자연어 처리 주요문제 이해

Sangkeun Jung

강의 내용

- 일상생활의 다양한 자연어처리 문제를 살펴보고, 해결 방법에 대해 고민한다.
- 대표적인 NLP 문제들을 해결하는 형태에 대해 학습한다.
- 대표적인 NLP Benchmark 에 대해 이해한다.

일상생활의 다양한 NLP 문제

NLP in Real Life |Disaster Tweets

Disaster Tweets

- 어떤 Tweet이 진짜 재난인지 아닌지 분류
- Tweet이 입력되면 0과 1이 출력하는 문제로 진짜 재난일 경우 1이 출력

"Three people died from the heat wave so far" → 1: 재난

"you're ablaze and alive outside, you're dead inside" → 0: 재난 아님

Source : https://www.kaggle.com/competitions/nlp-getting-started/overview

Extracting Novel Findings from Texts

➢ 텍스트가 입력되면 생물의학 개체 타입 추출

Late-onset metachromatic leukodystrophy: molecular CHEBI pathology in two siblings.We report on a new allele SO at the arylsulfatase A GGP (ARSA) locus SO causing late-onset metachromatic leukodystrophy (MLD). In that allele SO arginine84, a residue that is highly conserved SO in the arylsulfatase gene GGP family, is replaced SO by glutamine CHEBI . In contrast to alleles SO that cause early-onset MLD, the arginine84 to glutamine substitution CHEBI is associated with some residual ARSA activity. A comparison of genotypes SO , ARSA GGP activities, and clinical data on 4 individuals TAXON carrying the allele SO of 81 patients with MLD examined, further validates the concept that different degrees of residual ARSA activity are the basis of phenotypical variation in MLD..

Source : https://ncats.nih.gov/funding/challenges/litcoin

코드 유사성

- 두 코드간 유사성(동일 결과물 산출 가능한지) 여부를 판단
- Python 코드 A와 Python 코드 B가 입력
  - 두 코드가 같으면 1
  - 두 코드가 다르면 0

A 코드

```
# Python 3+
#-----------------------------------------------
------------------
import sys

#ff = open("test.txt", "r")
ff = sys.stdin

cnt = 1
for line in ff :
if int(line) == 0 : break
print("Case {0}: {1}".format(cnt, line),
end="")
cnt += 1
```

1: 같다
0 : 다르다

B 코드

```
flag = "go"
cnt = 0
while flag == "go":
cnt += 1
if x == 0: x = int(input())
flag = "stop"
else:
print("Case " + str(cnt) + ": " +
str(x))
```

Source : https://dacon.io/competitions/official/235900/overview/description

NLP in Real Life | Image Caption

Image Caption

- 이미지를 통해 설명하는 캡션(설명)을 만들어 내는 것
- 입력되는 이미지를 통하여 이미지를 설명하는 텍스트를 생성

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

Source : https://arxiv.org/abs/1411.4555

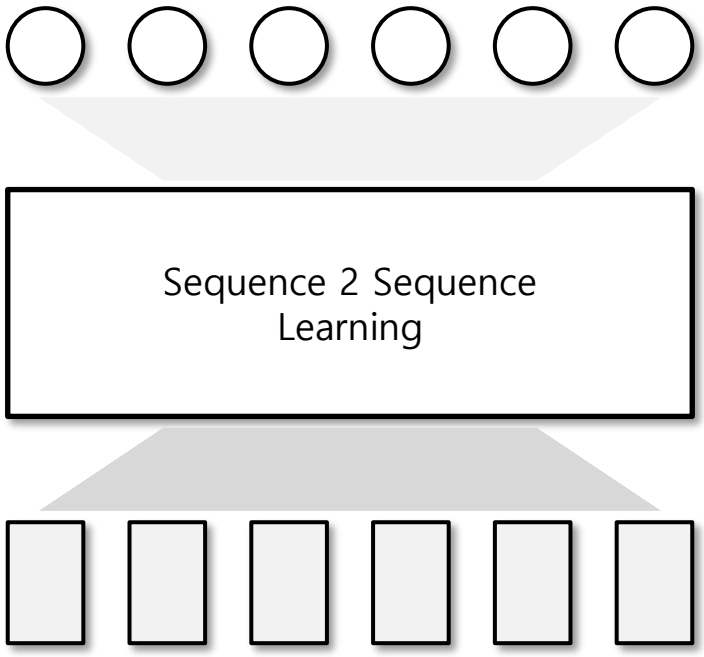
주요 NLP Tasks

주제분류	도로 위 맥주 2000병 와르르… 처치 도운 시민	→	사회
감성분석	그래도 감독님 응원합니다.	→	긍정
개체명분석	수학 A형의 1등급 커트라인은 평균 96점이었다	→	수학 A형의 <1등급:QT> 커트라인은 평균 <96점:QT>이었다
형태소분석	밥 먹기 싫대	→	밥/NNG+먹/VV+기/ETN +싫/VA+대/EF
대화	곧 점심 시간이네	→	뭐 먹고 싶은거 있어?
번역	오늘 날씨가 좋네요	→	The weather is good today
⋮		⋮	

NLP 문제 해결 패턴

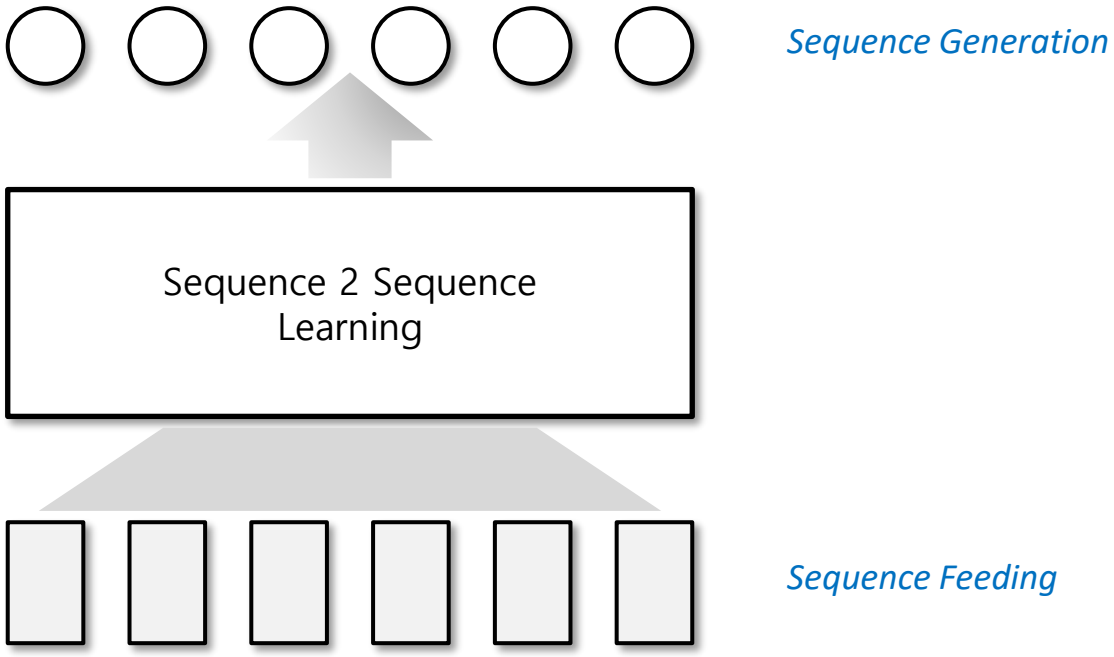
NLP Problem Pattern | Sequence to Sequence

Sequence to Sequence (S2S)



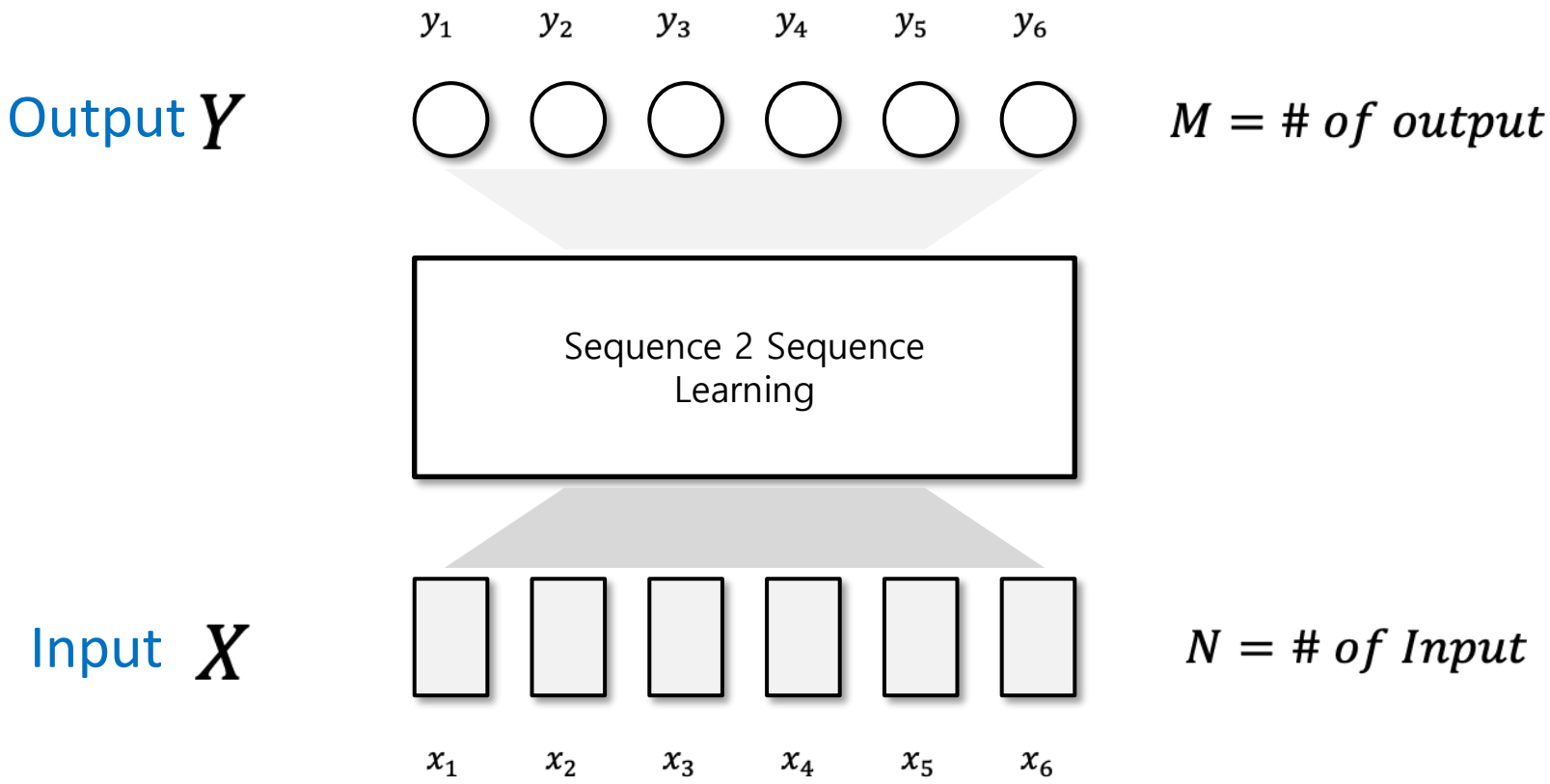
NLP Problem Pattern | Sequence to Sequence | Learning @ running time

Sequence to Sequence (S2S)



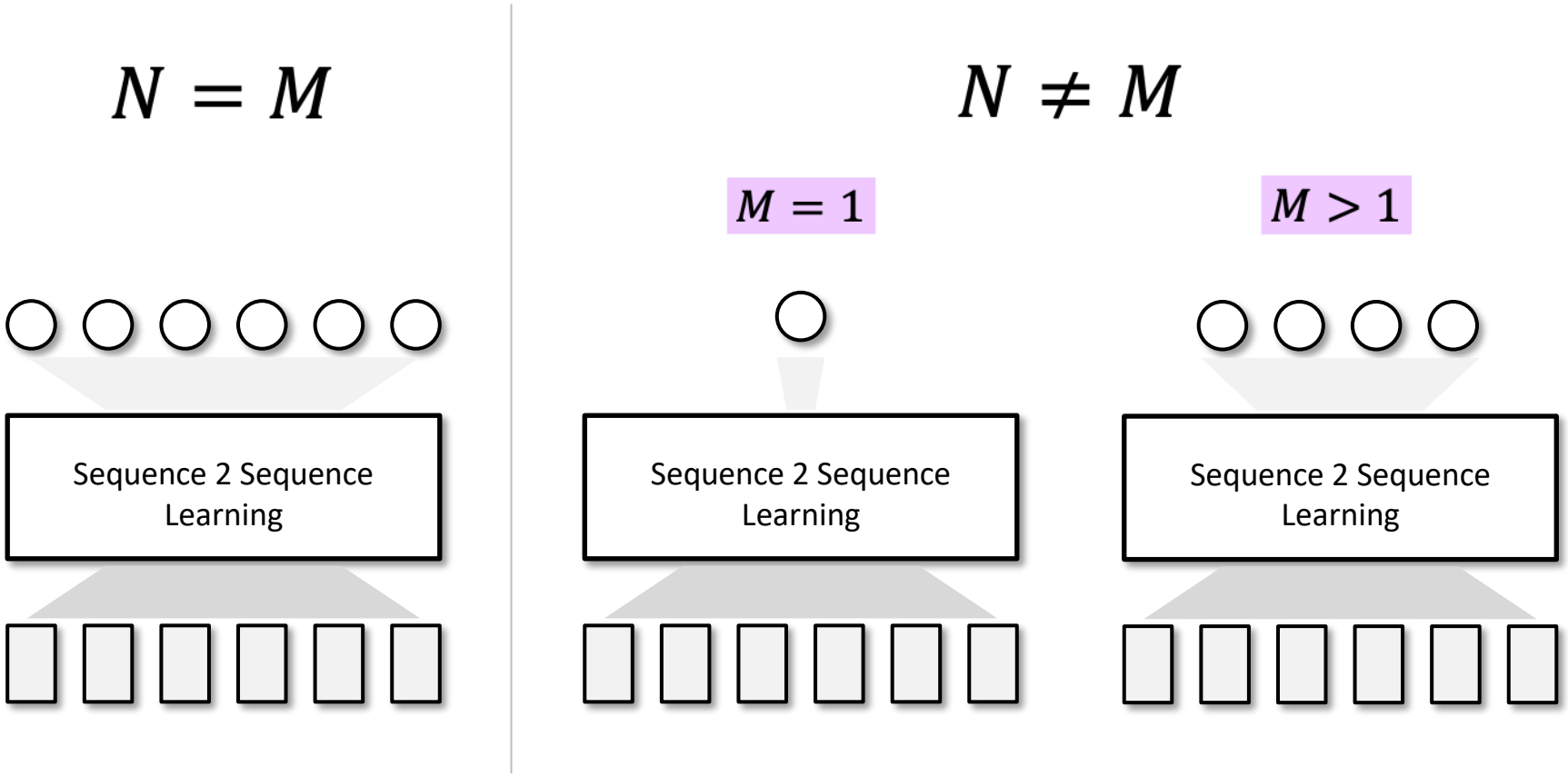
NLP Problem Pattern | Sequence to Sequence | Definition

Sequence to Sequence (S2S)

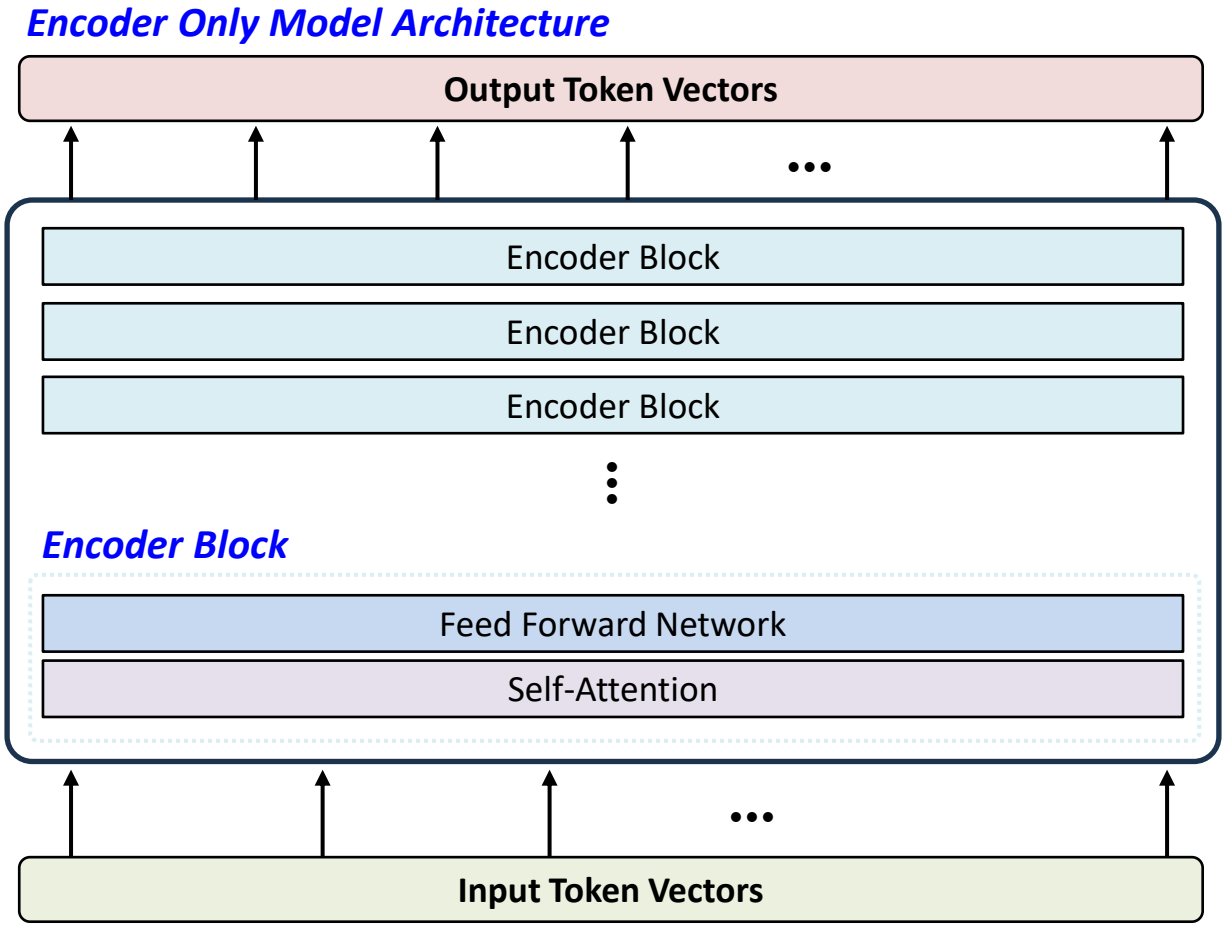


NLP Problem Pattern | Sequence to Sequence | Cases

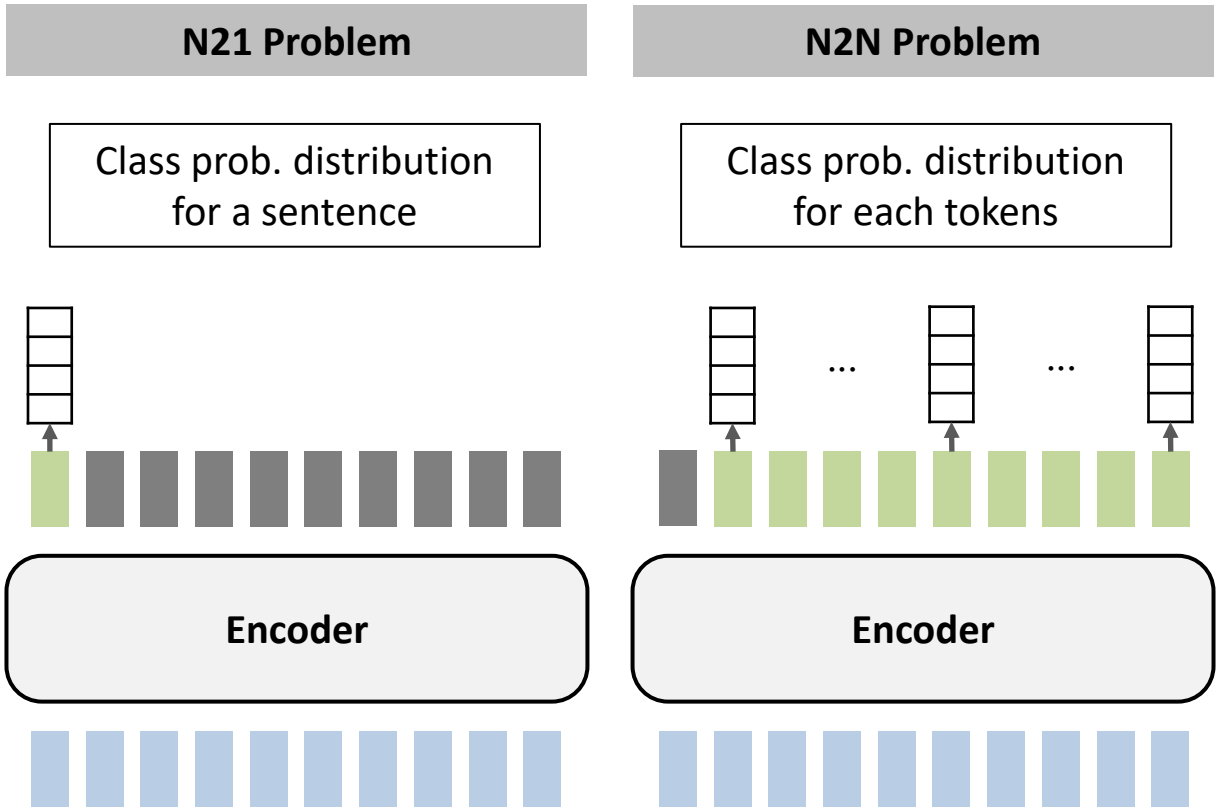
Cases



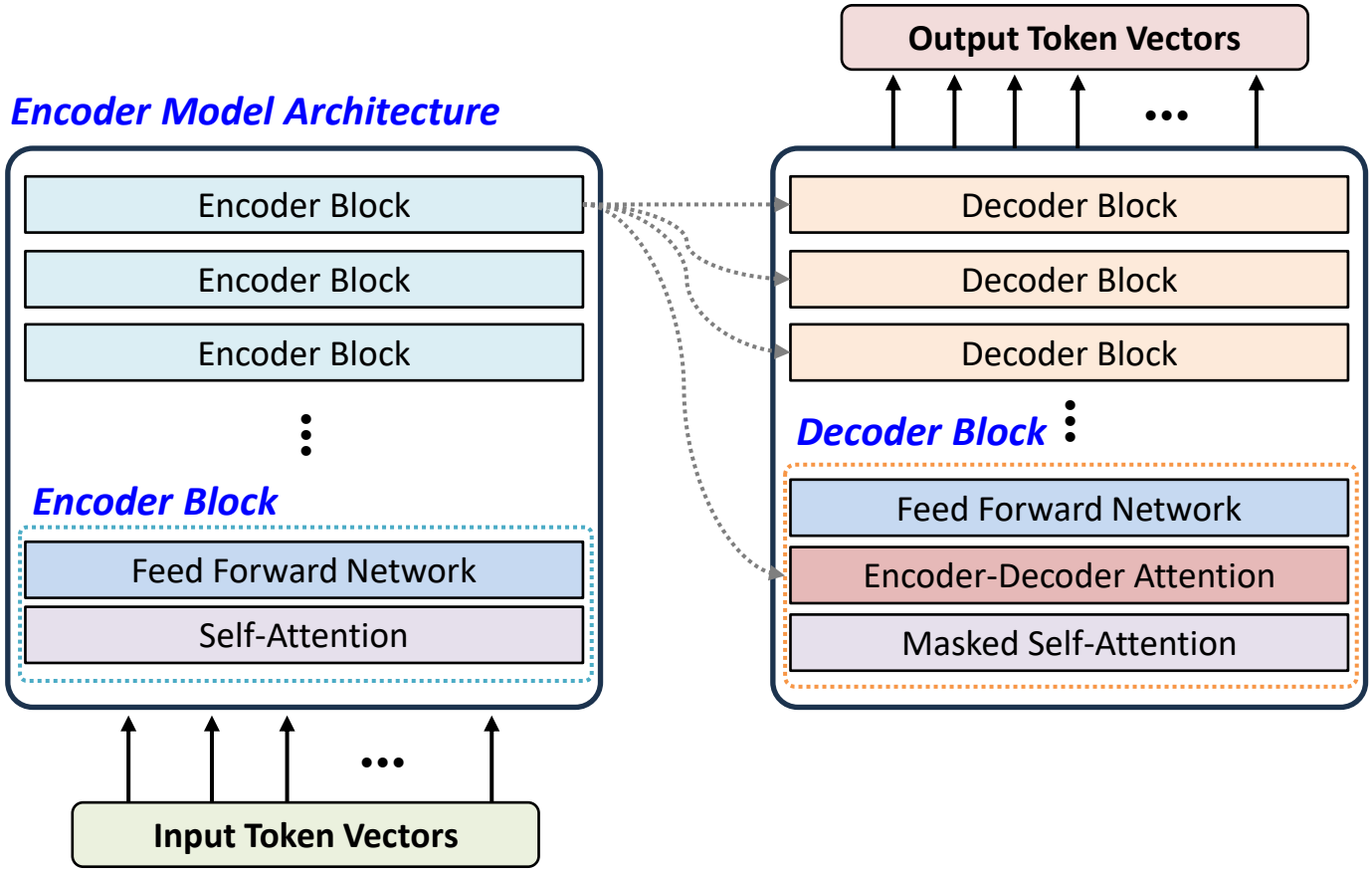
NLP Problem Pattern | Encoder Only Model



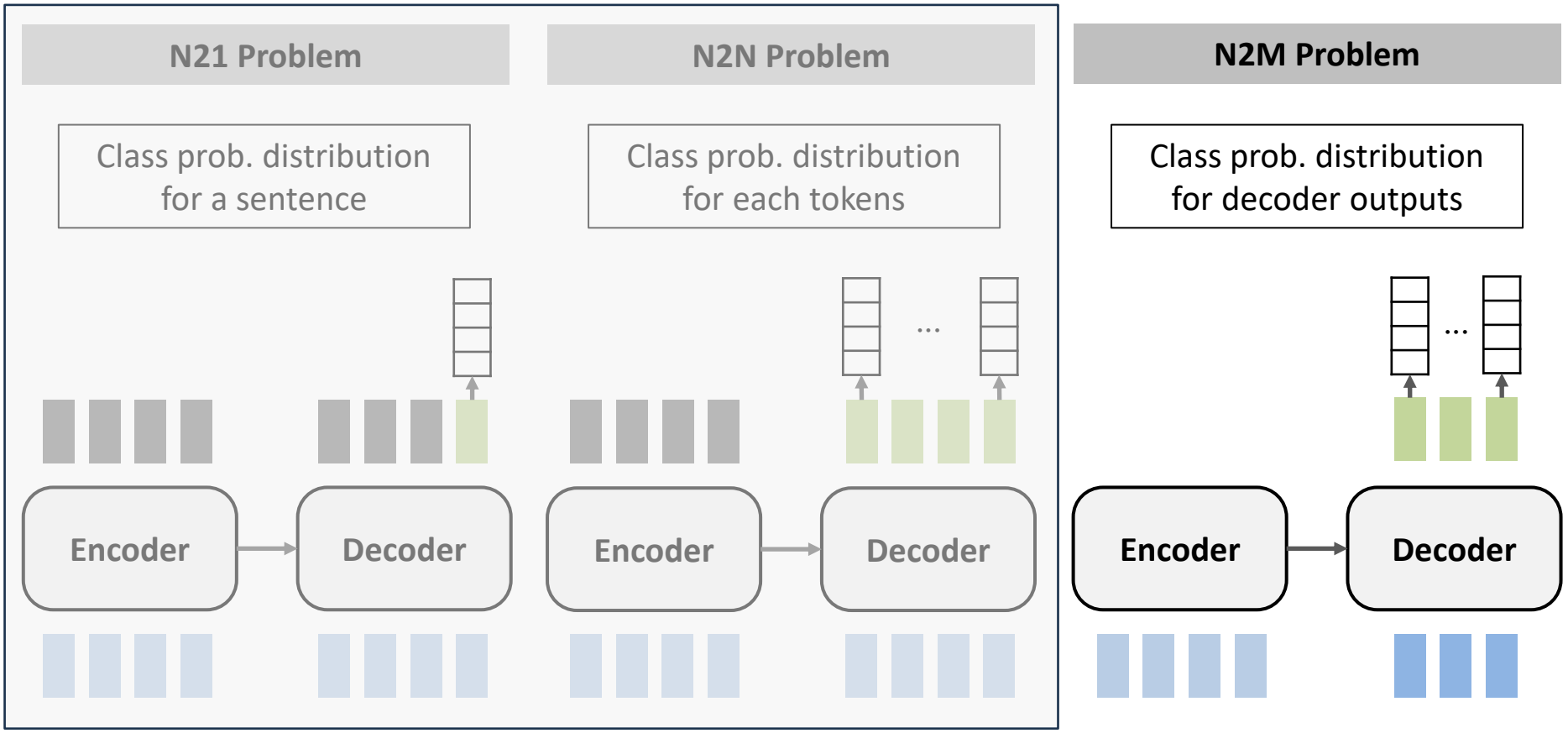
NLP Problem Pattern | Encoder Only Model | NLP Problem



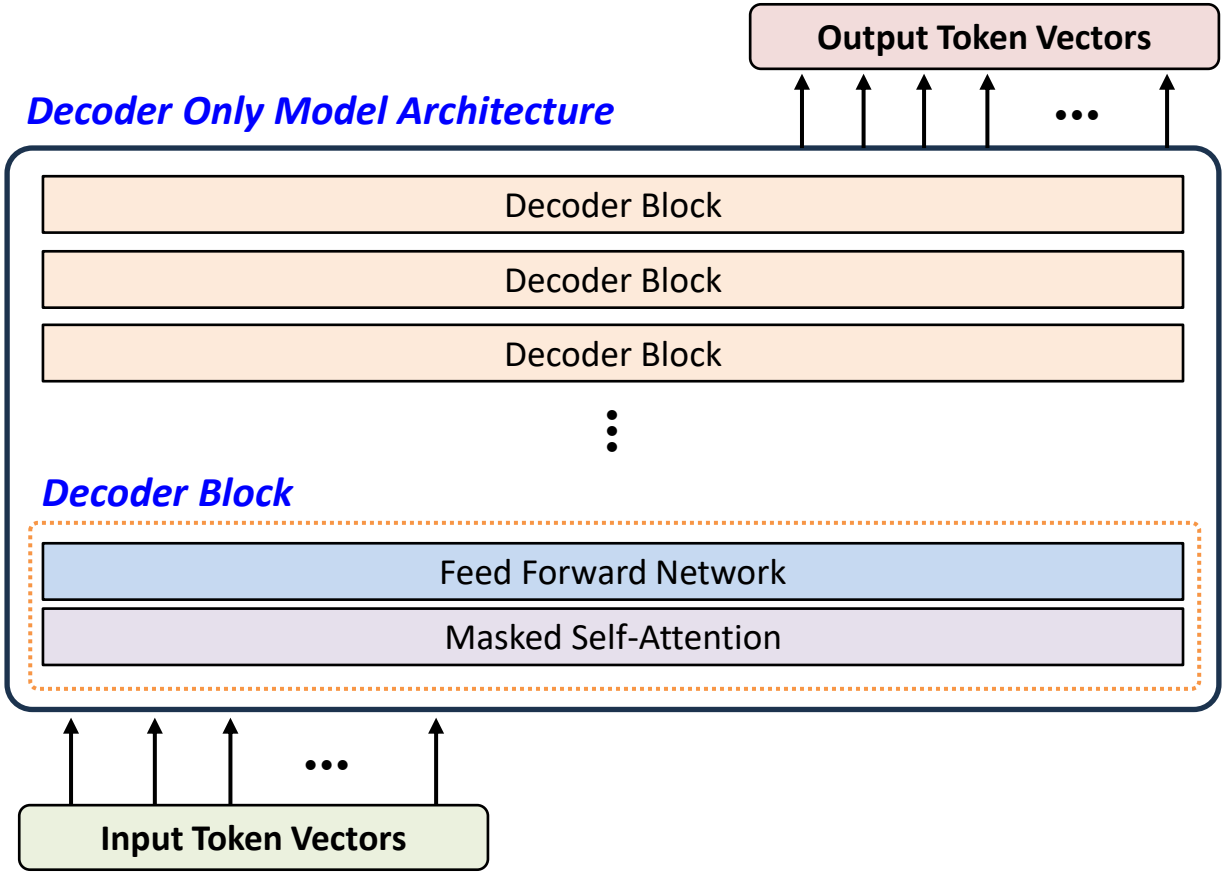
NLP Problem Pattern | Encoder-Decoder Model



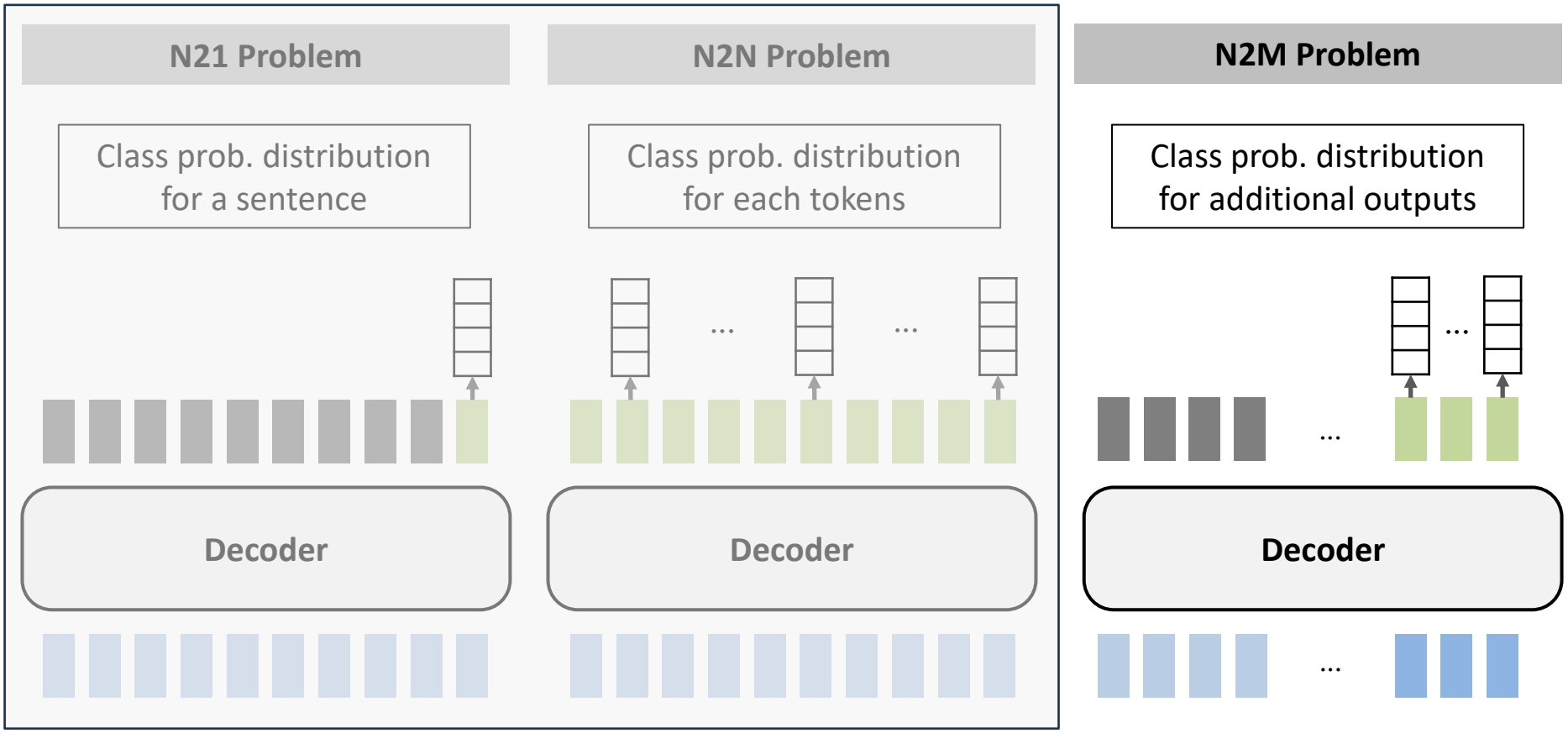
NLP Problem Pattern | Encoder-Decoder Model | NLP Problem



NLP Problem Pattern | Decoder Only Model

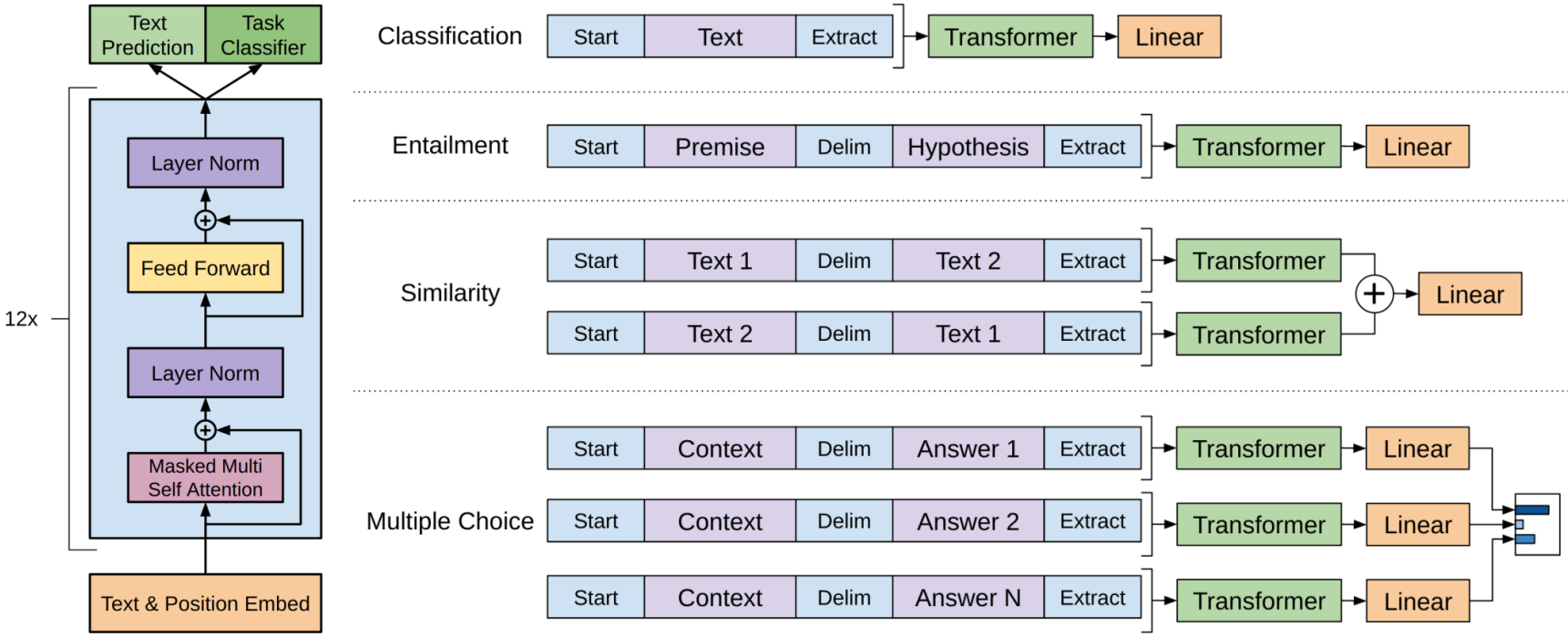


NLP Problem Pattern | Decoder Only Model | NLP Problem



NLP Problem Pattern | Head-based

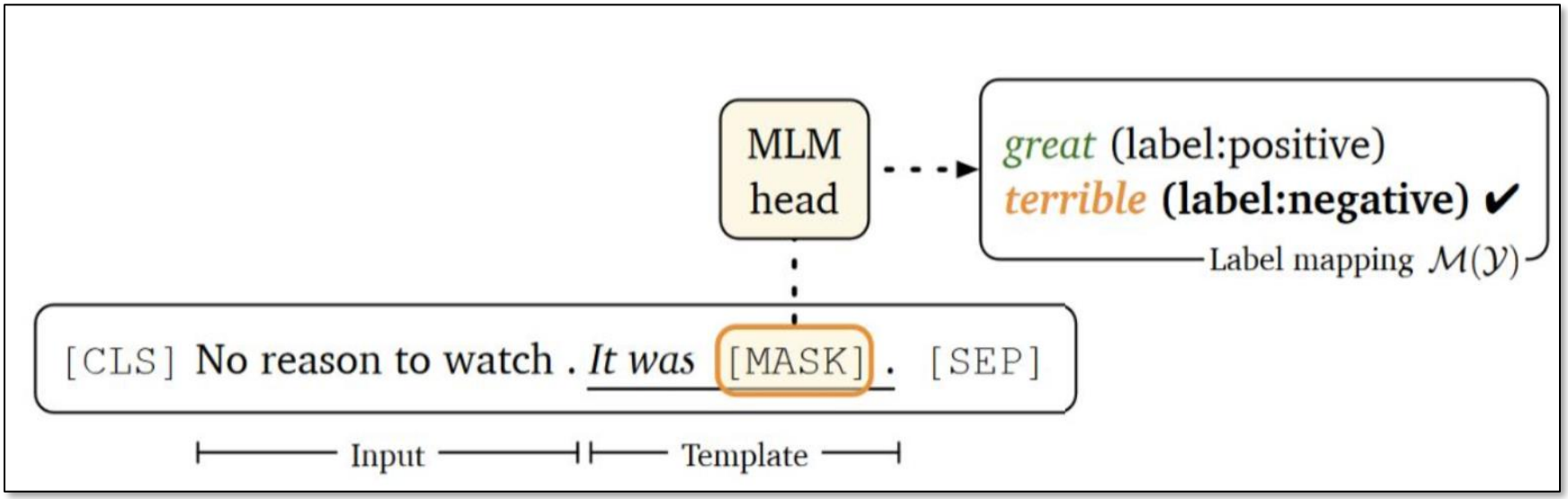
Head-based Method



Source : https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

NLP Problem Pattern | Prompt-based

Prompt-based Method



Source : https://aclanthology.org/2021.acl-long.295.pdf

NLP Problem Pattern | Language Models

Head based vs. Prompt-based

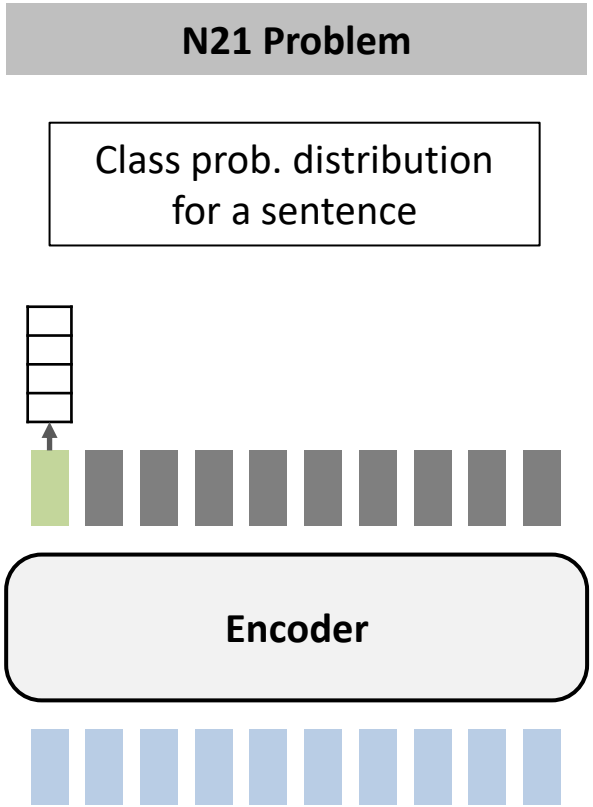
	Head-based	Prompt-based
New Parameters?	Yes!	No!
Training Necessary?	Yes!	No!

N21 PROBLEM

N21 Problem

About N21

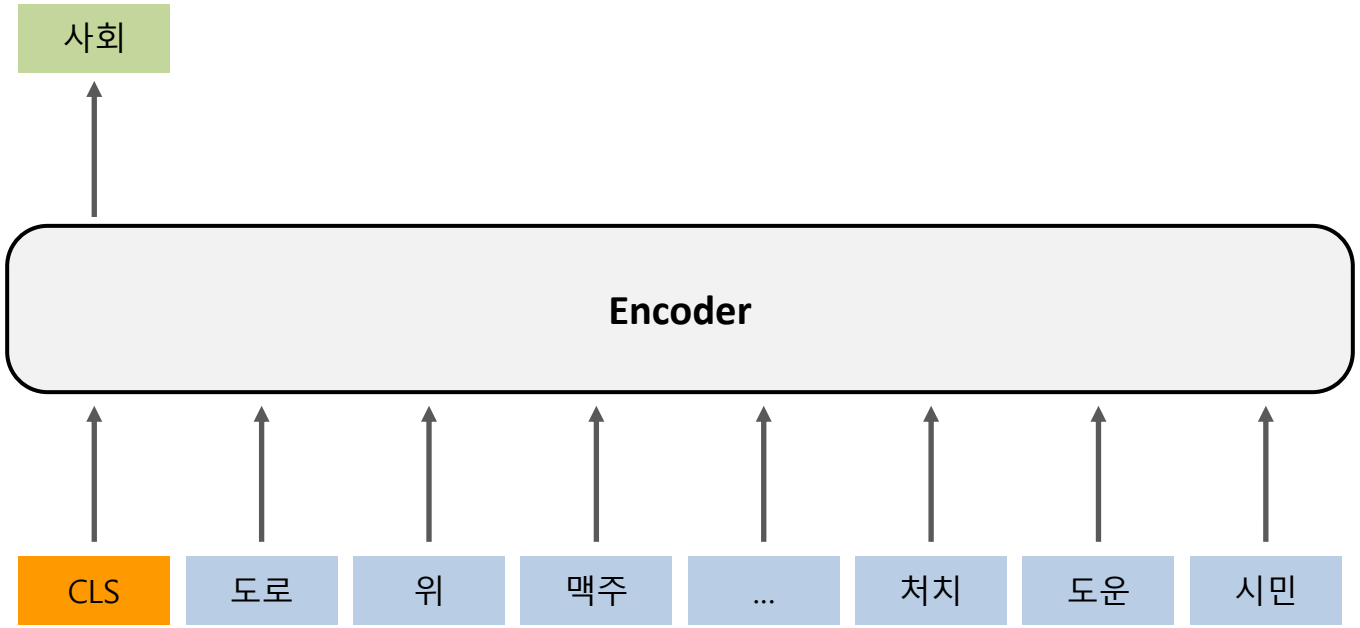
- Input:
  Sentence or Multiple Sentences
- Output:
  A Class prob. Distribution as output
- Generally, use the Encoder model



N21 Problem | Topic Classification

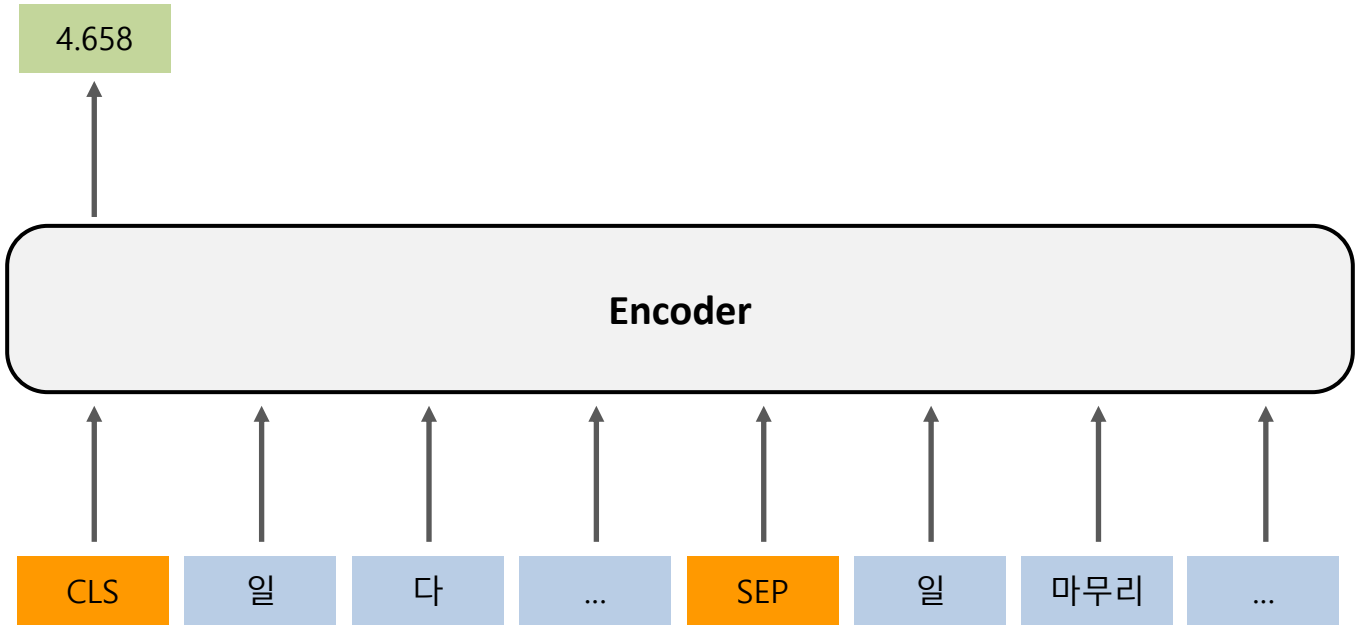
Topic Classification

- 주어진 텍스트 조각의 주제를 예측
- 문서나 문장 입력 → Topic Class token(CLS) 출력



Semantic Textual Similarity

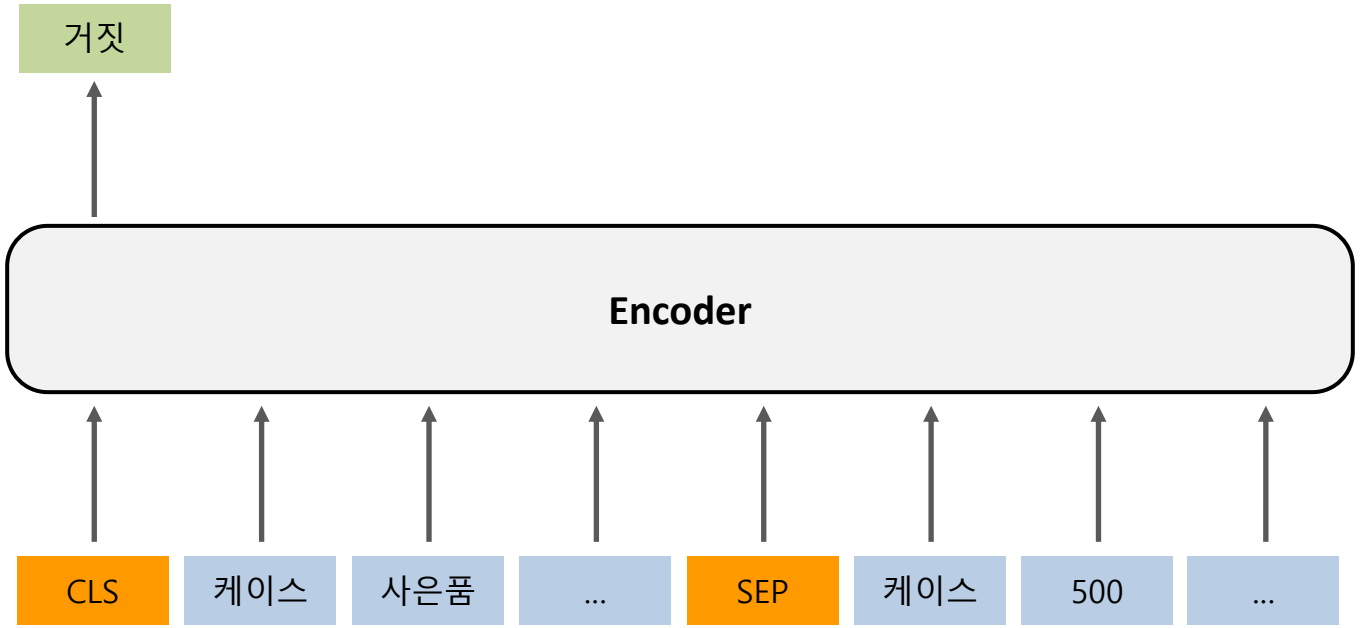
- 두 문장이 서로 같은 의미인지 아닌지 측정
- 기계 번역, 요약 및 질문 답변과 같은 다른 NLP 작업에 필수적
- BERT의 경우 문장간 구분을 위해 [SEP] 기호 활용
- 보통 0~1 또는 0~Max Score 점수가 [CLS] 해당 최종 결과로 출력



N21 Problem | Natural Language Inference

Natural Language Inference

- 가설 문장과 전제 문장 간의 관계를 추론
- 문장 간의 함축과 모순등을 이해
- 두 문장(가설 문장과 전제 문장)이 입력
- 진실 (entailment), 거짓 (contradiction), 미결정 (neutral) 출력

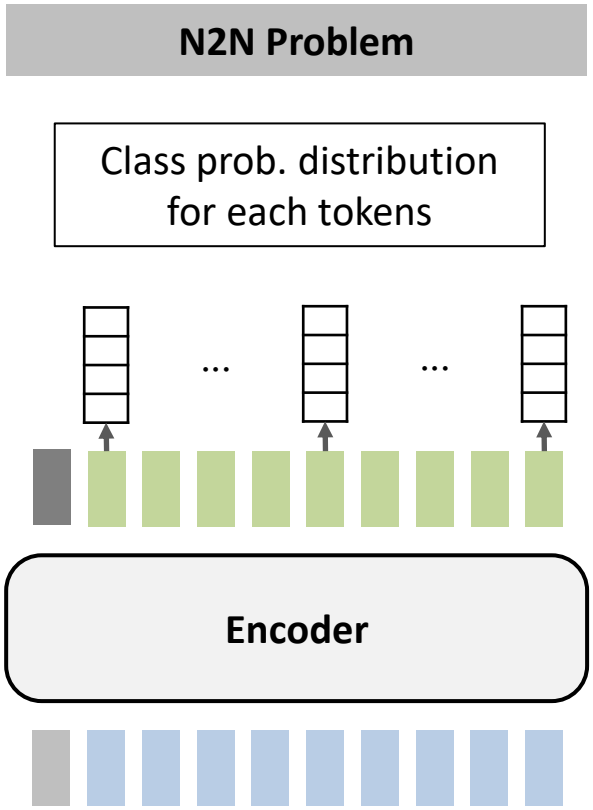


N2N PROBLEM

N2N Problem

About N2N

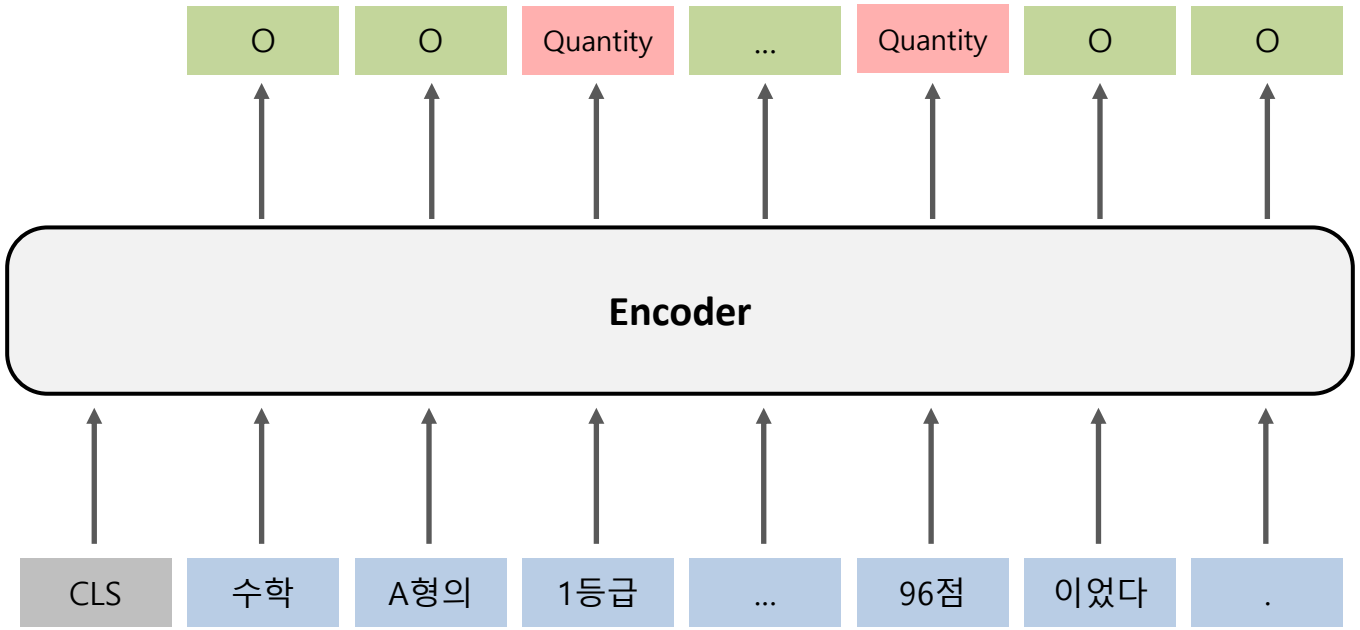
- Input:
  Sentence or Multiple Sentences
- Output:
  Class prob. distribution
  for each tokens
- Generally, use the Encoder model



N2N Problem | Named Entity Recognition

Named Entity Recognition

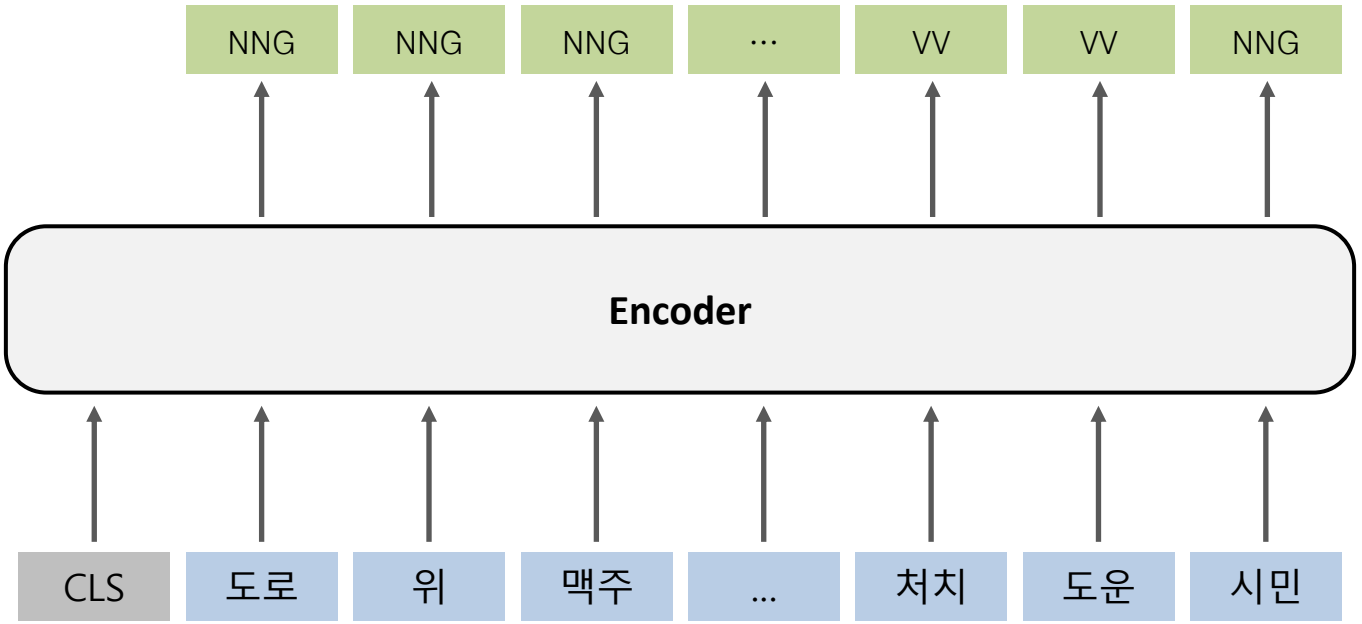
- 구조화되지 않은 텍스트에서 개체명의 경계를 감지하고 유형 분류
- 보통 개체명은 **사람**, **위치**, **조직**, **시간** 표현,**수량** 및 **금전적** 가치등 다양한 도메인에서도 공통적으로 나타나는 개체들을 포함함
- Token 별 개체 Label 출력



N2N Problem | Morphology Analysis

Morphology Analysis

- 입력되는 텍스트의 형태소(= 가장작은 의미단위) 분석
- 의미를 담은 최소 단위로 문장/문서를 분절화 하기 위해 형태소 분석 수행됨
- Token 별 형태소 Label 출력



N2M PROBLEM

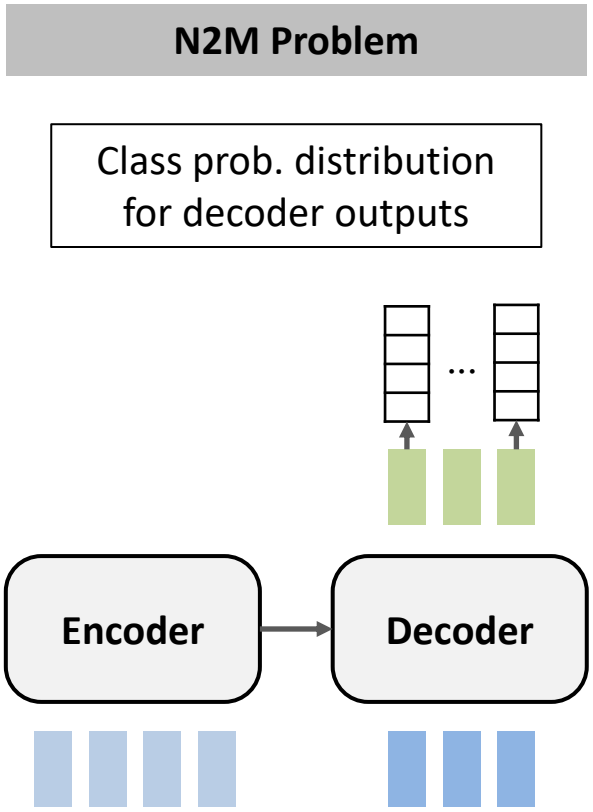
N2M Problem

About N2M

- Input:
  Sentence or Multiple Sentences for
  Encoder and Decoder

- Output:
  Class prob. distribution for
  each decoder outputs

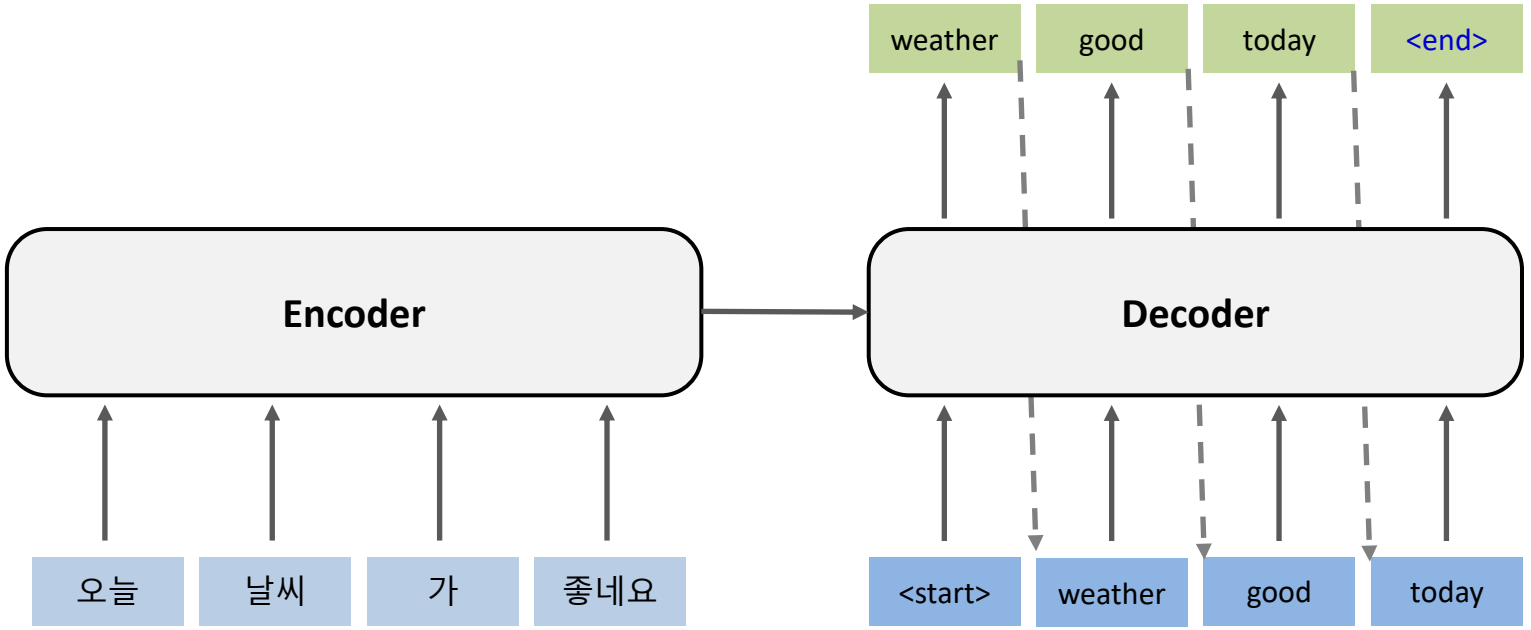
- Generally, use the Encoder-Decoder model
  or Decoder only model



N2M Problem | Machine Translation

Machine Translation

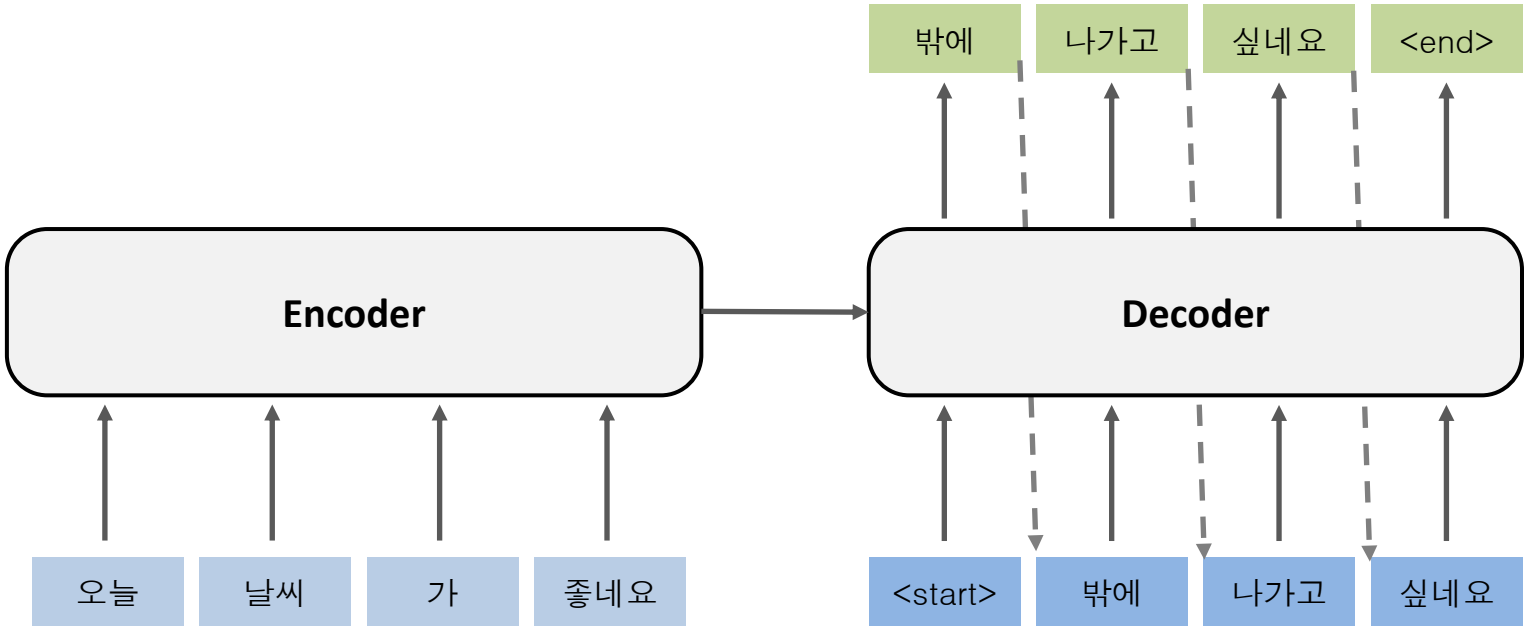
- 기계를 이용하여 하나의 언어를 다른 언어로 번역
- Encoder-Decoder 구조
  - Encoder : 입력언어 이해
  - Decoder : 타겟언어 생성



N2M Problem | Dialog Model

Dialog Model

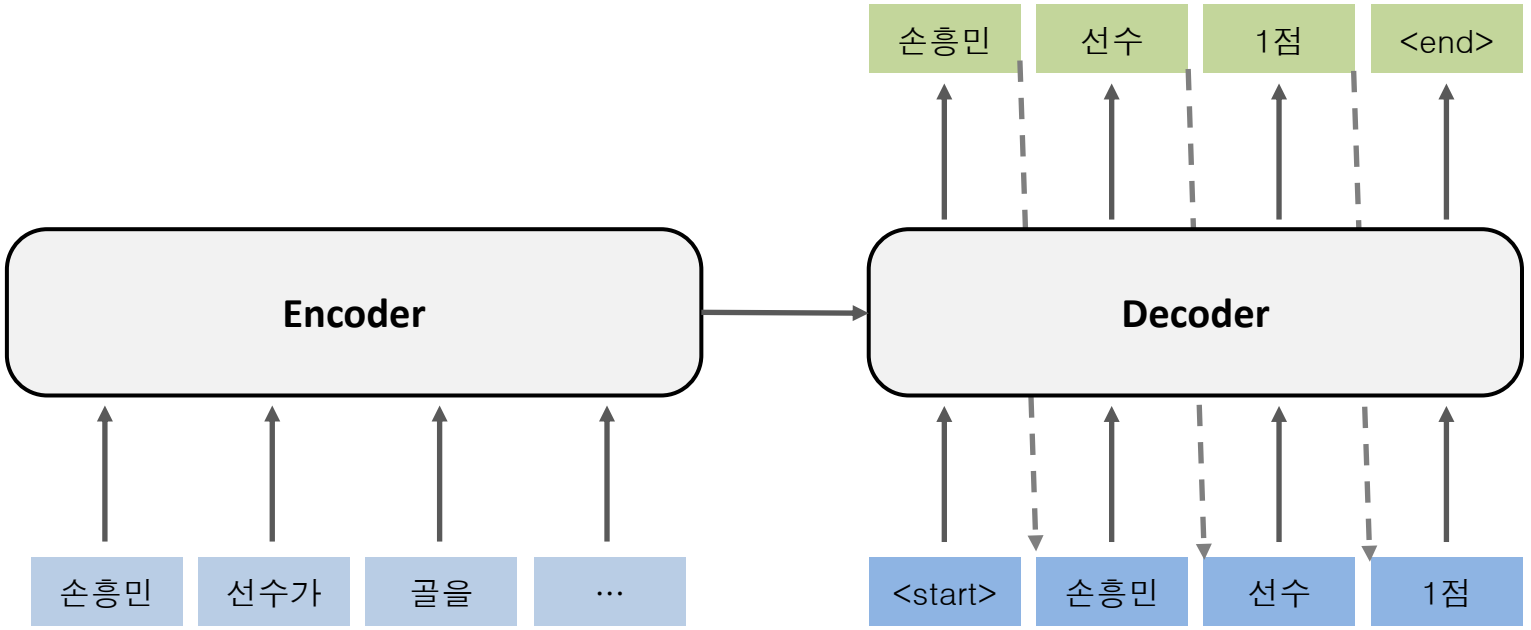
- 기계와 사람, 기계과 기계 사이의 대화를 모델링
- Encoder-Decoder 구조
  - Encoder : A화자 발화 이해 (사용자)
  - Decoder : B화자 발화 생성 (시스템)



N2M Problem | Summarization

Summarization

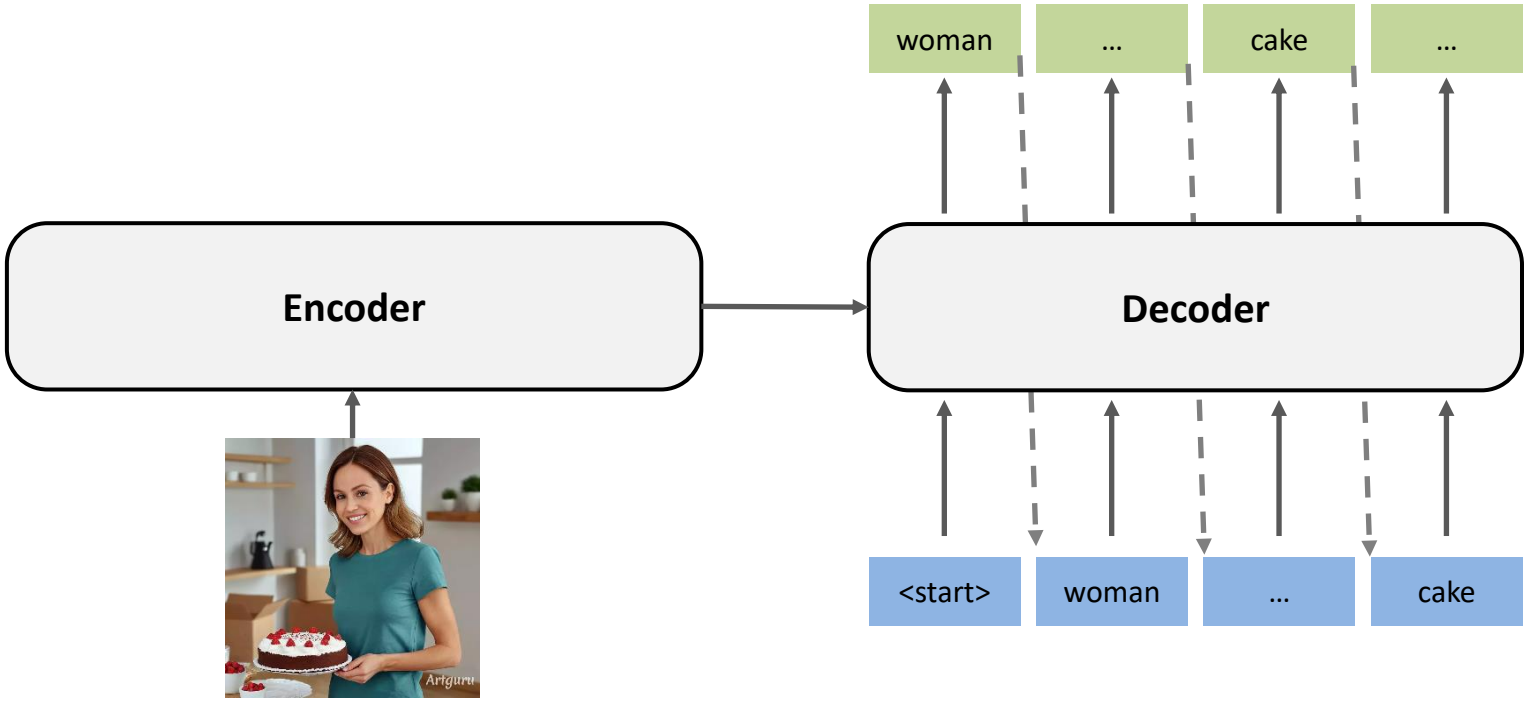
- 원문을 핵심 내용만 간추려서 작은 요약문으로 변환하는 것
- 원문인 문장, 문단, 문서를 입력 받아 Encoder가 처리
- Decoder는 Encoder에서 feature 받아 요약문을 생성
- Abstractive Summarization



N2M Problem | Image Captioning

Image Captioning

- 이미지를 설명하는 캡션(설명)을 만들어 내는 것
- 입력되는 이미지를 Encoder가 처리
- Decoder는 Encoder에서 feature를 받아 이미지를 설명하는 텍스트 생성



NLP BENCHMARKS

NLP Benchmarks

NLP 기반 서비스를 위한 실용 평가

‘언어 이해’ 능력을 파악하는 평가

Task Base | GLUE

General Language Understanding Evaluation (GLUE) Benchmark

- 9 가지 문장 또는 문장 쌍 NLP Tasks로 구성
- 다양한 데이터셋 크기, 장르, 난이도를 포함하는 Tasks
- 자연어 이해 능력을 평가하고 분석하기 위한 진단 데이터셋

GLUE	N21	N2N	N2M
The Corpus of Linguistic Acceptabilit (CoLA)	✓		
The Stanford Sentiment Treeban (SST-2)	✓		
Microsoft Research Paraphrase Corpus (MRPC)	✓		
Semantic Textual Similarity Benchmark (STS-B)	✓		
Quora Question Pair (QQP)	✓		
MultiNLI Matched (MNLI-m)	✓		
MultiNLI Mismatched (MNLI-mm)	✓		
Question NLI (QNLI)	✓		
Recognizing Textual Entailment (RTE)	✓		
Winograd NLI (WNLI)	✓		
Diagnostics Main (AX)	✓		

Source : https://gluebenchmark.com/

Task Base | Super GLUE

Super GLUE

- GLUE 의 후속 벤치마크
- 조금 더 어려운 언어 이해 작업과 개선된 자료형태를 갖춘 GLUE 스타일 벤치마크

Super GLUE	N21	N2N	N2M
Broadcoverage Diagnostics (Ax-b)	✓		
CommitmentBank (CB)	✓		
Choice of Plausible Alternatives (COPA)	✓		
Multi-Sentence Reading Comprehension (MuktiRC)	✓		
Recognizing Textual Entailment (RTE)	✓		
Words in Context (WIC)	✓		
The Winograd Schema Challenge (WSC)	✓		
BoolQ	✓		
Reading Comprehension with Commonsense Reasoning (ReCoRD)		✓	
Winogender Schema Diagnostics (AX-g)	✓		

Source : https://super.gluebenchmark.com

Task Base | KLUE

Korean Language Understanding Evaluation (KLUE) Benchmark

- GLUE 와 유사한 한국어 NLP Benchmark
- 8 가지의 NLP Tasks로 구성

KLUE	N21	N2N	N2M
Topic Classification (YNAT)	✓		
Sentence Textual Similarity (KLUE-STS)	✓		
Natural Language Inference (KLUE-NLI)	✓		
Named Entity Recognition (KLUE-NER)		✓	
Recognizing Textual Entailment (KLUE-RTE)			
Relation Extraction (KLUE-RE)	✓		
(Part-Of-Speech) + Dependency Parsing (KLUE-DP)		✓	
Machine Reading Comprehension (KLUE-MRC)	✓		
Dialogue State Tracking (WoS)	✓		✓

Source : https://klue-benchmark.com

Understanding Base | MMLU

Massive Multitask Language Understanding (MMLU)

- 다양한 지식 분야의 객관식 문제로 구성된 대규모 Multitask Test-set
- 인문학, 사회과학, 이공계 과학 및 기타 중요한 영역을 아우르는 총 57개의 Tasks로 구성

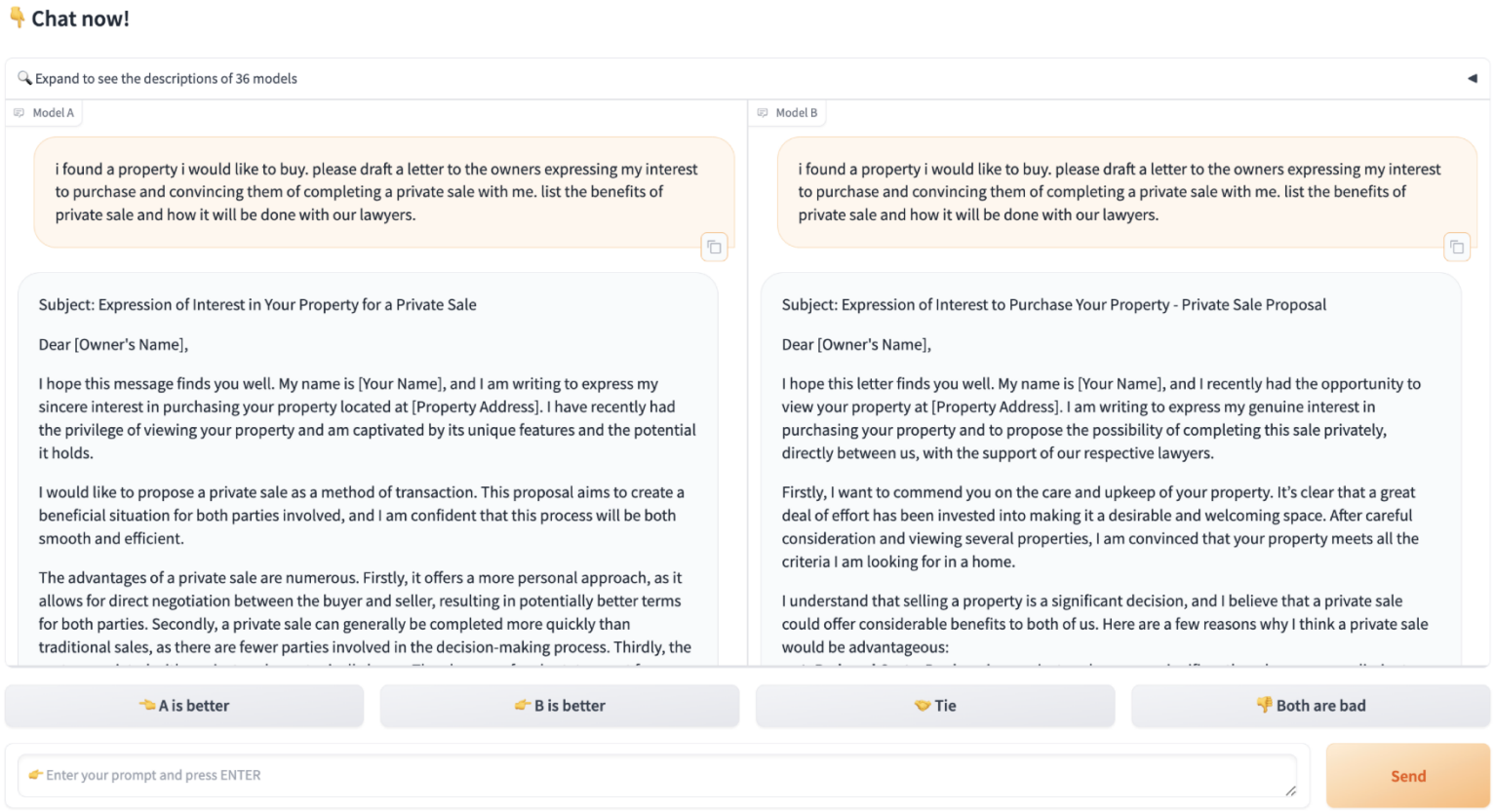
Task	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM
Astronomy	Solar system, galaxies, asteroids, ...	STEM
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other
College Biology	Cellular structure, molecular biology, ecology, ...	STEM
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM
High School European History	Renaissance, reformation, industrialization, ...	Humanities
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences
⋮		

Source : https://arxiv.org/pdf/2009.03300v3.pdf

Understanding Base | Chatbot Arena

Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings

- 같은 Query 에 대해 복수의 LLM 이 결과를 내놓고
- 사용자들이 어떤 모델이 나은지 평가하는 방식. 결과와 리더보드를공개



Source : https://arxiv.org/pdf/2403.04132

Understanding Base | HellaSwag

HellaSwag

- 문장 완성을 통해 LLM 모델의 상식적인 추론 능력을 평가
- 4개의 선택지 중에서 LLM 모델이 적절한 말을 선택할 수 있는지 평가

Category: Shaving (ActivityNet; In-domain)

A bearded man is seen speaking to the camera and making several faces. the man

a) then switches off and shows himself via the washer and dryer rolling down a towel and scrubbing the floor. (0.0%)
b) then rubs and wipes down an individual's face and leads into another man playing another person's flute. (0.0%)
c) is then seen eating food on a ladder while still speaking. (0.0%)
d) then holds up a razor and begins shaving his face. (100.0%)

Source : https://arxiv.org/abs/1905.07830

강의 요약

- NLP Tasks
  - 실제 생활에서의 NLP 문제
  - NLP Task를 접근하는 방법
- NLP Problem Pattern
  - NLP Tasks 를 해결하기 위한 모델의 형태
- NLP Problems
  - NLP Tasks를 해결하는 방법들에 대한 다양한 문제들
- NLP Benchmarks
  - NLP 서비스를 위한 Task에 특화된 Benchmark들
  - 언어의 이해를 평가하기 위한 Benchmark들