

DeepIntoDeep

Generative models in NLP

발표자: 박수빈

Generative models in NLP



박수빈


Artificial Intelligence in Korea University(AIKU)


Department of Computer Science and Engineering, Korea University


Introductions

- Is 'creation' only for human?

 What are the things 'only' human can do? 

다른 답안 보기 



 There are many things that only humans can do. Here are a few examples:

- **Feel emotions.** Humans are the only creatures known to experience a wide range of emotions, such as love, joy, sadness, anger, and fear. These emotions are essential to our social interactions and our ability to make decisions.
- **Be creative.** Humans are the only creatures known to create art, music, literature, and other forms of creative expression. This ability is a product of our complex brains and our ability to imagine and think abstractly.

Introductions

- Is 'creation' only made by individual's instinct?

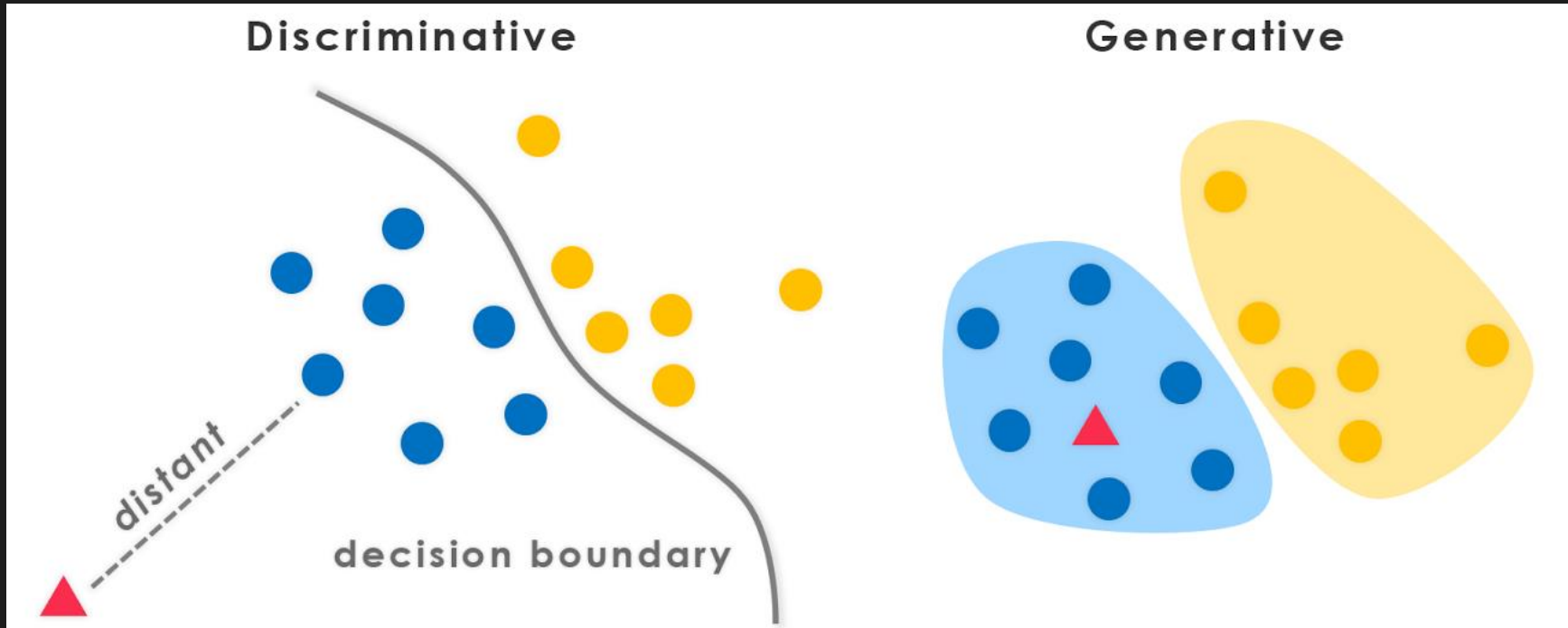
動詞

한다. 시작한다. 움직이기 시작한다. 온다. 온다. 온다. 온다. 소리난다. 울린다. 엎드린다. 연락한다. 포위한다. 좁힌다. 맞힌다. 맞는다. 맞힌다. 흘린다. 흐른다. 뚫린다. 넘어진다. 부러진다. 날아간다. 거꾸러진다. 패인다. 이그러진다. 떨어나간다. 뺏는다. 벌린다. 나가떨어진다. 떼다. 찢어진다. 갈라진다. 뽕개진다. 잘린다. 튼다. 튀어나가 붙는다. 금간다. 벌어진다. 깨진다. 부서진다. 무너진다. 붙든다. 깔린다. 긴다. 기어나간다. 붙들린다. 손 올린다. 묶인다. 간다. 끌려간다. 아, 이제 다가는구나. 어느 황토 구덕에 잠들까. 눈감는다. 눈 뜬다. 살아 있다. 있다. 있다. 있다. 살아 있다. 산다.

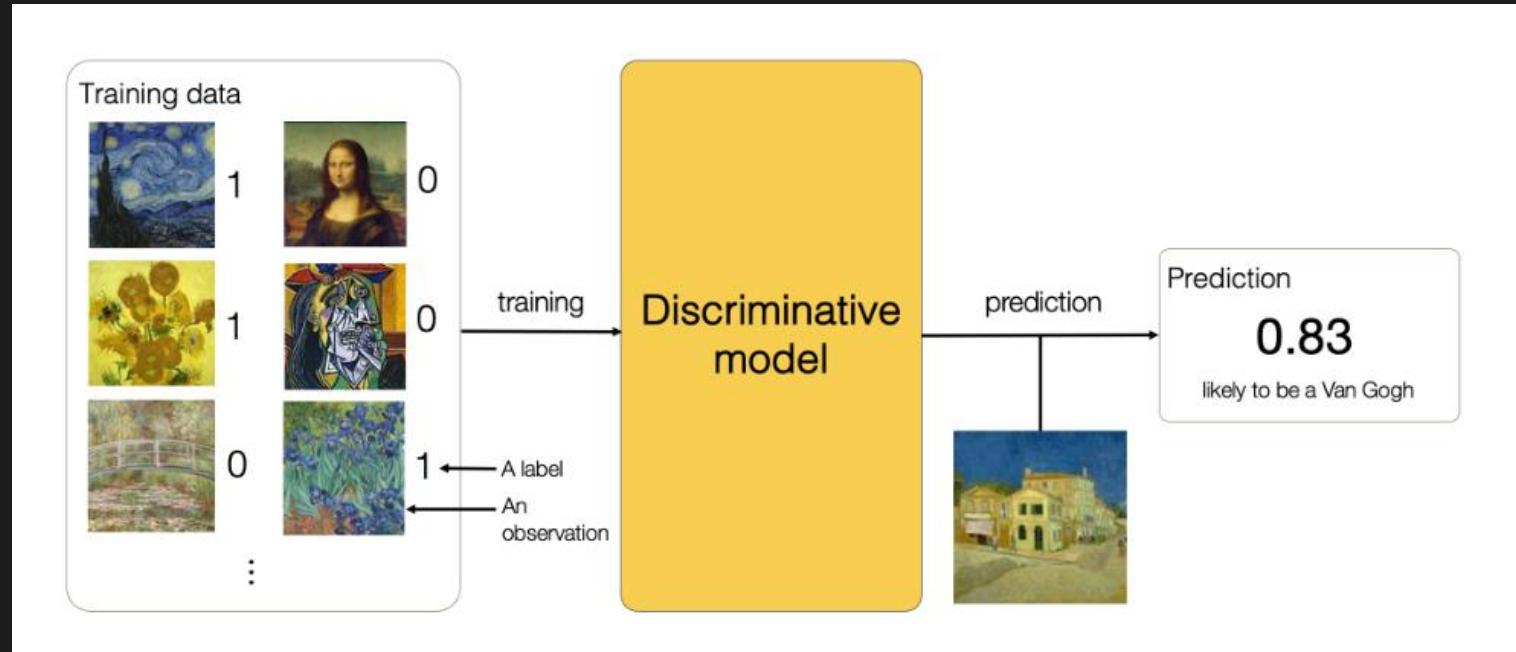
Contents

- Recap: Discriminative vs Generative
- Adaptations
- BART – Bidirectional Auto-Regressive Transformers
- T5 – Text to Text Transfer Transformer
- Generation methods – Greedy, Beam-search, Top-k, Top-p, temperature
- Evaluations example – BLEU
- Sum up
- Discussion

Recap: Discriminative vs Generative

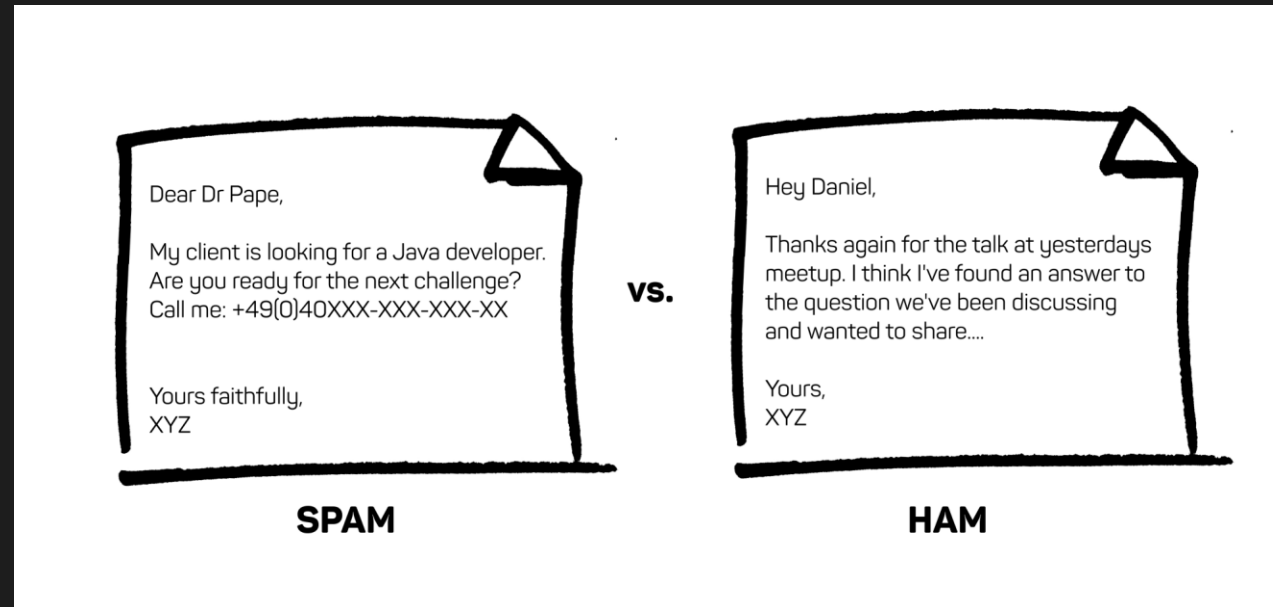


Discriminative



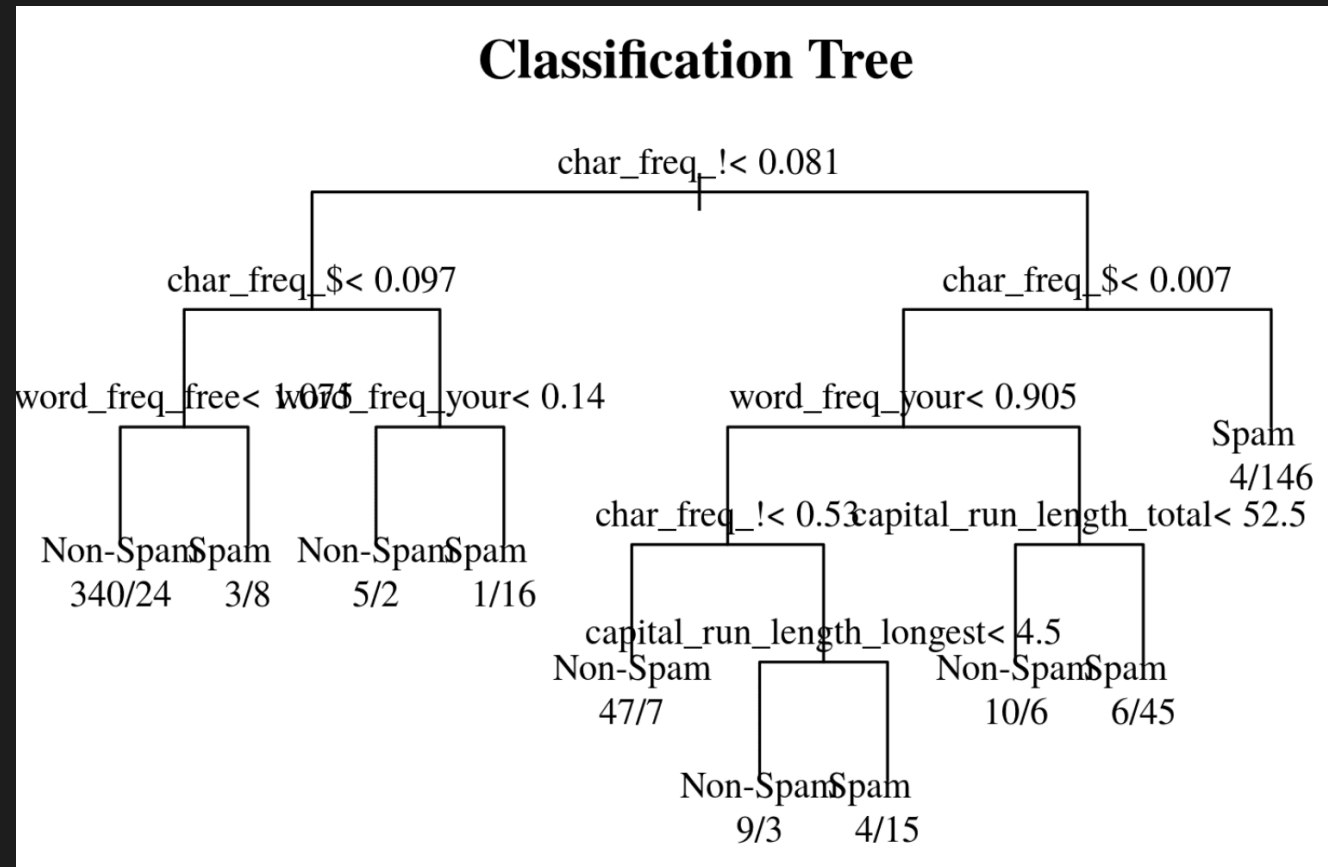
- Discriminative: Learning the ‘decision boundary’
- How? -> Identifying the patterns and correlations

Discriminative

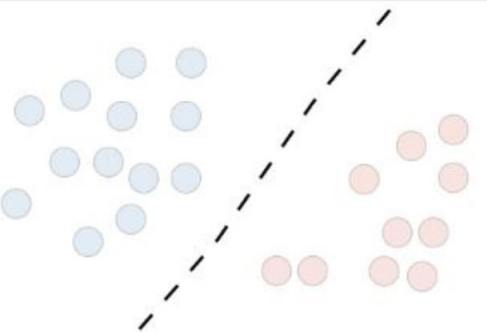
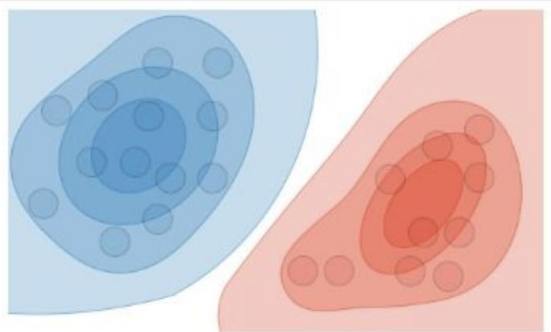


- Ex) Spam mail classification, sentimental analysis
- Simple compared to generation task, even possible with ML/statistics

Discriminative

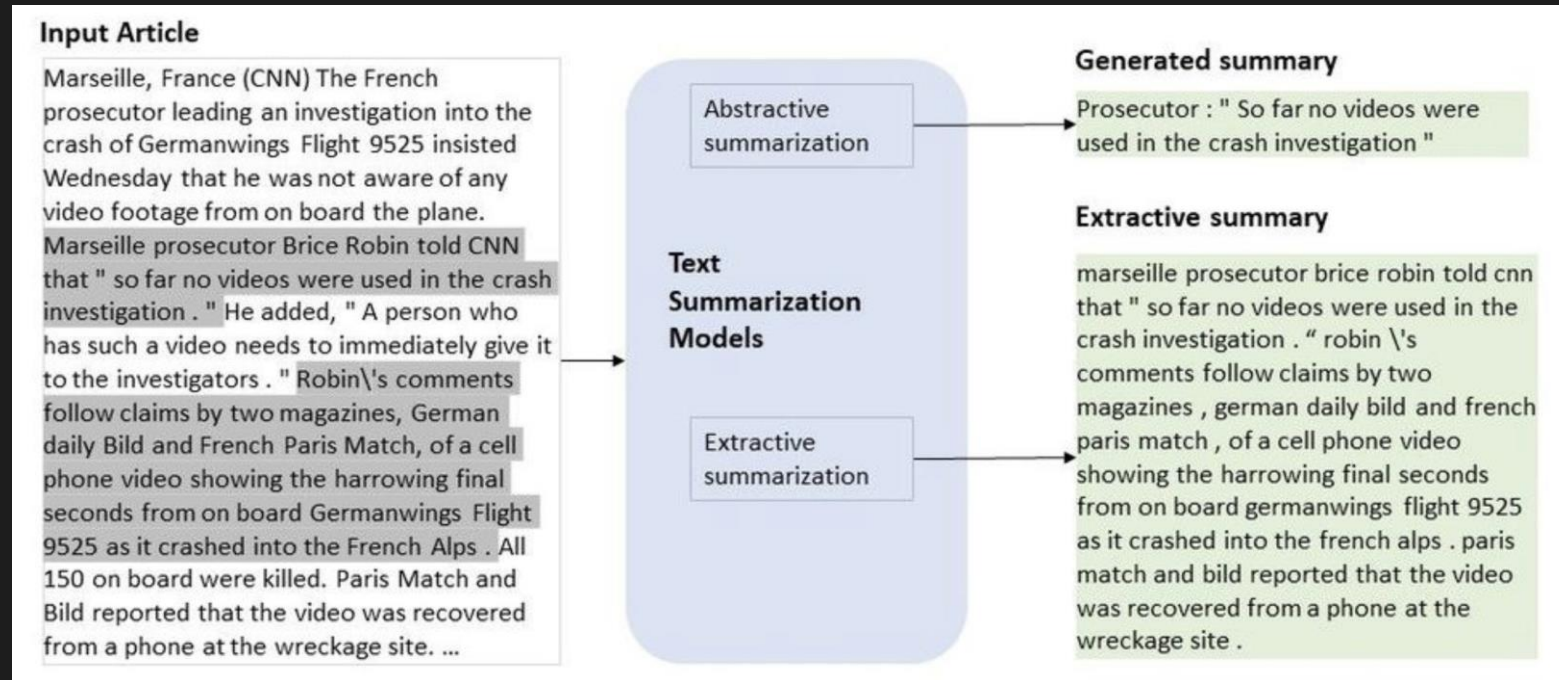


Generative

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

- By learning the probability distribution

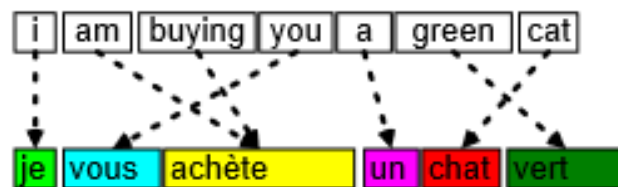
Generative



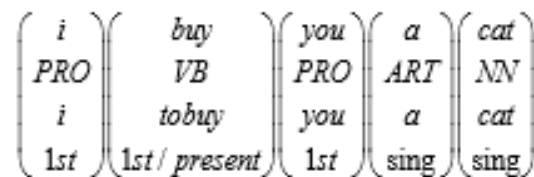
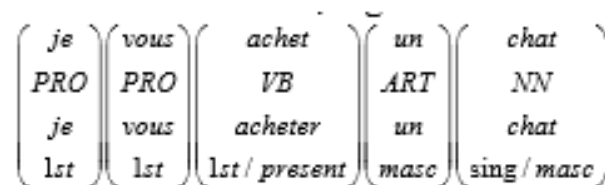
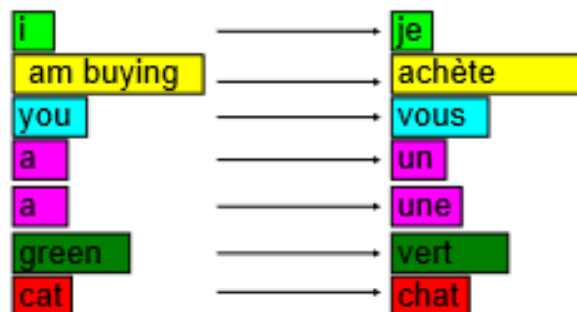
- Ex) Summarization, QA
- Needs a lot of data, resource, and quite difficult

Adaptations : Machine translation

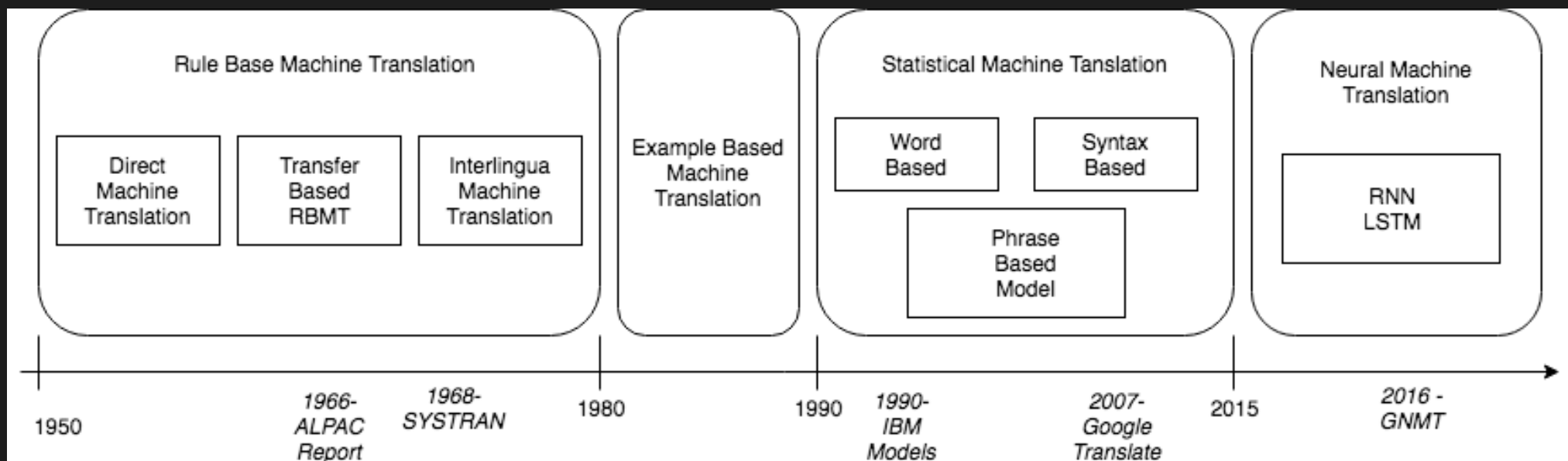
Translate:



using phrase dictionary:



Adaptations : Machine translation



Adaptations : 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

Adaptations : Next token prediction

Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:

Hannah is a ____

Hannah is a *sister*

Hannah is a *friend*

Hannah is a *marketer*

Hannah is a *comedian*

Masked-language-modeling

The model is given a sequence of words with the goal of predicting a 'masked' word in the middle.

Example

Jacob [mask] reading

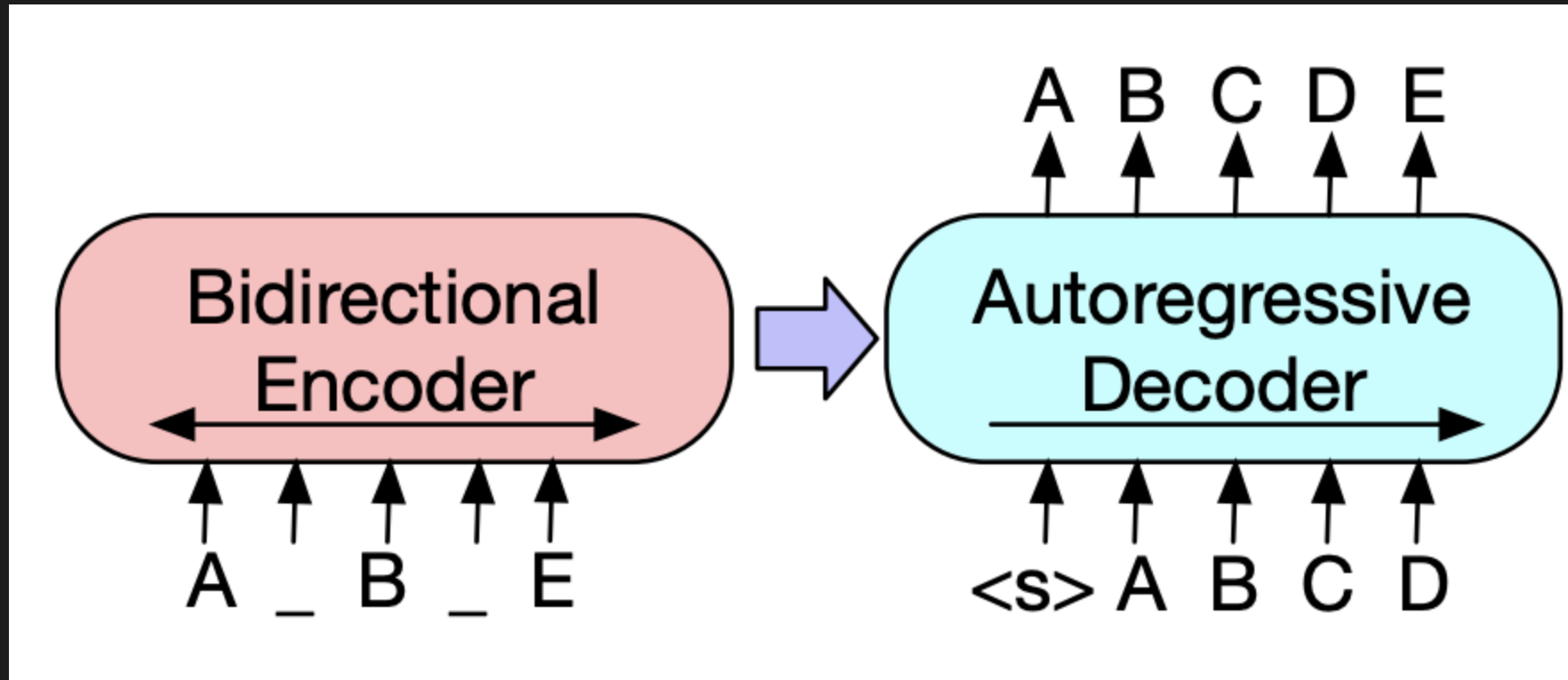
Jacob *fears* reading

Jacob *loves* reading

Jacob *enjoys* reading

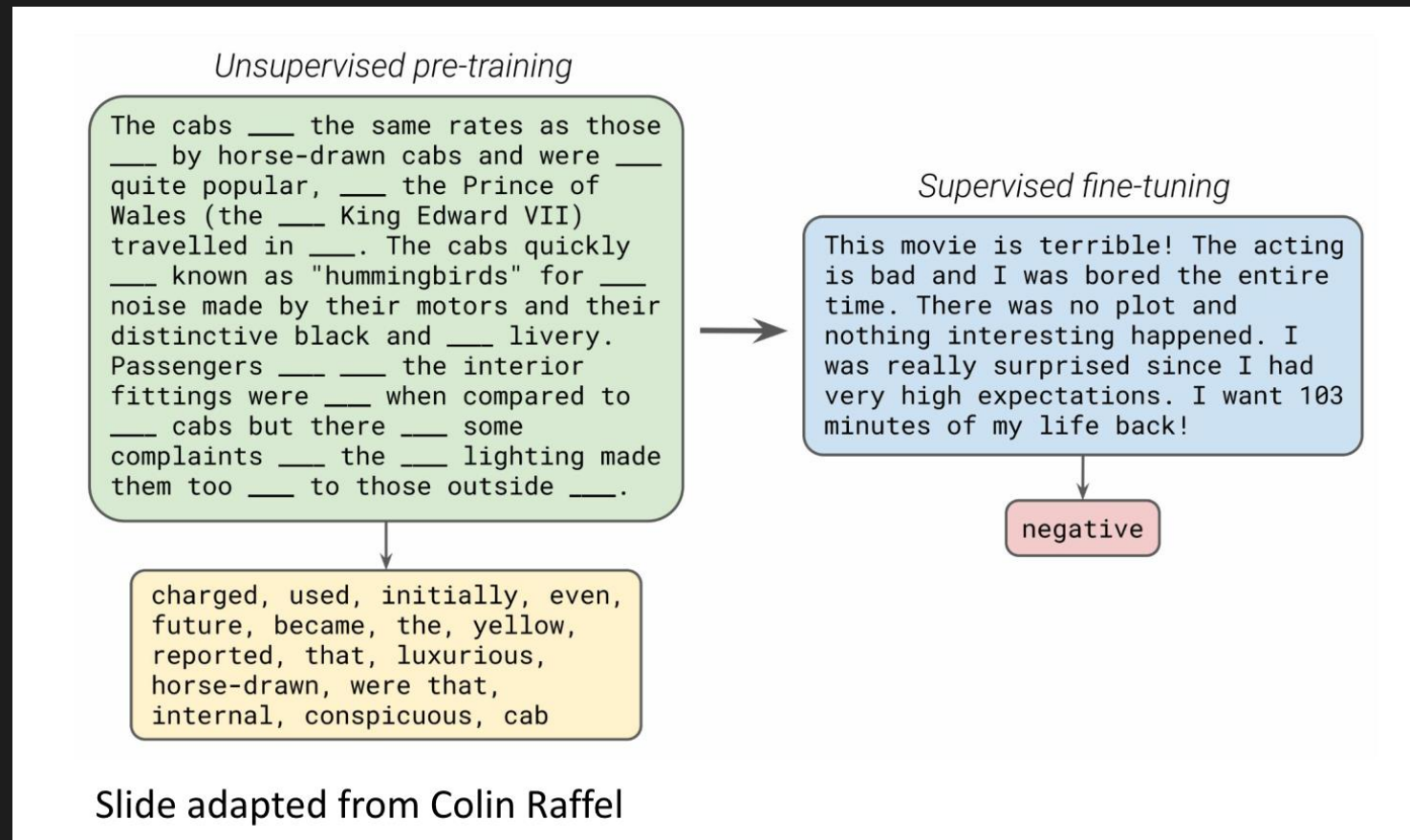
Jacob *hates* reading

BART – Bidirectional Auto-Regressive Transformers



Transfer learning?

- Supervised fine-tuning followed by unsupervised pre-training



Unsupervised pre-training

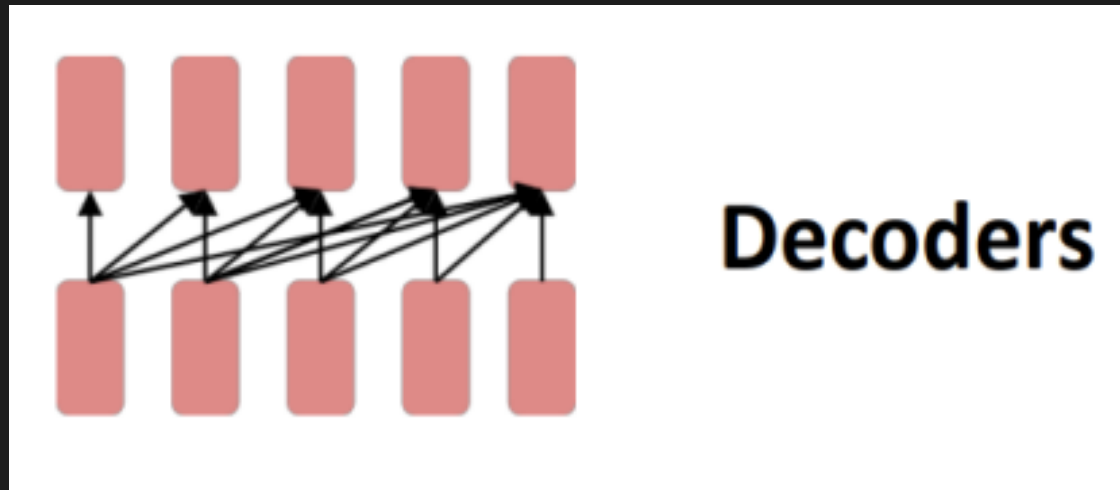
- First, to get an 'understanding' for extremely large corpus, pretrain it.
- After the pretraining, make it to do better job in specific task.

• Very general method!

- Then, how can we pretrain the models?

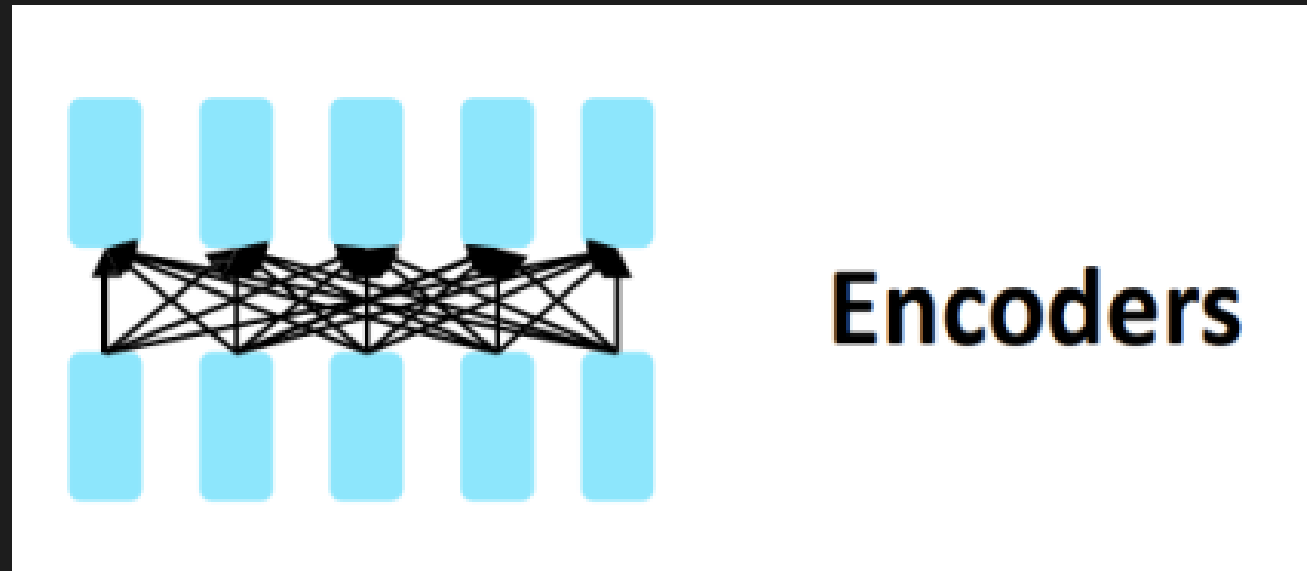
Decoder only (ex. GPT-2)

- So many LMs we see these days.
- We only can 'generate' the sequences.
- = Can't condition on future words.



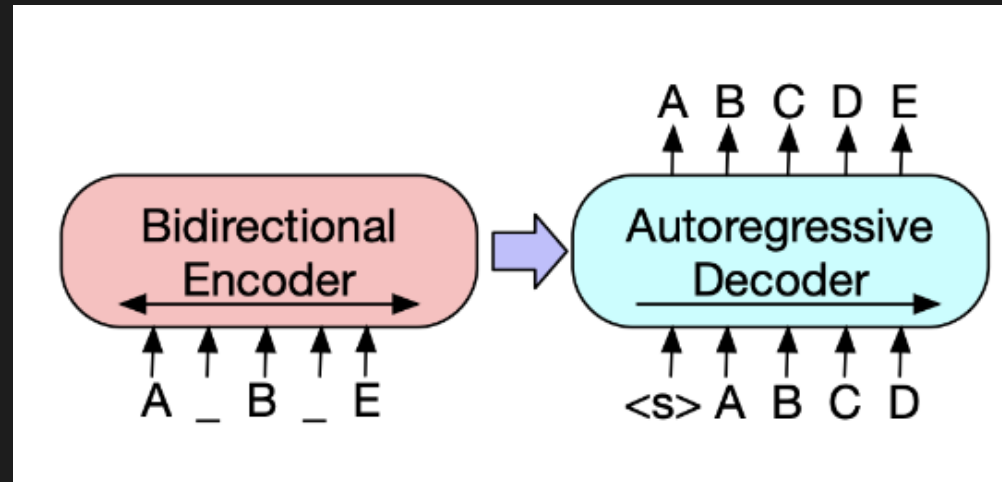
Bidirectional Encoder only(ex. BERT)

- We can get the context from both past and future.
- But saying so, this cannot be used well for pre-training.



Encoder-Decoder model (ex. BART)

- Let's get both advantages by taking these steps.



- 1. Corrupt text with several noising functions
- 2. Learn a model to reconstruct the original text
- 3. Finetune the model for diverse uses

How can we corrupt a text?

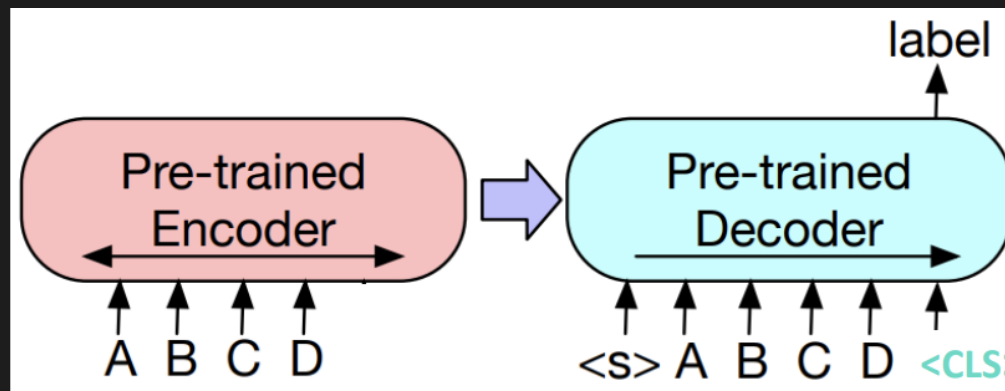
- Token masking: As always, masking random tokens.
- Text infilling : Fill a masked token
- Token deletion: Delete a random token
- Sentence permutation: Divide sentence and make permutations.
- Sentence Document rotation: `deque.rotate(n)`

How can pretrain BART?

- Of course, pretrain by corrupted text, BART tries to optimize reconstruction loss.
- Unlike previous models, we can use diverse corruption strategies to pretrain BART.

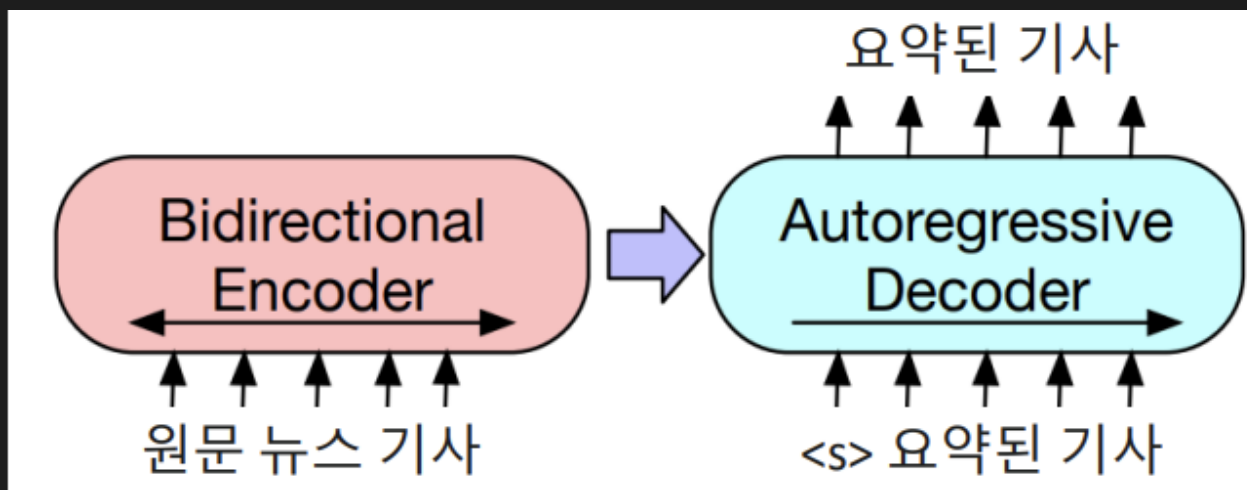
How can fine-tune BART?

- Because it has both encoder and decoder, BART can be fine-tuned for a lot of tasks, such as classification and generation.
- For sequence classification, we give the sequential data into both encoder and decoder.
- However, add <CLS> token at the last of the input of decoder.
- By this, put the output of decoder into linear classifier and get label.



How can fine-tune BART?

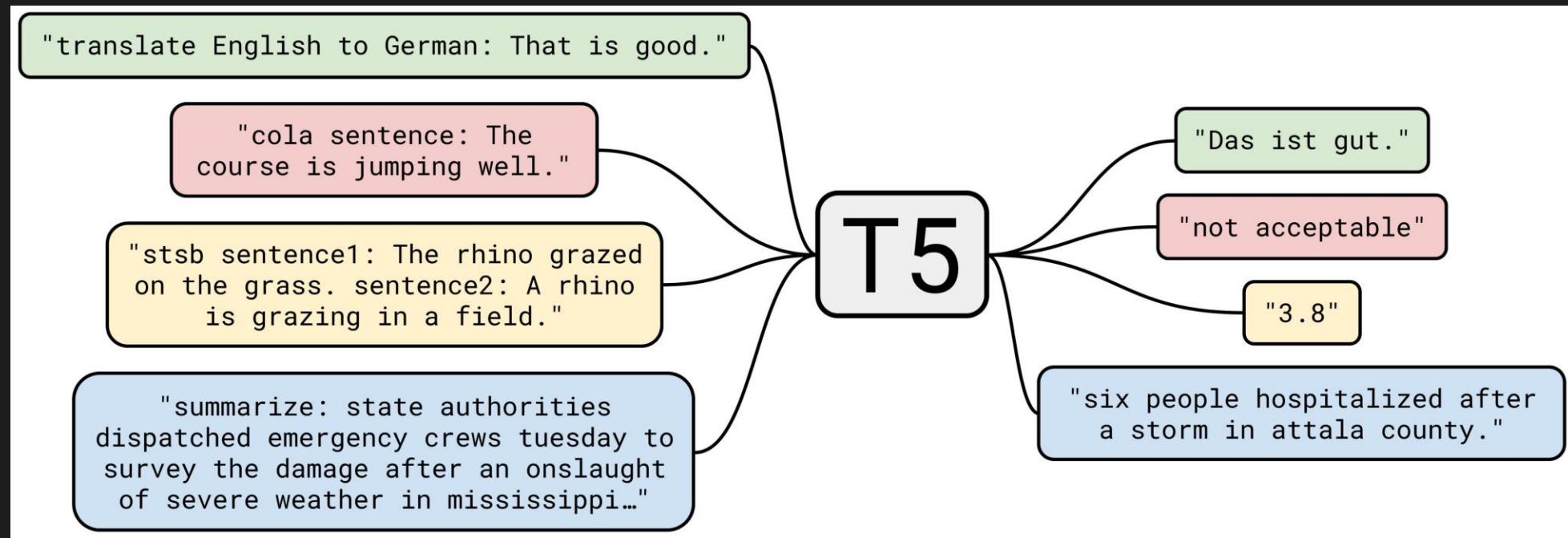
- For sequence generation, we again give the sequential data into both encoder and decoder.
- In this case, the things go much easier – pretraining objective was literally ‘generation’, creating sequence from corrupted text
- Ex) summarization, qa, etc...



How is the model structure different?

- Basically similar with Seq2Seq transformer architecture
- Activation function adjusted from LeLU to GeLU
- Base model has 6 layers for both encoder and decoder
- Large model has 12 layers for both encoder and decoder

T5 - Text to Text Transfer Transformer



Again, transfer learning

- Unsupervised pretraining + supervised fine-tuning
 - Except for some multi modal problem, NLP tasks are mostly text-to-text.
 - Question – Answer sets are both texts.
 - Raw text – Summarization are both texts.
 - Text – Sentiment are both texts.
-
- Then, can't we just train a single model to solve text-to text problem?

Multiple tasks done in single model

- Same model, same object, same training / decoding procedure
- By adding the 'prefix' into the input, model can identify the task.

summarize_topic: “처음에는 ‘금방 끝나겠지’라고 생각했는데 어느덧 100일이 됐네요. \ 그동안 춥고 아프고 힘들었지만 인간으로서 대우를 받을 수만 있다면 끝까지 버틸 수 있습니다.” \ LG트윈타워 청소 노동자들이 고용승계를 주장하며 파업에 나선지 100일째를 하루 앞둔 24일 \ 서울 여의도 LG트윈타워 앞 ‘행복한 고용승계 텐트촌’에서 만난 박상설(63)씨는 힘들었던 투쟁 과정을 \ 회상하며 눈시울을 붉혔다. 박씨는 2017년부터 LG트윈타워에서 청소 노동을 했지만 지난 1월 1일부로 \ 계약이 종료돼 직장을 떠났다. 자동차 소음과 불편한 잠자리로 텐트에서 매일 밤잠을 설치지만 투쟁을 \



'LG트윈타워 청소 노동자가 고용승계를 주장하며 파업에 나선지 100일째를 하루 앞둔 24일 서울 \ 여의도 LG트윈타워 앞 ‘행복한 고용승계 텐트촌’에서 만난 박상설(63)씨는 힘들었던 투쟁 과정을 \ 회상하며 눈시울을 붉혔다. 반면 노동자들은 2019년 노조를 결성하고 권리를 주장하기 시작하면서 사측 \ 는 밖에 났다고 주장한다. 때문에 반드시 LG트윈타워에서 정당한 권리를 인정받고 노동을 이어가야 \

But, how?

- The concept is explicit and persuasive.
- Maximizing the utility of unsupervised pre-training
- But “EXTREMELY LARGE” dataset should be provided... how?

C4 dataset (The Colossal Clean Crawled Corpus)

- Let's use the web-crawled texts! (20TB -> 750GB)
- So much useless data in web, such as gibberish, placeholder, htmls...
- Clean the data from heuristics.
 - 1. Only use sentence with punctuation in the end.
 - 2. Do not use pages less than 5 lines.
 - 3. Lines must be consisted with more than 3 letters.
 - 4. Exclude Dirty, Naughty, Obscene or Otherwise Bad Words
 - 5. Exclude lines including 'Javascript' (to exclude caution..)
 - 6. Exclude pages including 'lorem ipsum' (to exclude placeholder)

C4 dataset (The Colossal Clean Crawled Corpus)

Menu

Lemon

Introduction

The lemon, Citrus Limon (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Article

The origin of the lemon is unknown, though lemons are thought to have first grown in Assam (a region in northeast India), northern Burma or China. A genomic study of the lemon indicated it was a hybrid between bitter orange (sour orange) and citron.

Please enable JavaScript to use our site.

Home
Products
Shipping
Contact
FAQ

Dried Lemons, \$3.59/pound

Organic dried lemons from our farm in California. Lemons are harvested and sun-dried for maximum flavor. Good in soups and on popcorn.

The lemon, Citrus Limon (L.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae. The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses. The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur in tempus quam. In mollis et ante at consectetur. Aliquam erat volutpat. Donec at lacinia est. Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit. Fusce quis blandit lectus. Mauris at mauris a turpis tristique lacinia at nec ante. Aenean in scelerisque tellus, a efficitur ipsum. Integer justo enim, ornare vitae sem non, mollis fermentum lectus. Mauris ultrices nisl at libero porta sodales in ac orci.

```
function Ball(r) {  
  this.radius = r;  
  this.area = pi * r ** 2;  
  this.show = function(){  
    drawCircle(r);  
  }  
}
```

Slide adapted from Colin Raffel

Fine – tuning with multiple tasks

- After pre-training with the C4 dataset, fine-tune it with multiple tasks.
- SQuAD : Extractive QA task. (Answer included in the document given)
- CoLA : Sentence acceptability
 - Ex) 나는 집을 샀다 : unacceptable
- STS-B : Sentence similarity
 - Ex) 투명 드래곤은 크아앙 울부짖었다. / 투명 드래곤은 크앙 부르짖었다 – 유사

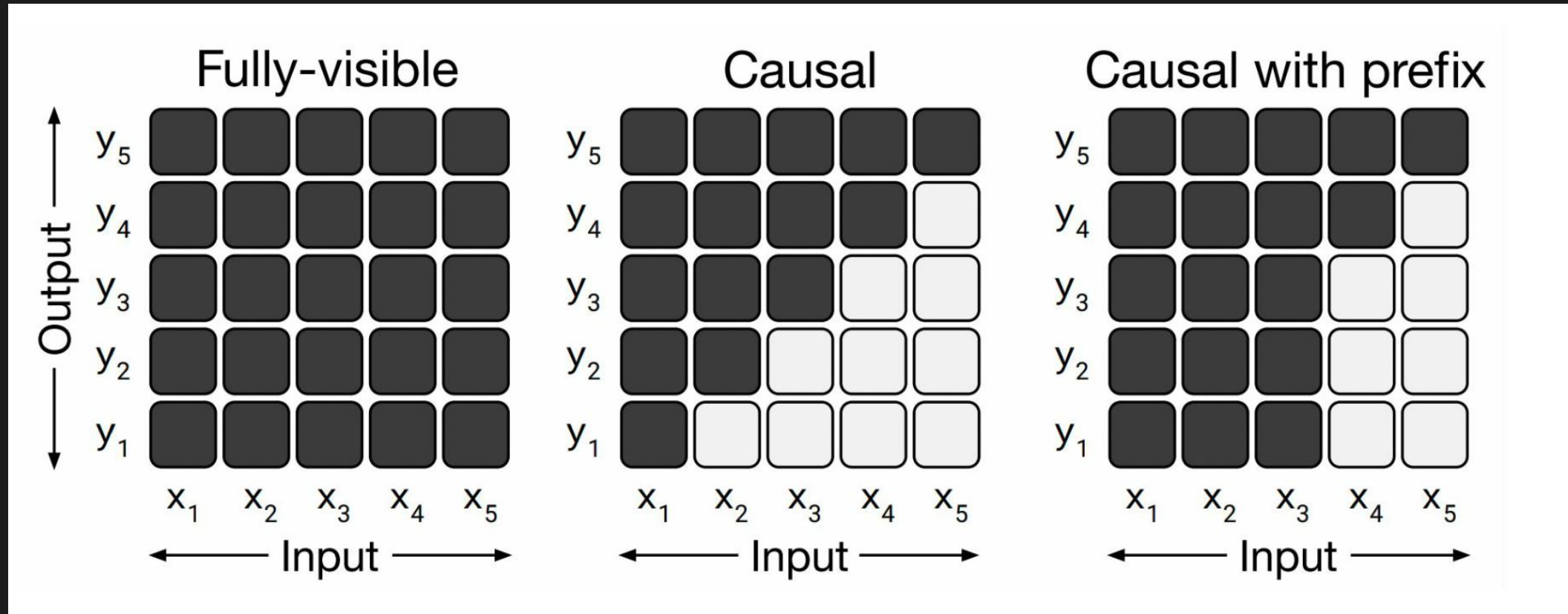
Fine – tuning with multiple tasks

- CoPA : Casual reasoning
 - Ex) 나는 길을 가다가 넘어졌다. 1과 2 중무슨 일이 일어나겠는가?
 - 답 1: 억만장자가 되었다.
 - 답 2: 무릎이 까졌다.
- ReCoRD/MultiRC : QA/Reading comprehension
 - Ex) 투명 드래곤은 크아양 울부짖었다. / 투명 드래곤은 크앙 부르짖었다 – 유사

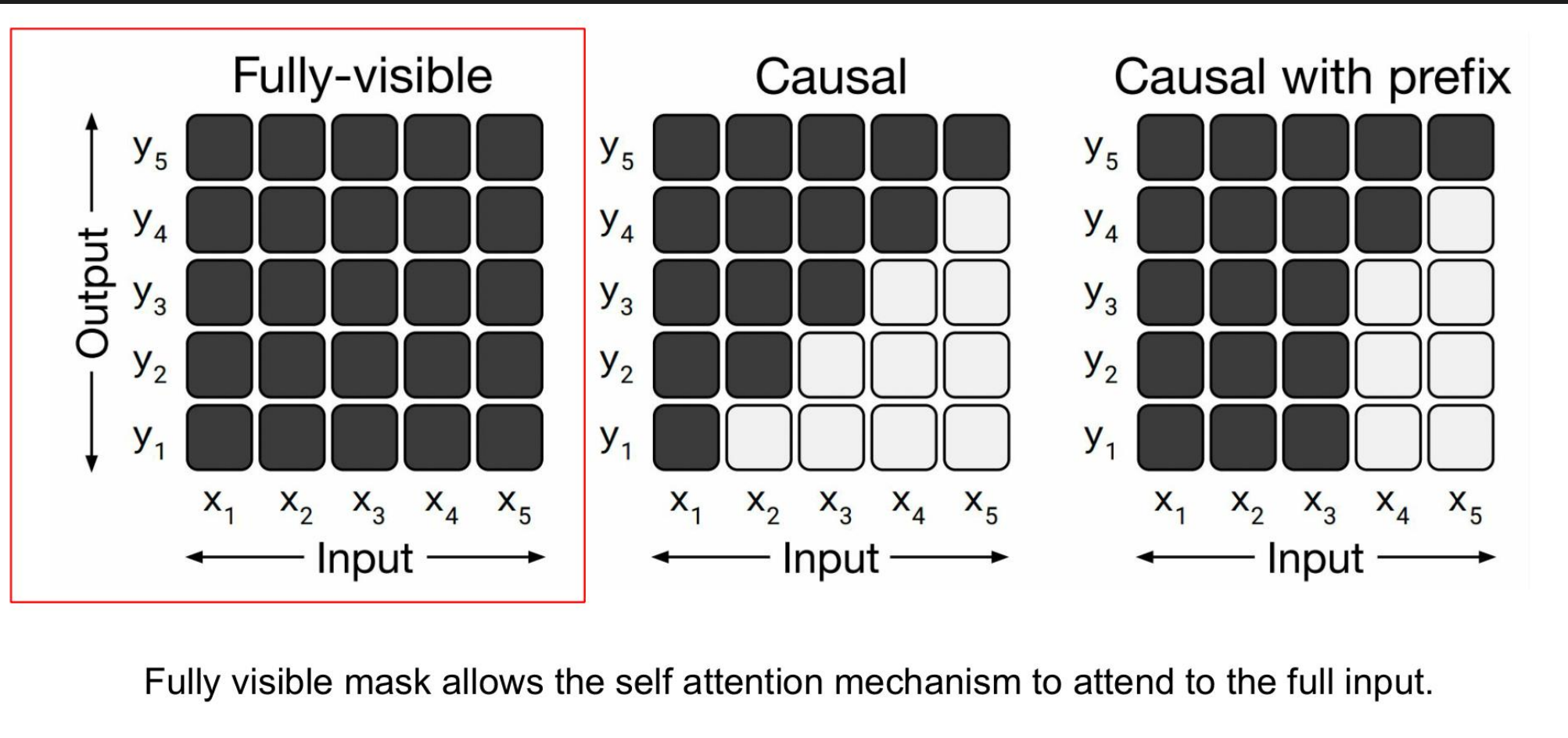
Model architecture

- Basic Encoder – decoder transformer model (Vaswani et al. (2017))
- Both encoder and decoder is similar with BERT base model.
- Encoder and decoder is consisted with 12 blocks.
- Attention structure differ to process 'prefix'!

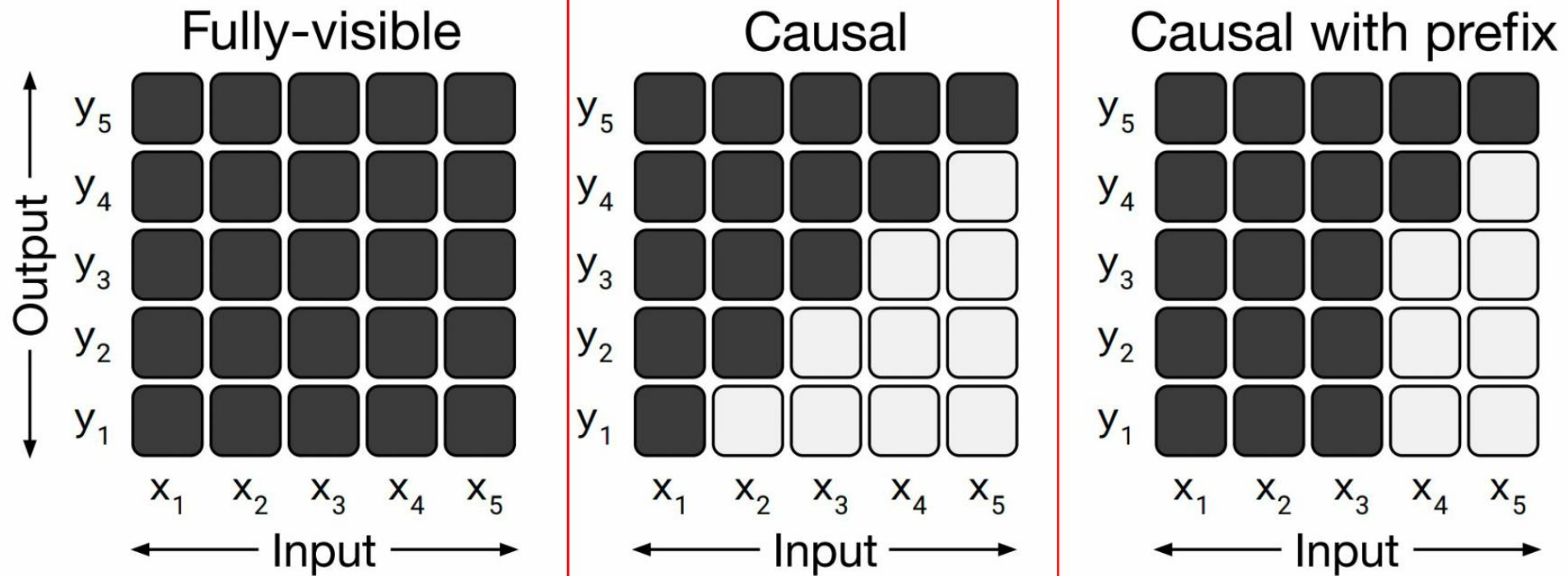
Different attention mask patterns



Fully – visible attention



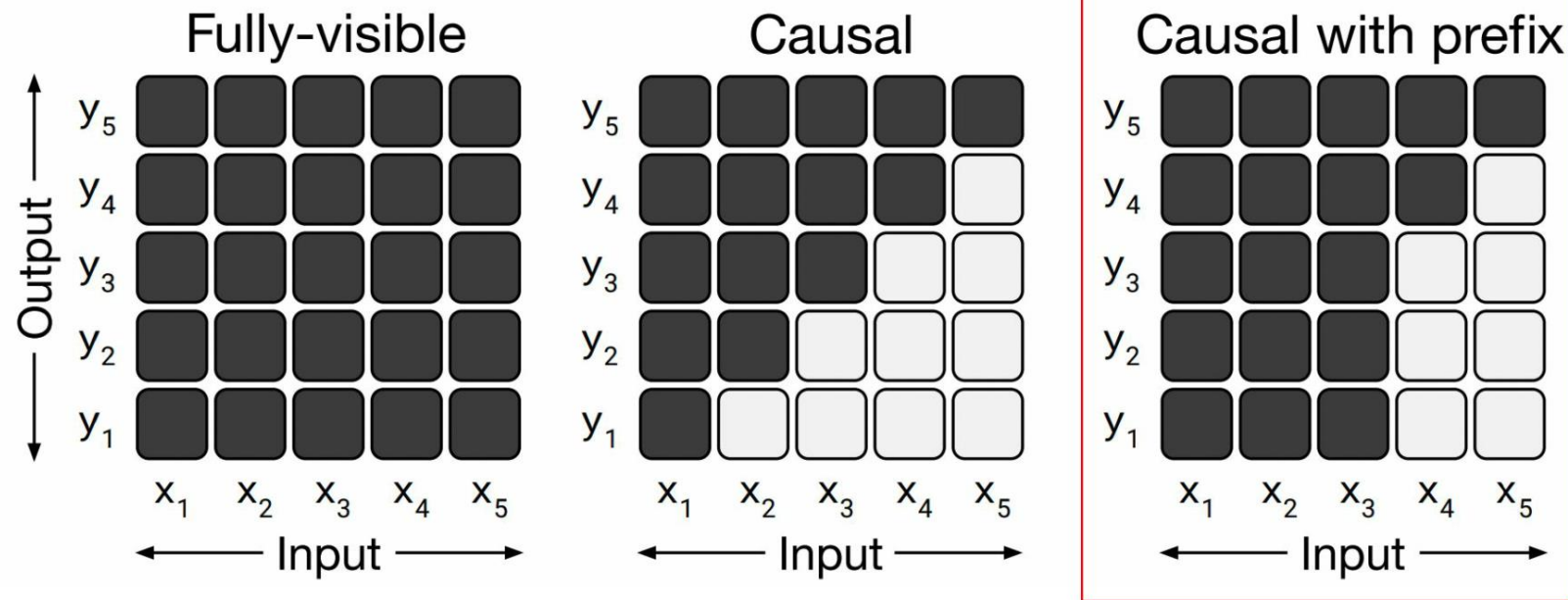
Causal attention



A causal mask doesn't allow output elements to look into the future.

Causal attention with prefix

Different Attention Mask Patterns



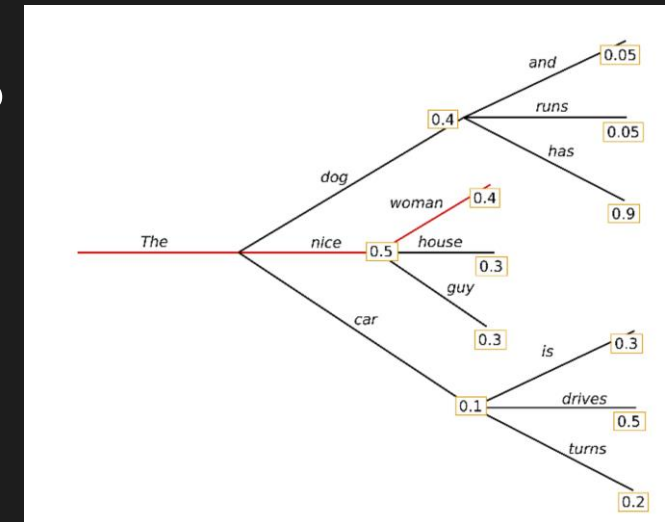
Causal mask with prefix allows to fully-visible masking on a portion of input.

How can we infer?

- Let's say we got the distribution of the next word.
- But in sequential data, one inferences affect the other.
- How can we infer the 'overall' correct answer?

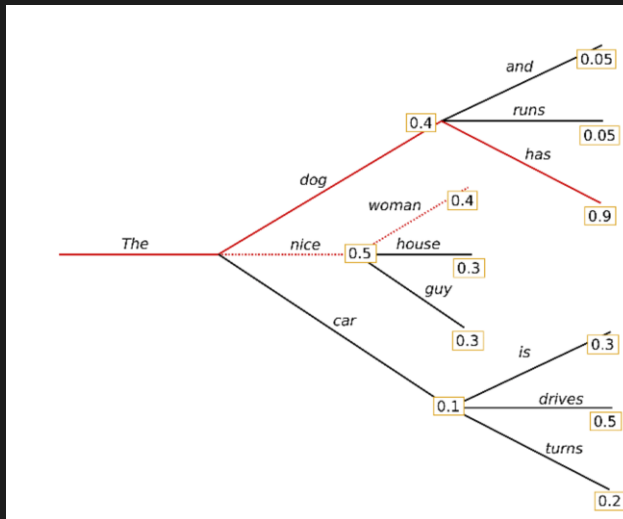
Greedy method

- Basic method : choose the most-likely word at certain point.
- Fast, easy to understand and make.
- However, what if the probability of most-likely word and the second doesn't differ a lot?
- What if my wrong selection affects the later ones?
- We must infer for timesteps of times-
in this method, one wrong answer is critical

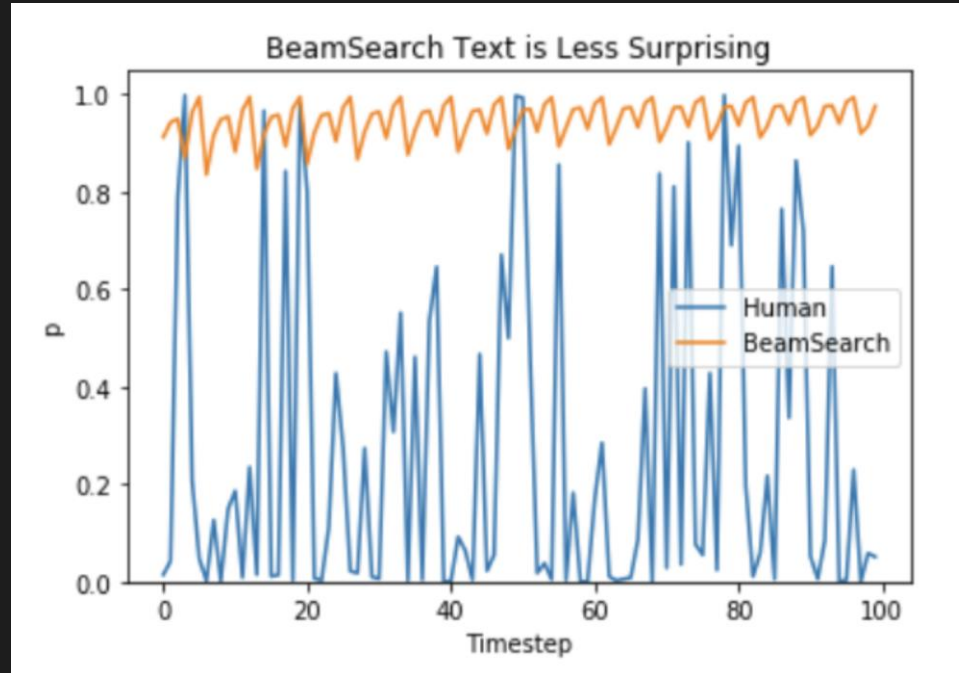


Beam Search

- For k-beams, search the most likely beam at the time.
- By this, we avoid one wrong answer resulting critical error.
- It makes the inference slow, especially when k is large.
- However, the overall answer might become more likely.

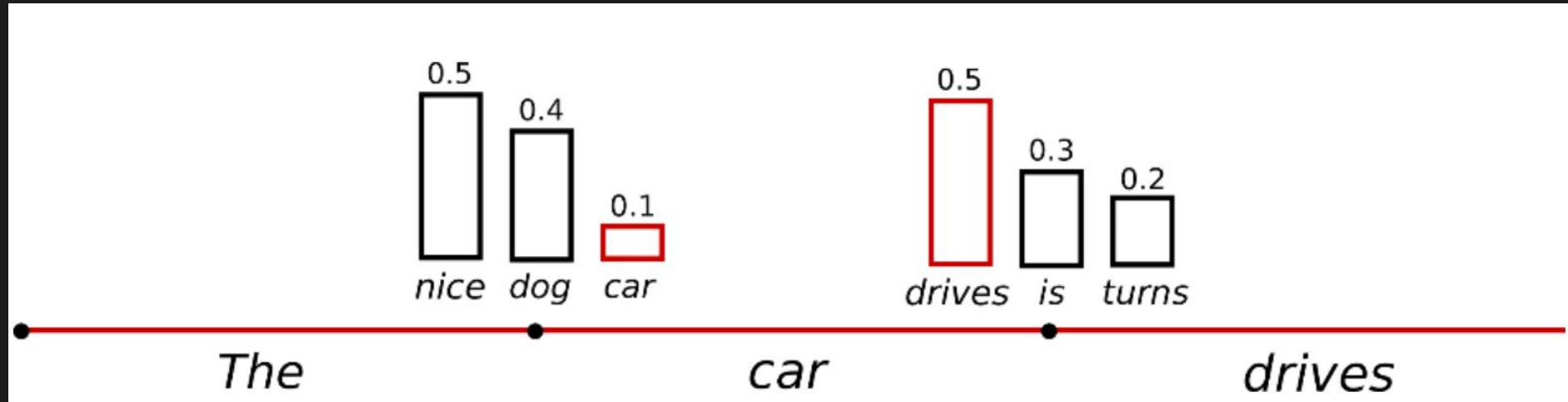


Critical weakness : Beam search is boring



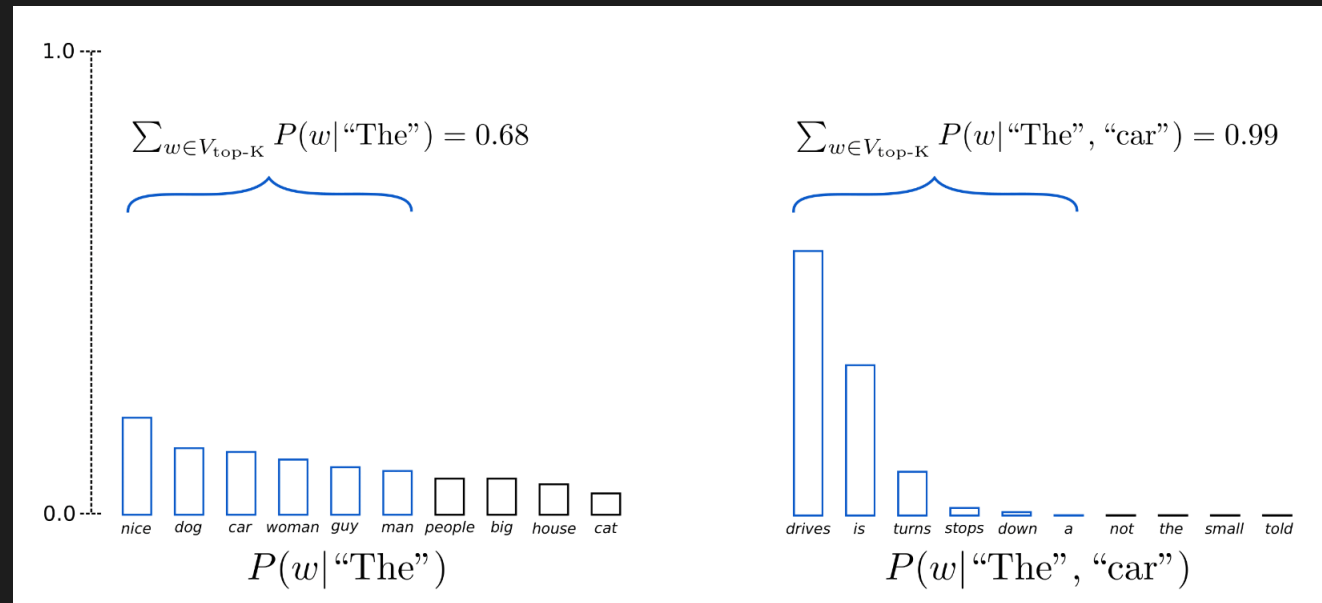
- This might be good for machine-translation, but not for situation when open-ended answer is needed.
- Let's add the 'Randomness'!

Sampling



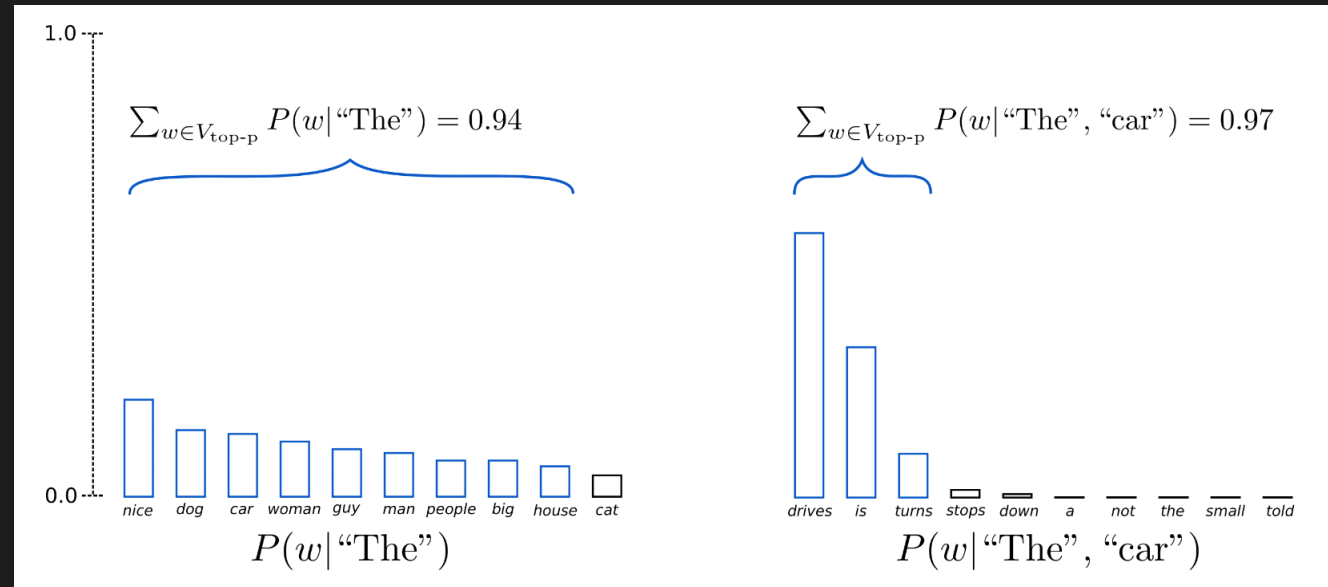
- Previously, the selection made was by choosing the word with 'highest probability' 100%.
- Now, let's see the probability as real 'probability'.
- Ex) 50% probability for choosing 'drives' as answer.

Top-k Sampling



- Even so, including all words seems dangerous...
- Let's only include top-k tokens for sampling.
- Again, might be boring.

Top-p Sampling



- Let's choose the words from the set which has accumulative probability p
- This gives us random answer compared to top-k, but not always better.

BLUE score and ROUGE score

- BLEU(Bilingual Evaluation Understudy) score
How many words of the generated sentence are in reference sentence?
- ROGUE(Recall-Oriented Understudy for Gisting Evaluation)
How many words of the reference sentence are in generated sentence?
- Generated sentence: I was generated by the model
- Reference sentence: I was referenced by human

$$\text{ROUGE: } \#\{w_{ref} \in S_{gen} | w_{ref} \in S_{ref}\} / |S_{ref}| = 3/5$$

$$\text{BLEU: } \#\{w_{gen} \in S_{ref} | w_{gen} \in S_{gen}\} / |S_{gen}| = 3/6$$

Sum up

- Generative models structure
- Inference methods are important, too
- But the most important part is 'DATA'.
- Mostly, what matters is data size...
- Model size are big, and costly training.
- -> Not recommended for toy project / or individual project
- Though, very attractive and promising
- LLM?

Discussion

- 1. Top-k가 Top-p보다 좋은 경우는 언제일까요? 간단하게 예시를 들어 주세요.
- 2. C4 dataset을 정제하는 기준이 ppt에 기재한 것 외에도 더 존재합니다. 어떤 기준이었을지 생각해보고 이유와 함께 답해주세요.

감사합니다.