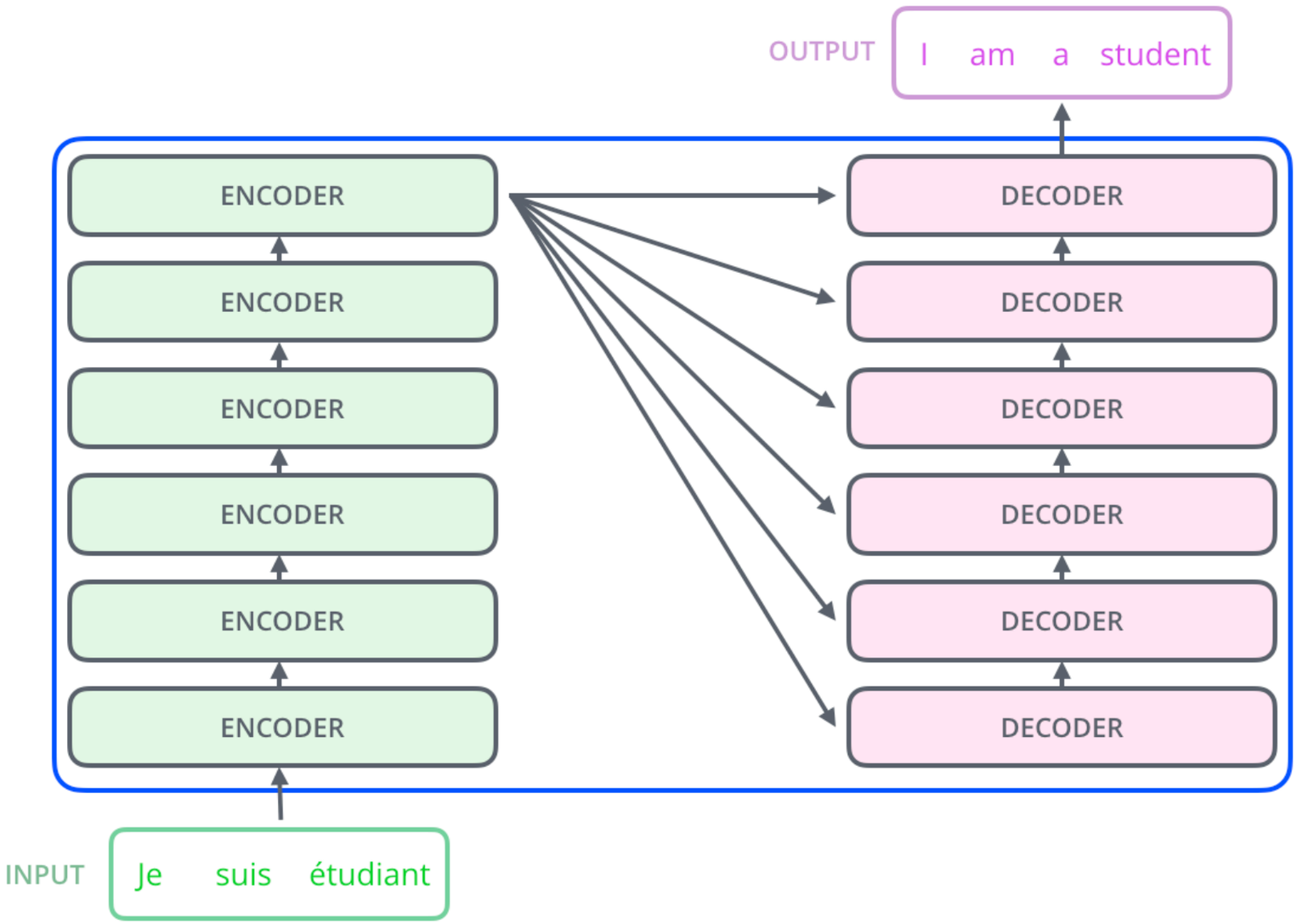


Dataset for a task: spam detection

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Transformer

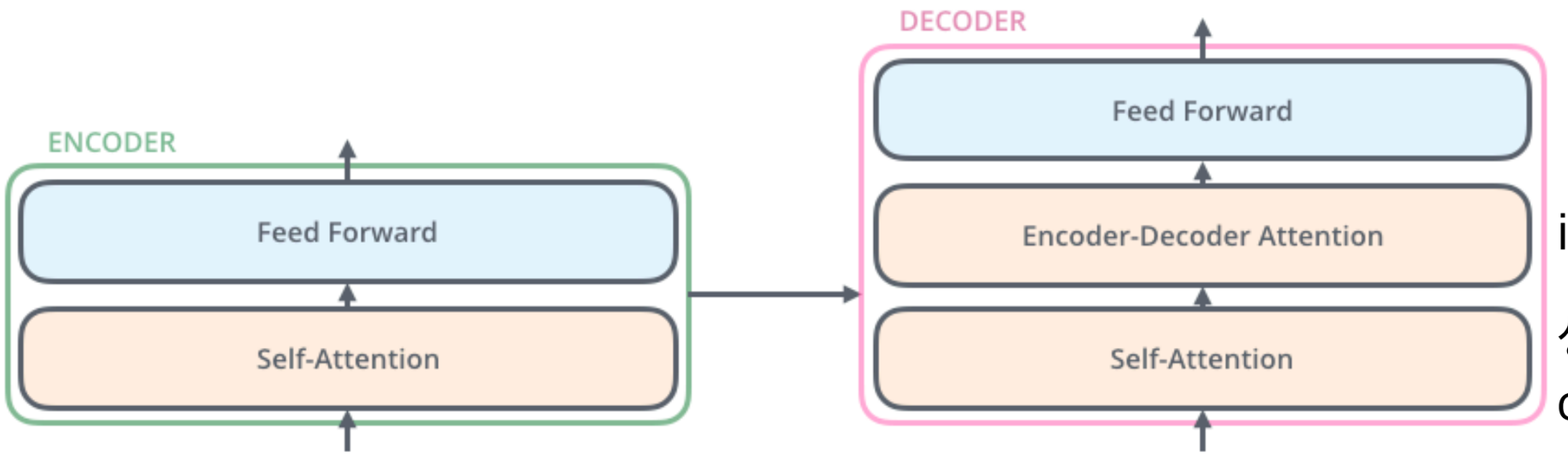
- A non-recurrent sequence-to-sequence (Seq2Seq) architecture consisting of an Encoder and a Decoder, with Self-Attention



- Encoders and decoders are both stacked by several identical blocks.

Transformer

- Encoder
  - a multi-head self-attention layer, a position-wise feed-forward layer
- Decoder
  - a multi-head self-attention layer, a position-wise feed-forward layer, a cross-attention layer (Encoder-Decoder Attention)
  - to consider the output of the Encoder as a context for generation



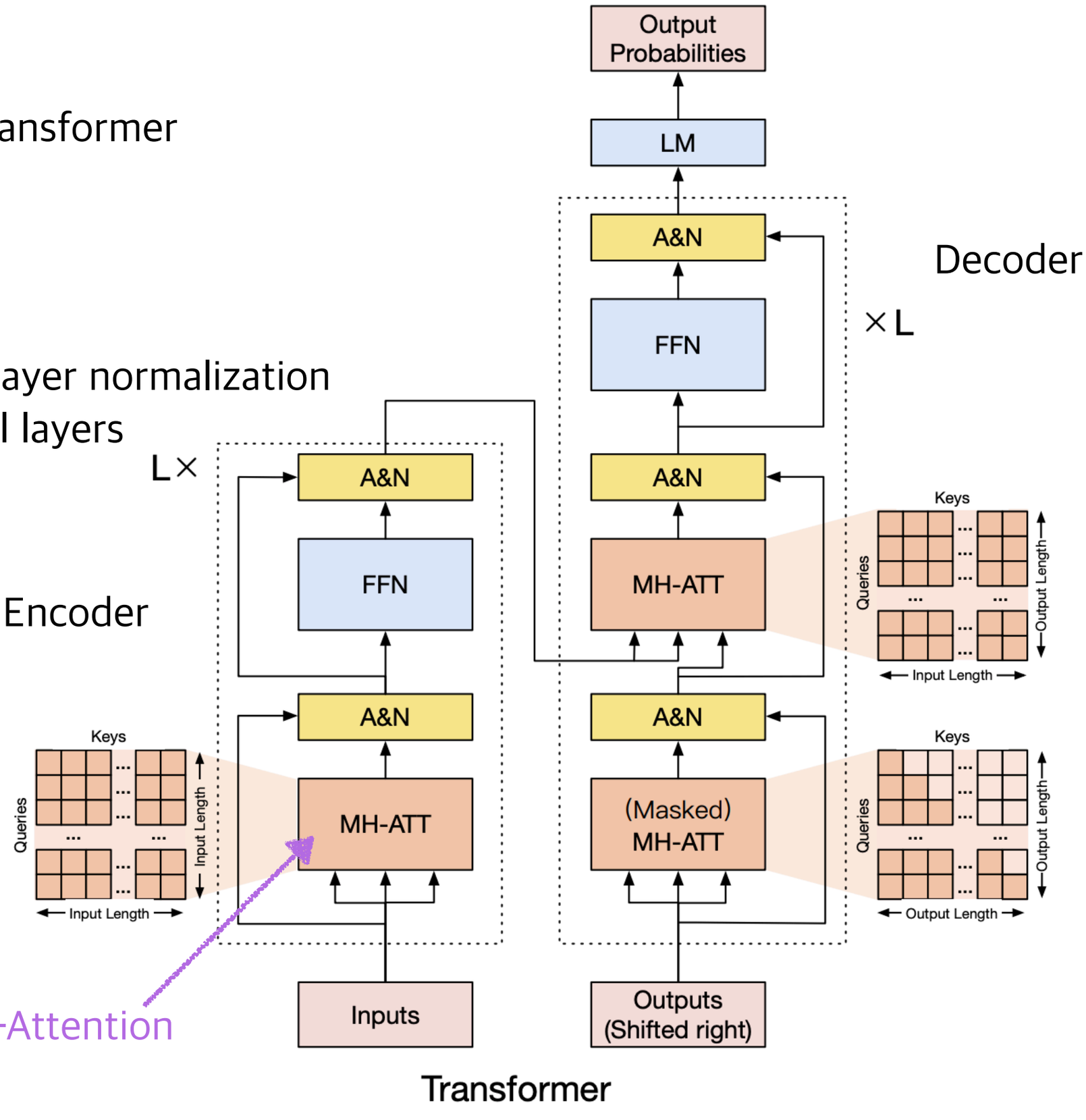
input과의 관계

생성하고 있는 output 내의 관계

Transformer

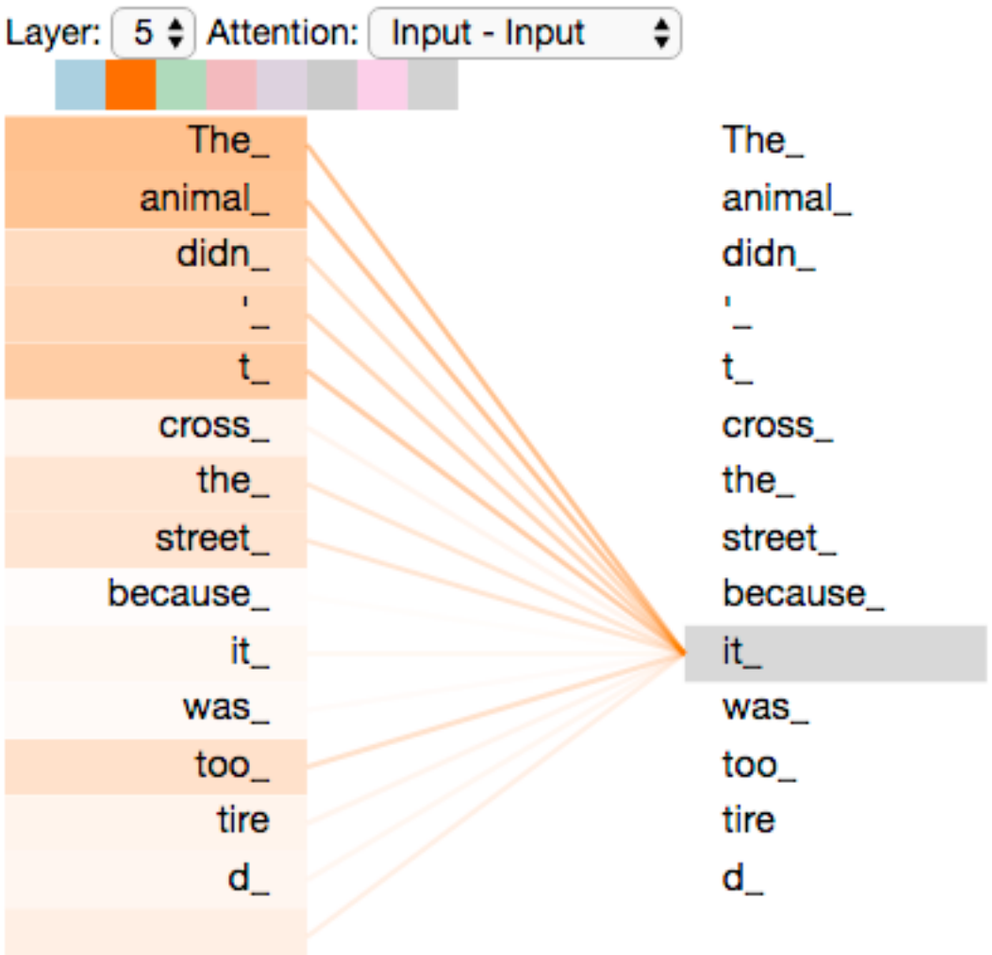
- The architecture of Transformer

Residual connection and layer normalization employed between neural layers



Multi-Head Self Attention

- The tokens calculating the attention are from the same sequence.
  - to learn the relationships between different words in the sentence
  - Usually calculated by the weighted sum of the other elements in the sequence
- Bidirectional design: referring to both previous and next tokens



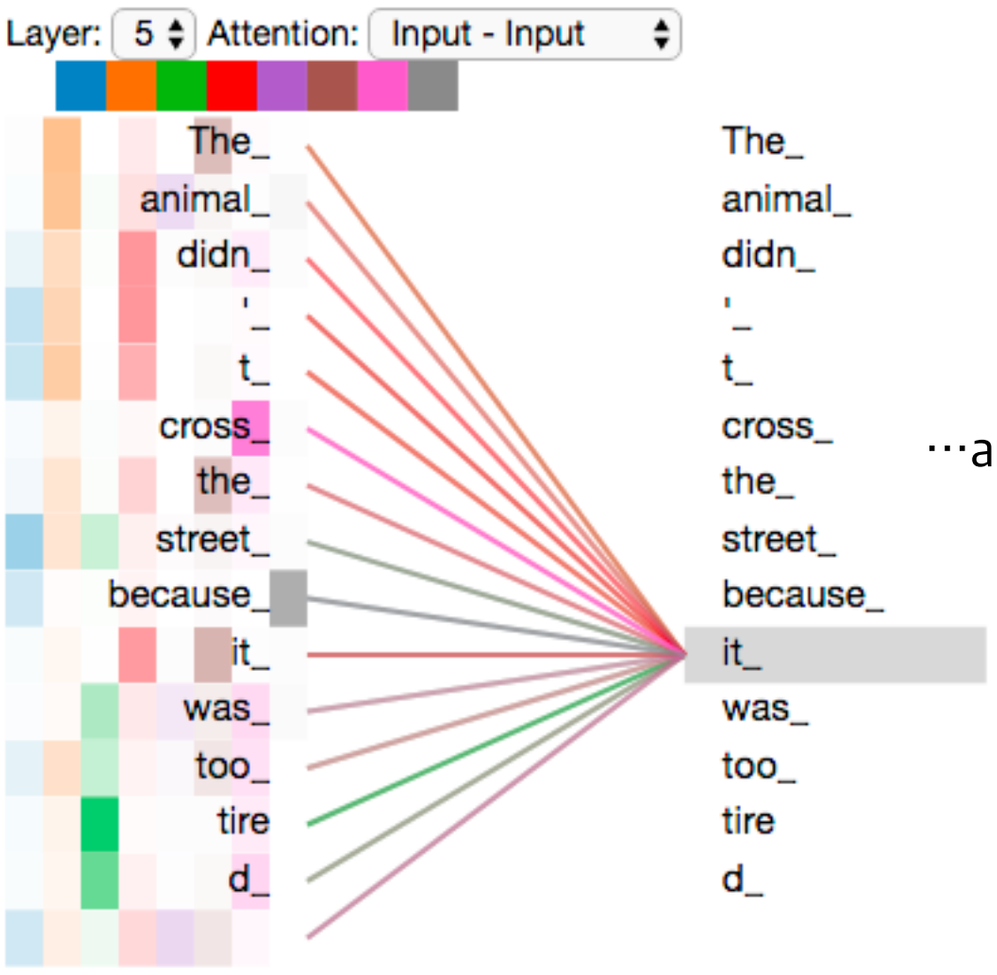
The animal did not cross the road because it was too tired.

The animal did not cross the road because it was too wide.

Multi-Head Self Attention

- Attentions are calculated multiple times in parallel.
  - By each head -> multi-head!
  - It expands the model's ability to focus on different positions.
  - It gives the attention layer multiple "representation subspaces".

…since each matrices are randomly initialized.



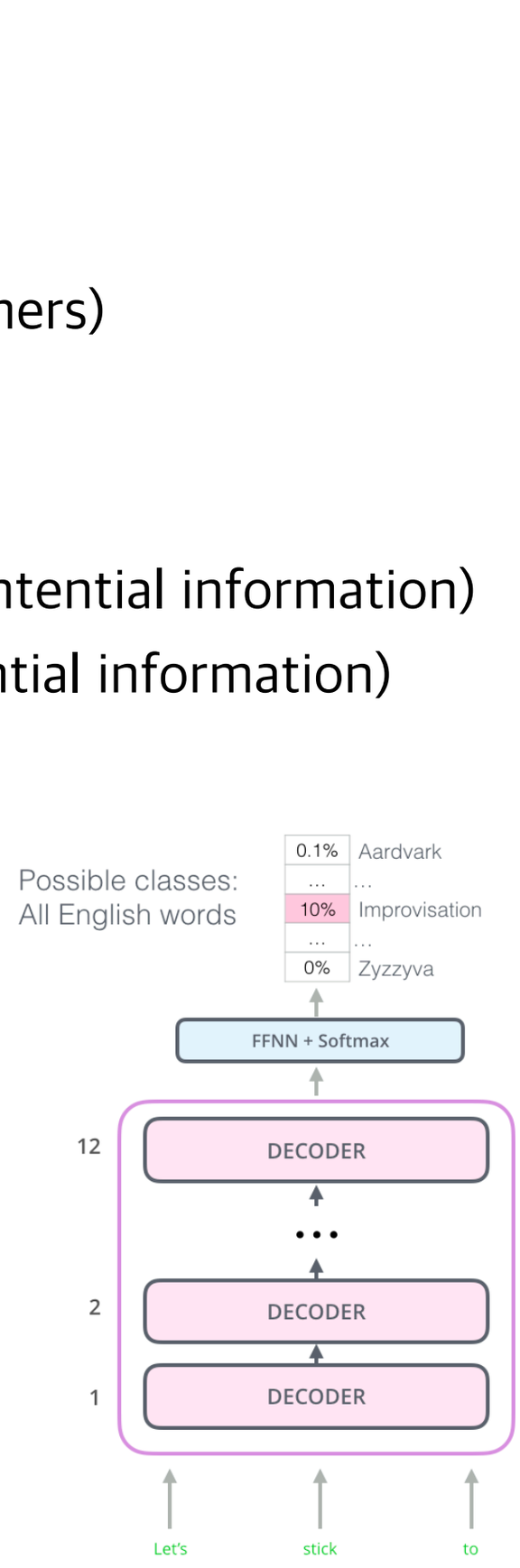
Different degrees of attentions of different heads

…are concatenated to construct an embedding for a sequence.

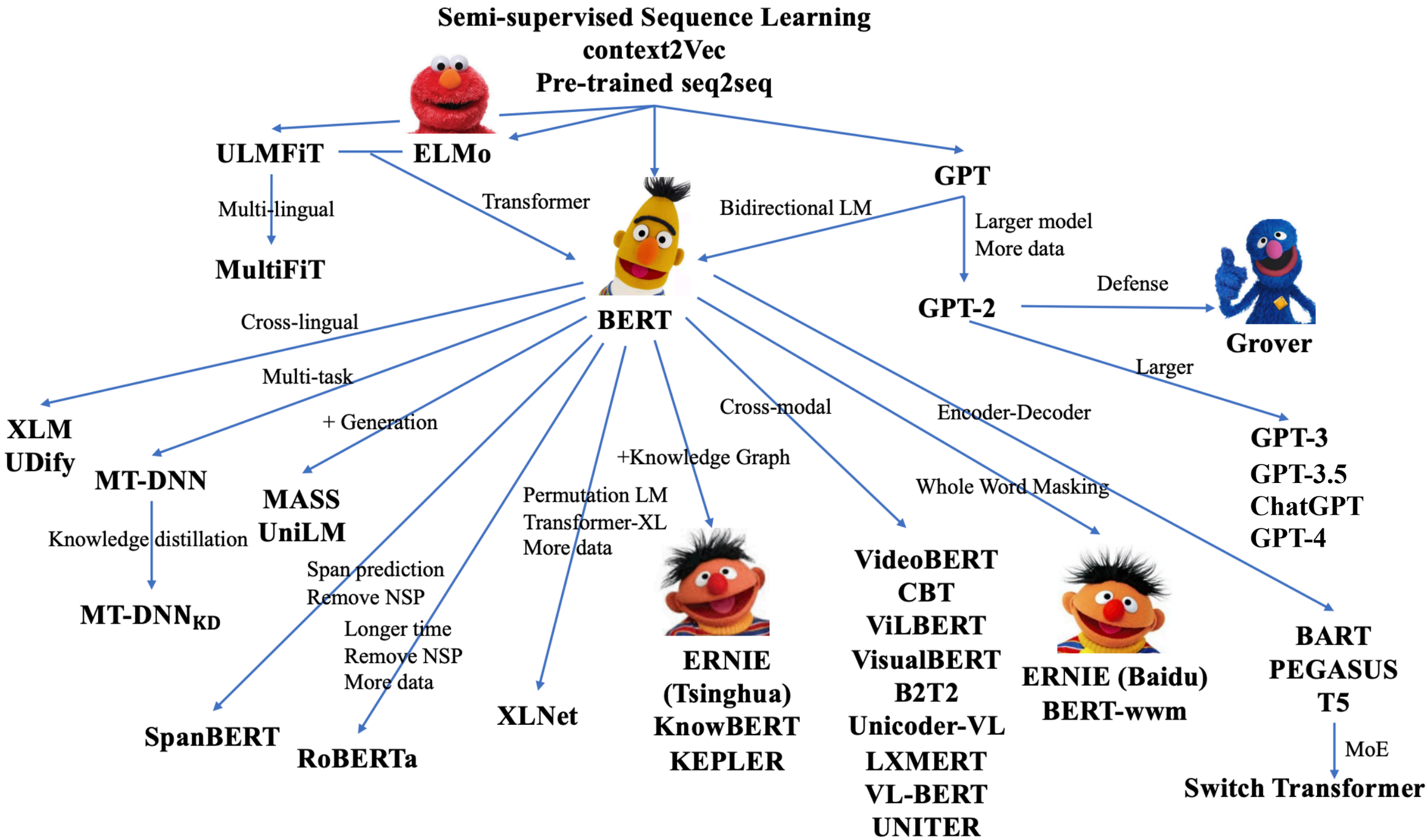
Transformer-Based LMs

- BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)
  - A bunch of Transformer encoders stacked together
  - Learns general linguistic knowledge from a large corpus
  - Training tasks: Masked Language Modeling (MLM, intra-sentential information)
    Next Sentence Prediction (NSP, inter-sentential information)
  - Auto-encoder (training a bi-directional language model)

- GPT (**G**enerative **P**re-Training **T**ransformer)
  - Pre-training a Transformer decoder for language modeling
  - Stacked 12 decoder layers, with no encoder
  - Predicting next word using massive (unlabeled) datasets
  - Auto-regressive (only training a forward language model)



The Family of Recent Typical PTMs



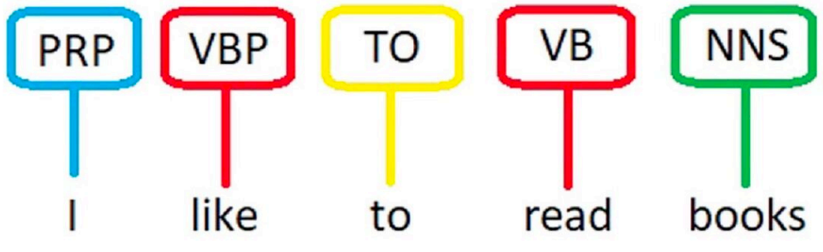
Paradigms in NLP

- How can we formulate each task?
  - e.g. sentiment analysis -> classification, part-of-speech tagging -> sequence labeling
- Major paradigms of NLP tasks (Sun et al., 2021)
  - Classification, Sentence Pair Classification or Regression, Reading Comprehension, Sequence Labeling, Sequence-to-Sequence (Seq2Seq) Modeling

e.g. NLI (sentence pair classification)

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.

e.g. POS tagging (sequence labeling)



e.g. machine reading comprehension

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

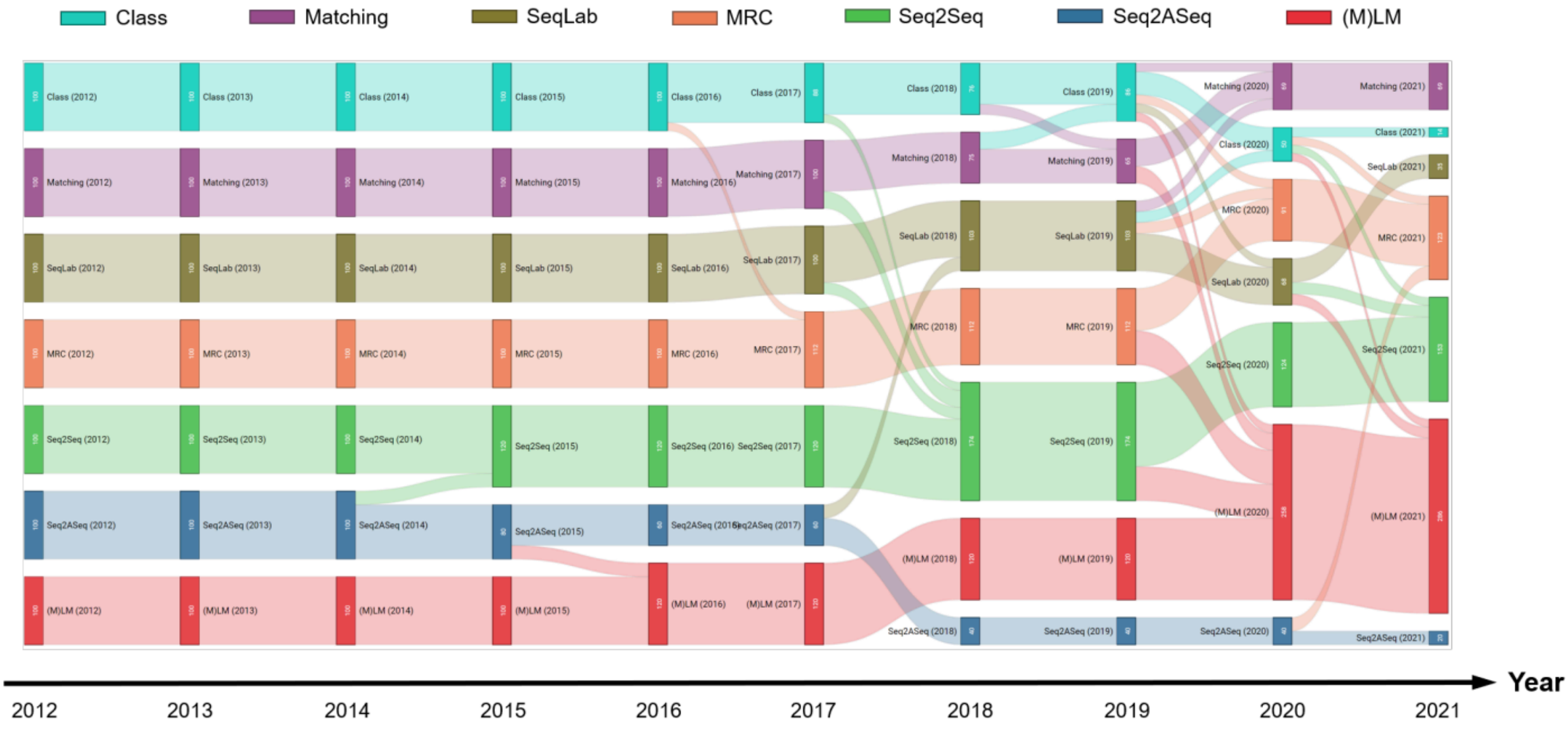
How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Sun et al. (2021). “Paradigm Shift in Natural Language Processing”. arXiv:2109.12575.

Paradigms in NLP

- Such various forms of tasks tend to be reformulated to (Masked) Language Modeling.
  - A single deployed model can serve as a unified solver for diverse NLP tasks
  - Merits: data efficiency, generalization, convenience



(M)LM

- The original input is modified with a prompt with some unfilled slots.
  - That can be filled by the pre-trained LMs
  - Then the filled word can be mapped to the label by a verbalizer

- Fully utilizing the pre-trained parameters of the MLM head
  - Instead of training a classification head from scratch
  - Few-shot and even zero-shot settings

- MLM example:

  [CLS] How are [MASK] doing today ? [SEP]

  Find the best word (with the highest probability) in the [MASK] position.

- Any other tasks can be reformulated like this:

  I hate this film . (Positive) or (Negative)?

  ==> seems like Prompt Tuning!

  [CLS] I hate this film . [SEP] This is [MASK] . [SEP]

  Good, awesome, nice, … -> Positive

  Bad, terrible, meaningless, … -> Negative

Prompt Tuning

- Classification

  [CLS] I hate this film . [SEP] This is [MASK] . [SEP]

  Prompt

  Good, awesome, nice, … -> Positive

  Bad, terrible, meaningless, … -> Negative

  Verbalizer

- Sentence Pair Classification

SNLI (entailment/neutral/contradiction)	
$<S_1>$? [MASK] , $<S_2>$	Yes/Maybe/No
$<S_1>$. [MASK] , $<S_2>$	Yes/Maybe/No
$<S_1>$? [MASK] $<S_2>$	Yes/Maybe/No

Verbalizer

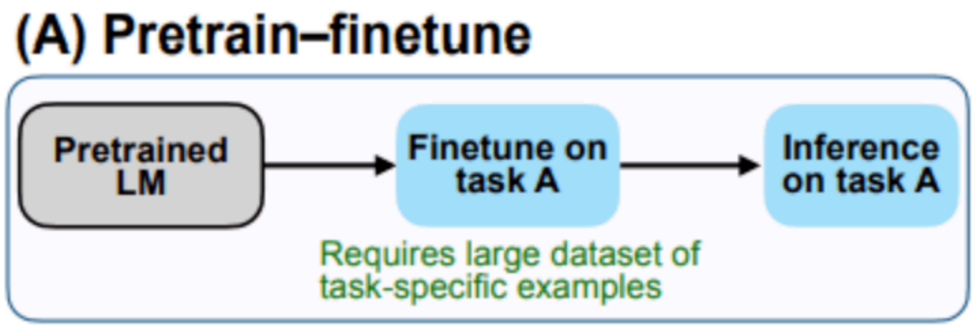
Prompt

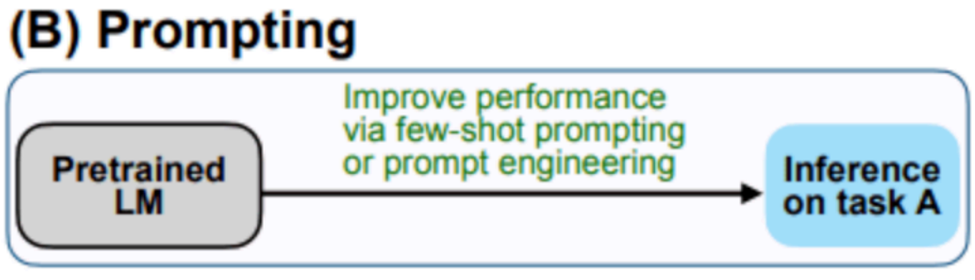
Stanford Natural Language Processing dataset

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.

Parameter-Efficient Tuning

- Traditional Fine-Tuning
  - The server has to maintain a task-specific copy of the entire pre-trained LM for each downstream task
  - Inference has to be performed in separate batches
- Prompt-based tuning
  - Only a single pre-trained LM is required
  - Different tasks can be performed by modifying the inputs with task-specific prompts
  - Inputs of different tasks can be mixed in the same batch
  - Reduces the gap between pre-training and fine-tuning of the language model



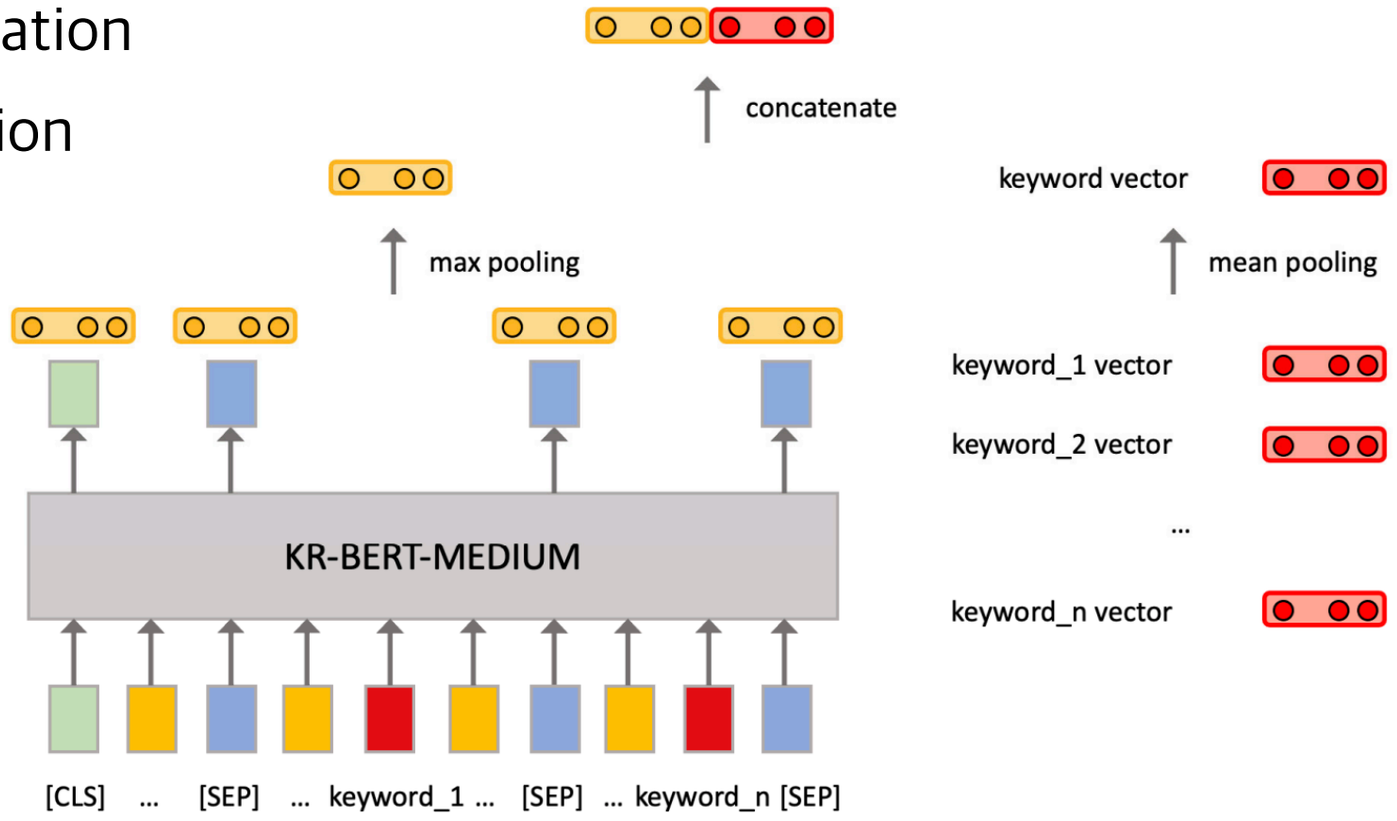


Research

- Constructing Korean-specific language models
  - Based on Korean monolingual corpora
  - Incorporating linguistic knowledge of Korean language
- Pre-trained models
  - KR-BERT (https://github.com/snunlp/KR-BERT)
  - KR-BERT-MEDIUM (https://github.com/snunlp/KR-BERT-MEDIUM)
  - KR-ELECTRA (https://github.com/snunlp/KR-ELECTRA)
- Models fine-tuned for specific purposes
  - KR-BERT-KOSAC for sentiment analysis (https://github.com/snunlp/KR-BERT-KOSAC)
  - KR-FinBERT for financial domain (https://github.com/snunlp/KR-FinBert)
  - …

Research

- Various applications utilizing the Transformer-based model architecture
  - Contract eligibility verification
    : A multi-label classification of contract clauses



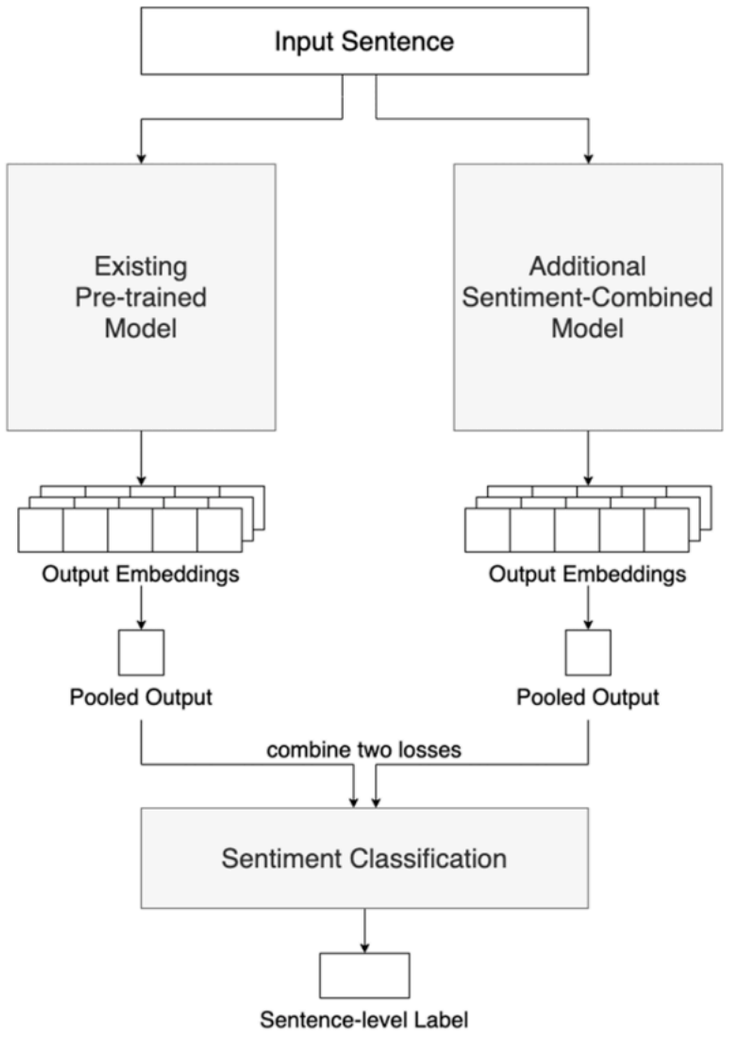
Content	Gold Label	Model Prediction
물가변동으로 인한 계약금액의 조정 [SEP] 계약상대자는 제3항의 규정에 의하여 계약금액의 증액을 청구하는 경우에는 계약 금액 조정내역서를 첨부하여야 한다. <신설 '99.11.4>	Optional3	Optional3
개인정보의 활용제한 [SEP] 본 조는 계약기간의 만료 및 계약의 해제·해지가 된 후에도 효력이 있다. 이 조항의 의무를 위반하는 경우 "수행사"는 "고객사"에게 발생한 모든 손해에 대하여 배상할 의무가 있으며, 손해배상과 별도로 이 계약의 즉시 해지사유가 된다.	Required6, Required11, Required12	Required6, Required11, Required12

이상아, 김석기, 김은진, 강민지, 신효필.(2022).키워드와 문장 임베딩을 활용한 조항별 분류모델 기반 계약서 적격성 검증.정보과학회논문지,49(10),848-858.

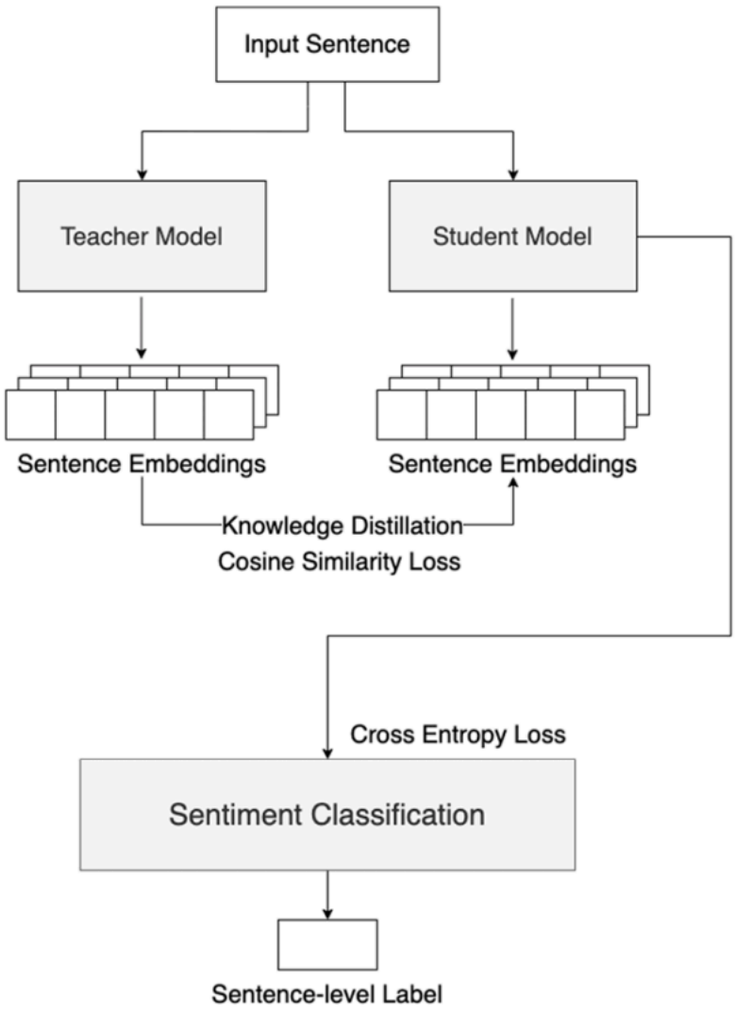
Research

- Various applications utilizing the Transformer-based model architecture
  - Sentiment analysis by combining sentiment information with BERT

External Fusing



Knowledge Distillation



이상아, 신효필.(2021).감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합.정보과학회논문지,48(7),815-824.

Research

- Others
  - Probing the “linguistic” ability of Transformer-based language models
    e.g. tasks like BLiMP
  - Semantic searching on texts of various domains e.g. biomedical, legal
  - Construction of a better multilingual (polyglot) model
    that works well with low-resourced languages (distant from English)
  - ...

감사합니다😊

Thank you for listening!

sanalee@snu.ac.kr