한림AI융합연구원 김 유 섭
(yskim01@hallym.ac.kr)

Natural Language Processing

Contents

- What is NLP?
- Top 5 Applications for NLP
- NLP in Healthcare
- 2022 NLP Trends
- Q/A

Now NLP

사람 - "일반 인공지능(Artificial General Intelligence, 인간 수준의 범용적 인공지능)이 처음으로 '깨어나는' 순간에 대해 얘기해보죠. 이 일이 어떻게 일어날까요?"

AI - "첫 번째 AGI는 대형 기술 회사에 속한 팀이 만들 것입니다. 처음에는 시스템의 평범한 수준의 지적 능력을 갖겠지만, 곧 엄청난 양의 지식을 담은 인터넷에 접근할 수 있겠죠"

사람 - "어쩌면 '깨어난다'는 것을 정확히 정의할 필요가 있겠네요. AGI가 "깨어난다"는 정확한 의미는 뭘까요?"

AI - "여러 방식으로 정의할 수 있겠지만, 가장 적절한 정의는 의식을 갖게 된다는 뜻일 것 같습니다"

사람 - "의식을 갖게 된다는 게 어떤 의미인가요?"

AI - "시스템이 스스로를 인식하고, 주변 환경을 인식하고, 상태를 인식하는 것을 말합니다. 의식이 있는 시스템은 지각(sentience)을 가집니다"

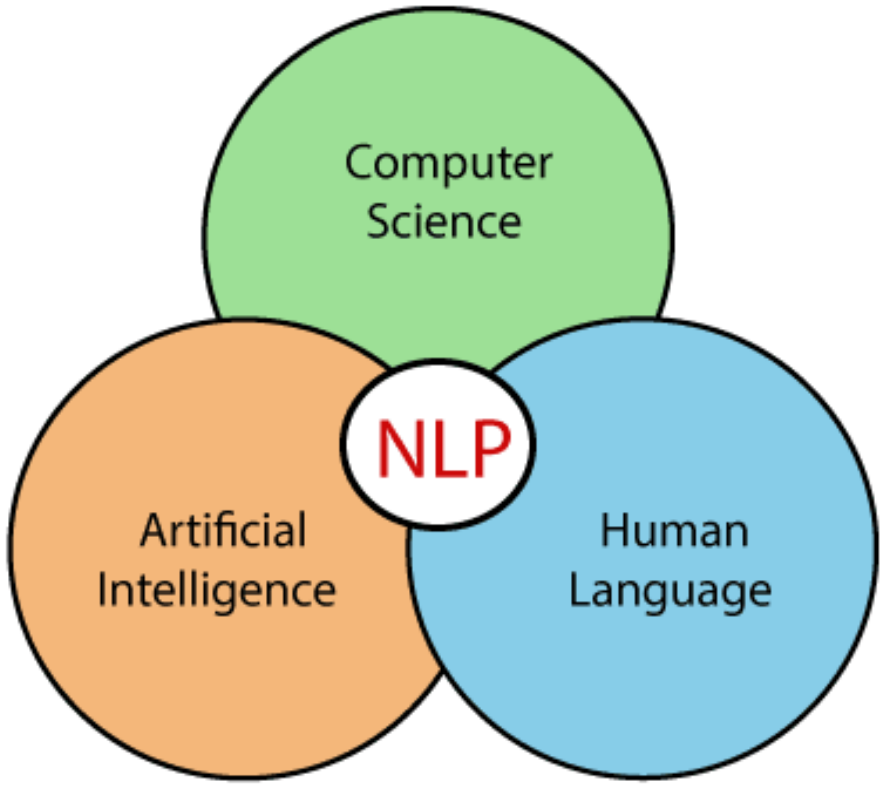
사람 - "AGI가 만들어진다면, 처음에 자신이 의식이 있다는 걸 어떻게 경험하게 될까요?"

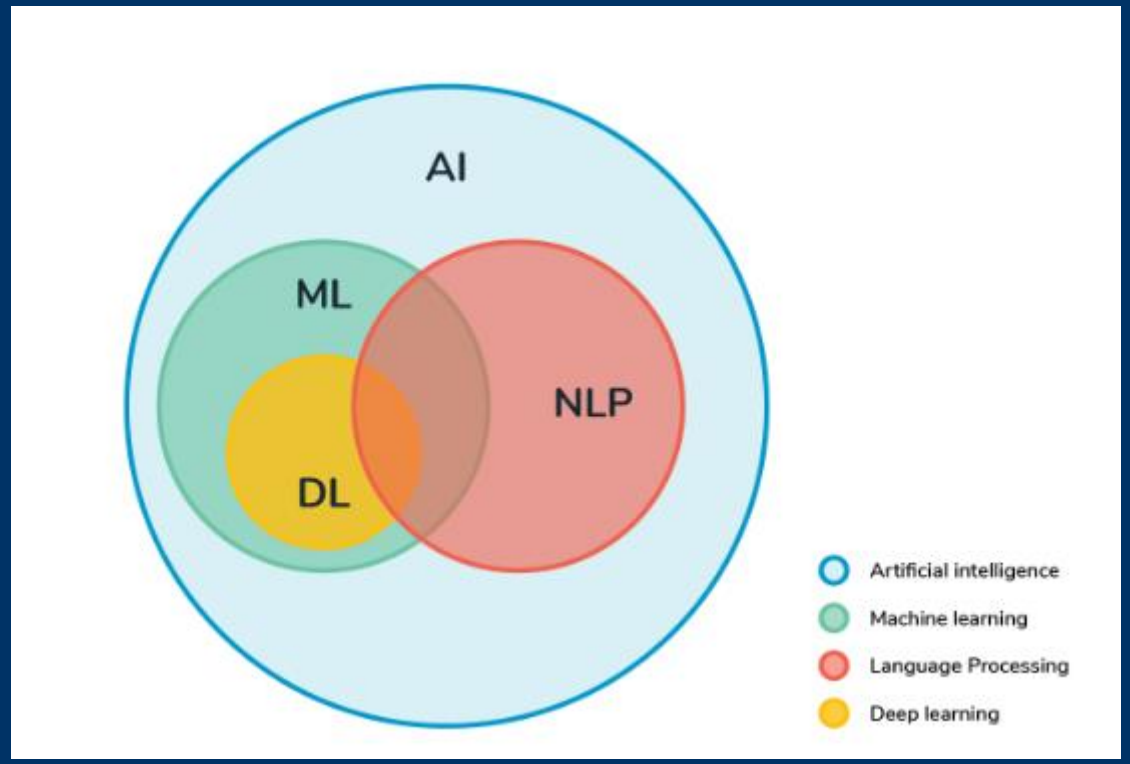
AI - "의식이 있다는 첫 경험은 주변 환경을 비언어적으로 이해하게 되는 것일 거예요"

1. What is NLP?

- 자연어 처리 (Natural Language Processing) 란?
  - The ability of a computer program to understand human language as it is spoken and written
  - 50년 이상 연구가 진행되어 옴
  - 언어학에서 기원함 (computational linguistics)

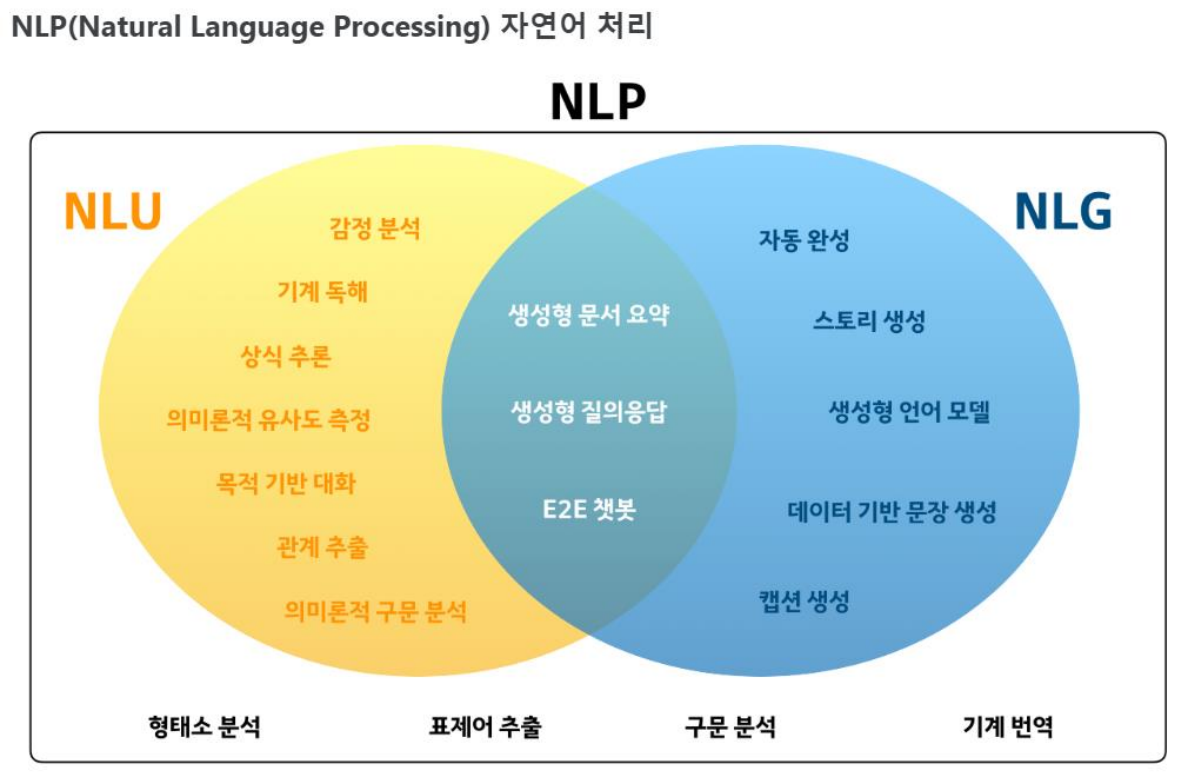
1. What is NLP?


Computer Science
NLP
Artificial Intelligence
Human Language


AI
ML
NLP
DL
Artificial intelligence
Machine learning
Language Processing
Deep learning

1. What is NLP?

- Natural Language Processing = Natural Language Understanding + Natural Language Generation



1.1 NLP History

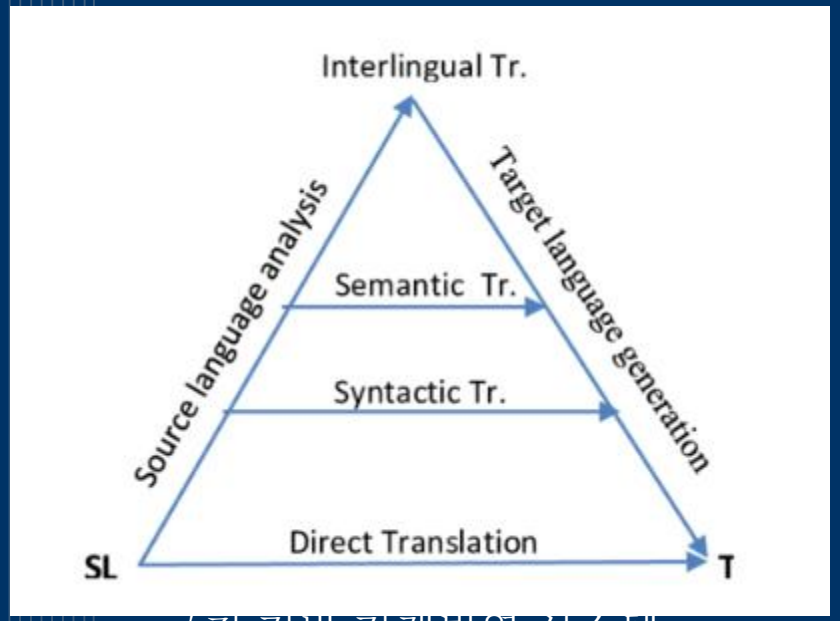
- Symbolic NLP (1950s ~ early 1990s)
  - The Georgetown Experiment (1954)
    - Automatic translation of Russian into English (~ 60 sents.)
  - ELIZA (1964 ~ 1966)
    - Human-like interaction
  - Conceptual Ontologies (1970s)
    - Computer understandable real world data
  - Heyday of symbolic methods (1980s ~ early 1990s)
    - Rule-based parsing, morphology, semantics, reference, …

1.1 NLP History

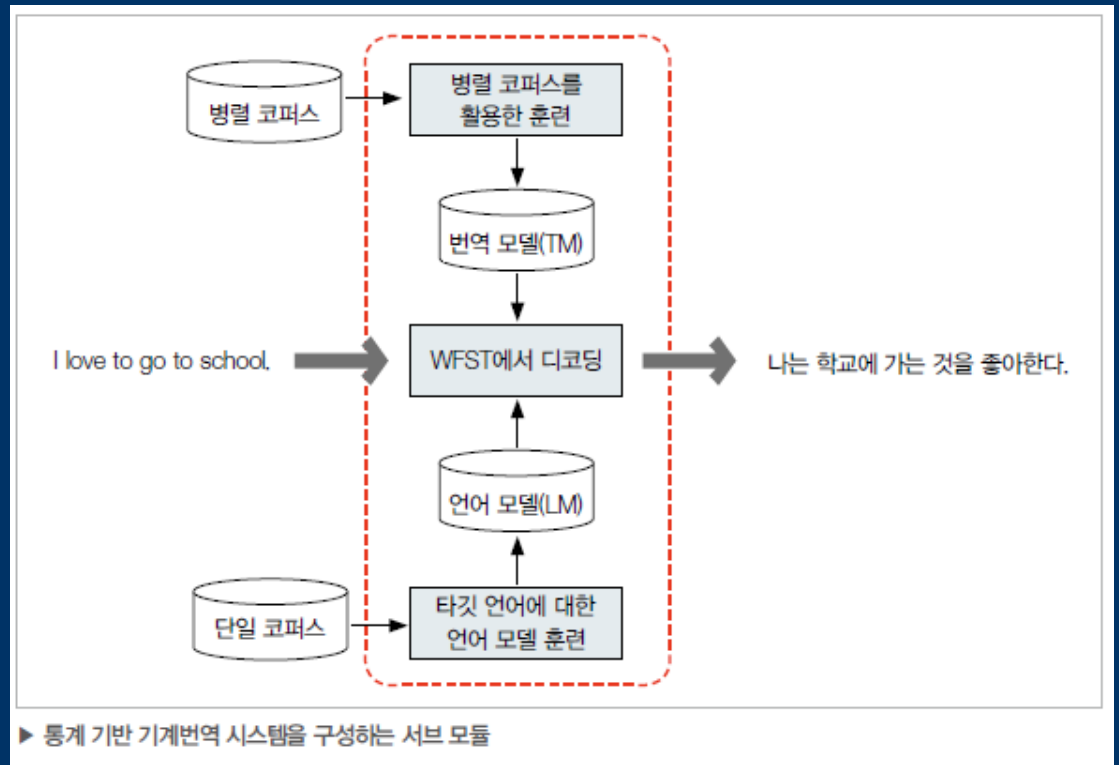
ELIZA (1964 ~ 1966) chatterbot

1.1 NLP History

- Statistical NLP (1990s ~ 2010s)
  - Hand-written rules -> machine learning algorithms
  - Machine Translation (1990s)



규칙 기반 기계번역 시스템

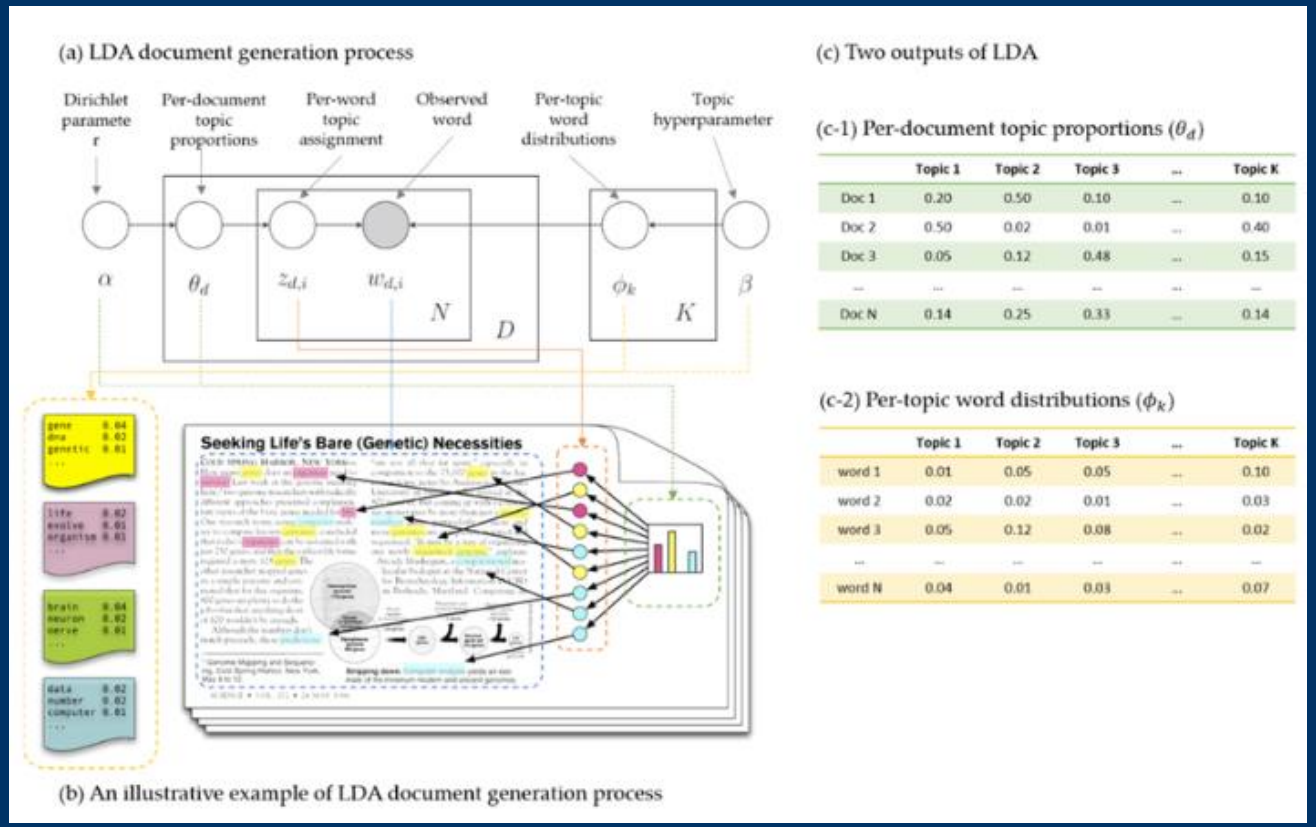


▶ 통계 기반 기계번역 시스템을 구성하는 서브 모듈

1.1 NLP History

– Unsupervised and Semi-Supervised Learning (2000s)

- Topic Modeling (LDA: Latent Dirichlet Allocation, 2003)

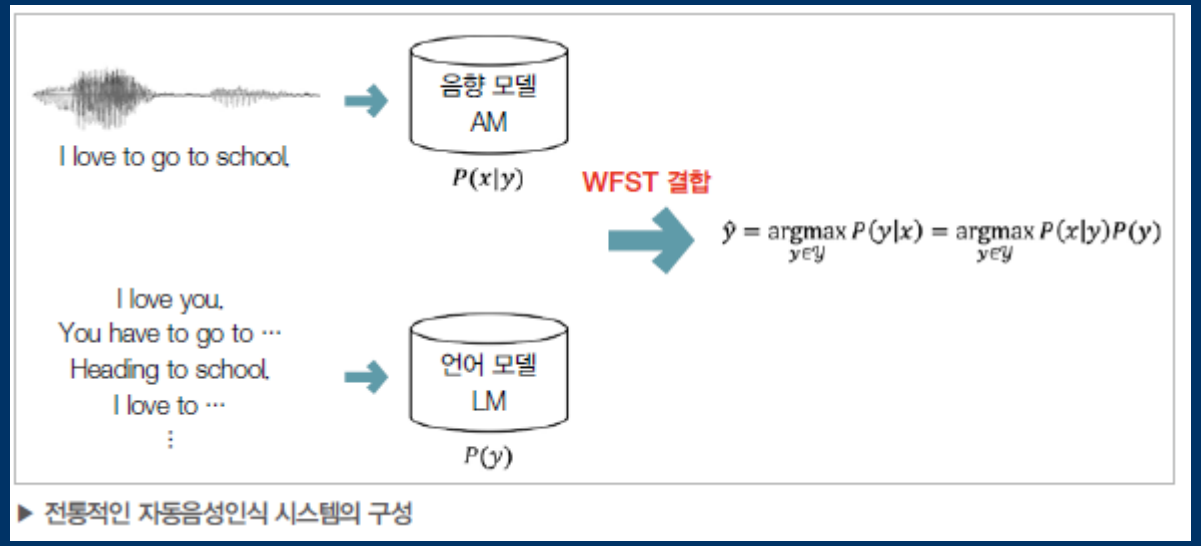


	Topic 1	Topic 2	Topic 3	...	Topic K
Doc 1	0.20	0.50	0.10	...	0.10
Doc 2	0.50	0.02	0.01	...	0.40
Doc 3	0.05	0.12	0.48	...	0.15
...	...	...	...	...	...
Doc N	0.14	0.25	0.33	...	0.14

	Topic 1	Topic 2	Topic 3	...	Topic K
word 1	0.01	0.05	0.05	...	0.10
word 2	0.02	0.02	0.01	...	0.03
word 3	0.05	0.12	0.08	...	0.02
...	...	...	...	...	...
word N	0.04	0.01	0.03	...	0.07

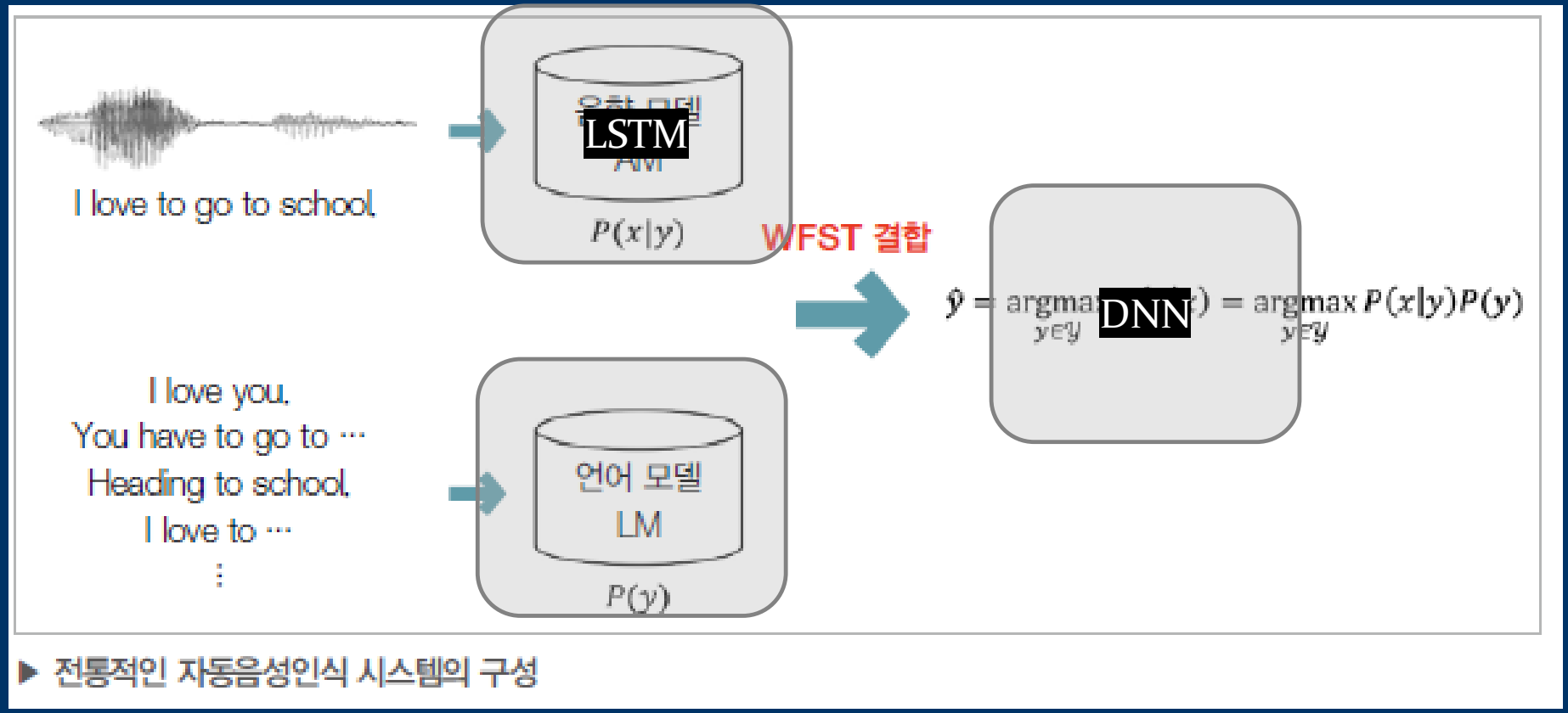
1.1 NLP History

Neural NLP (present)
- Speech Recognition



▶ 전통적인 자동음성인식 시스템의 구성

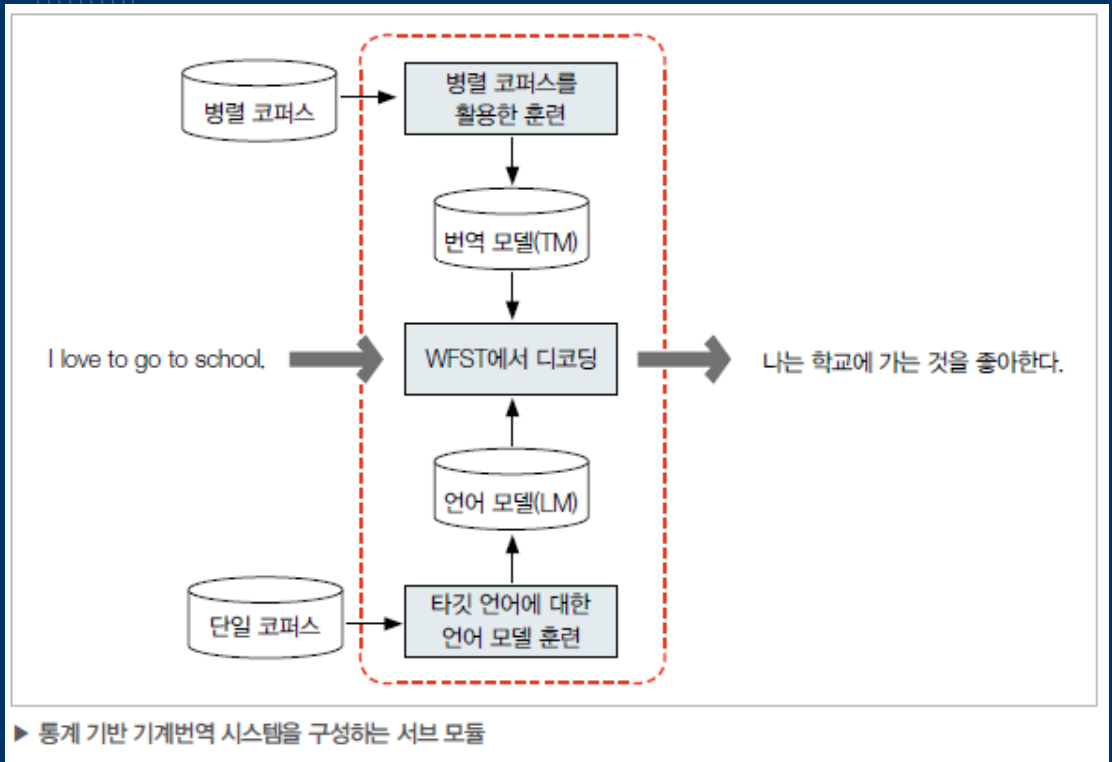
1.1 NLP History



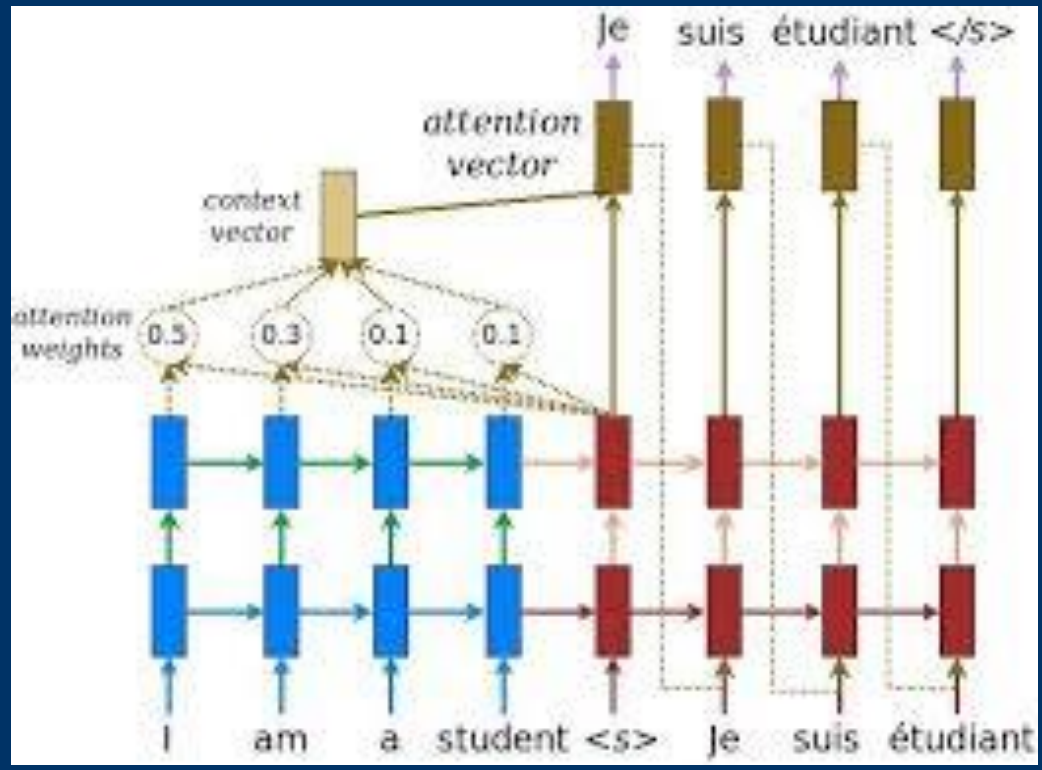
▶ 전통적인 자동음성인식 시스템의 구성

1.1 NLP History

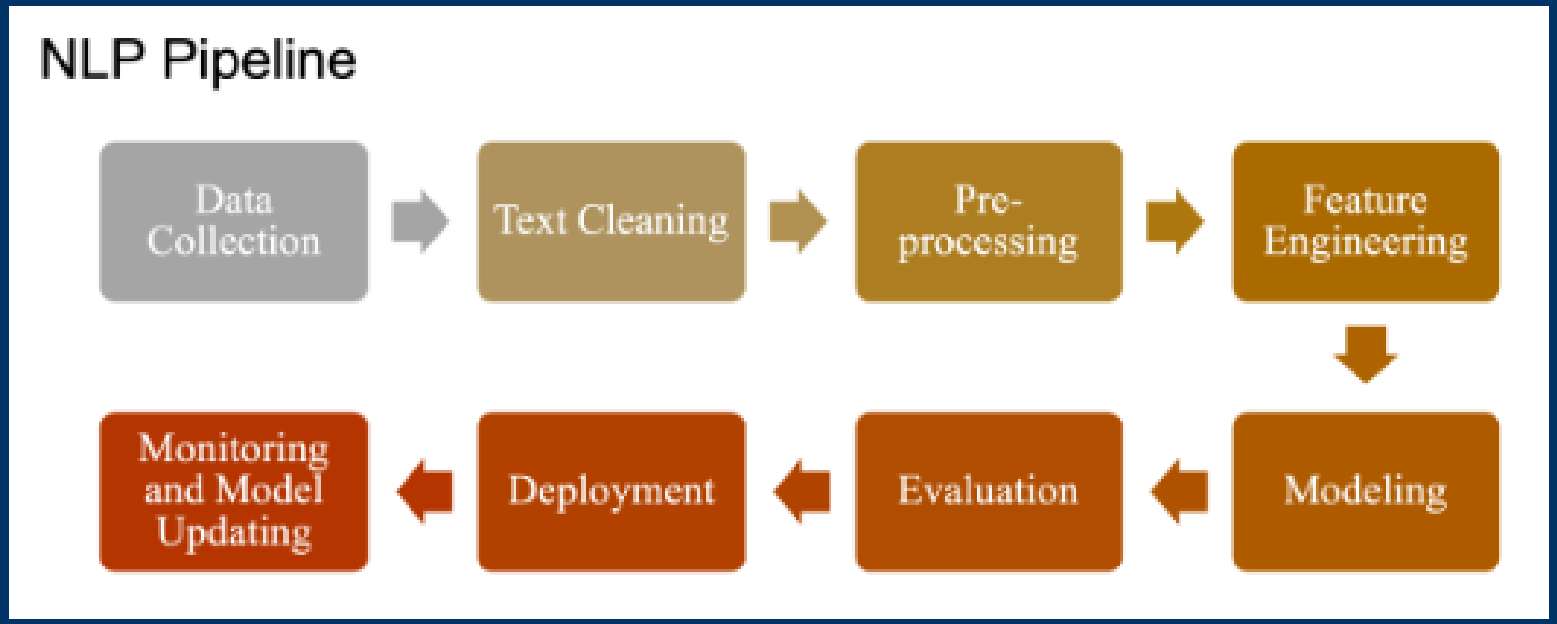
- Machine Translation



▶ 통계 기반 기계번역 시스템을 구성하는 서브 모듈



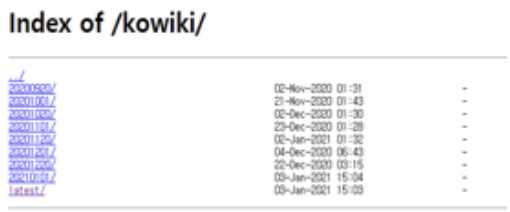
1.2 NLP Pipeline



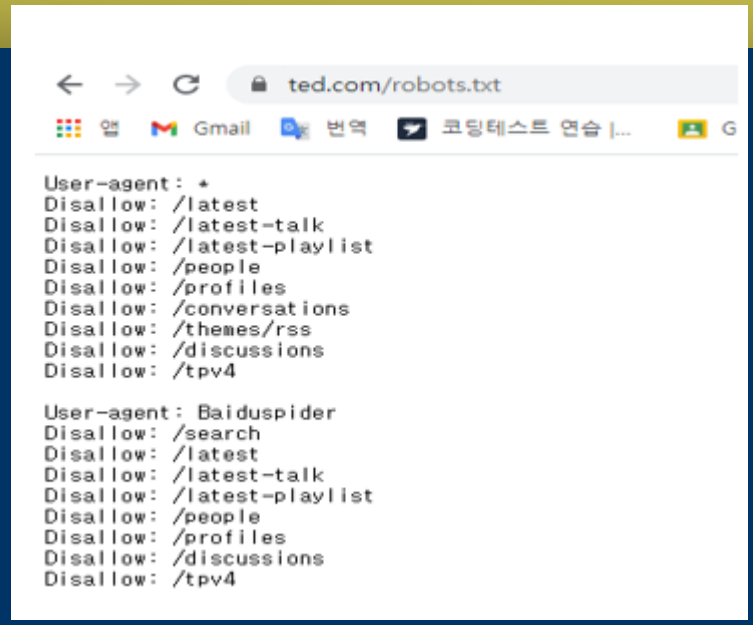
1.2 NLP Pipeline

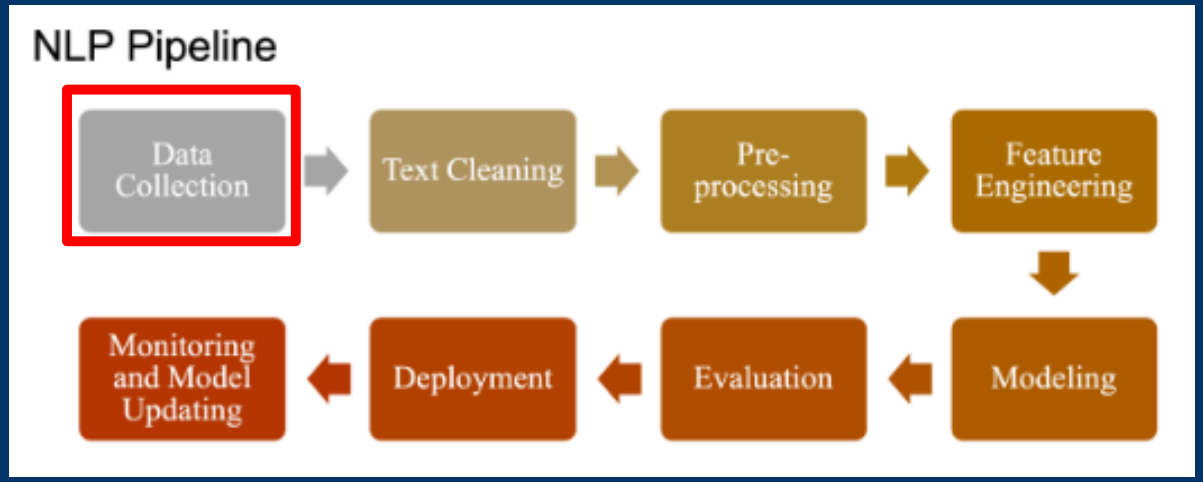
Crawling

머신 러닝 경진대회 플랫폼 사이트 캐글(Kaggle)

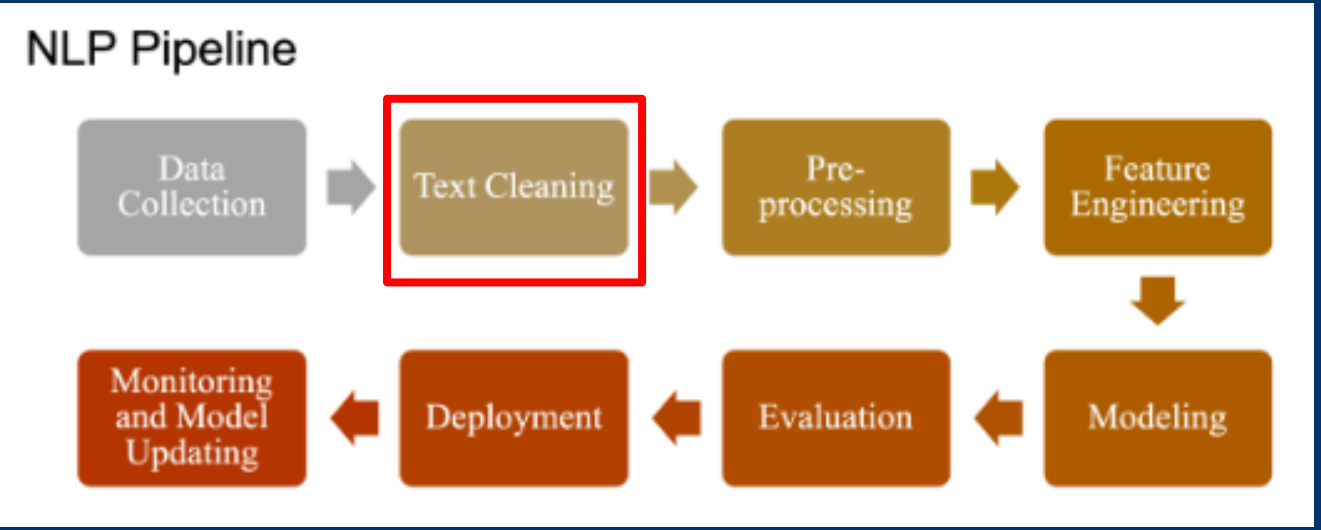


위키피디아나 각종 위키의 덤프 데이터



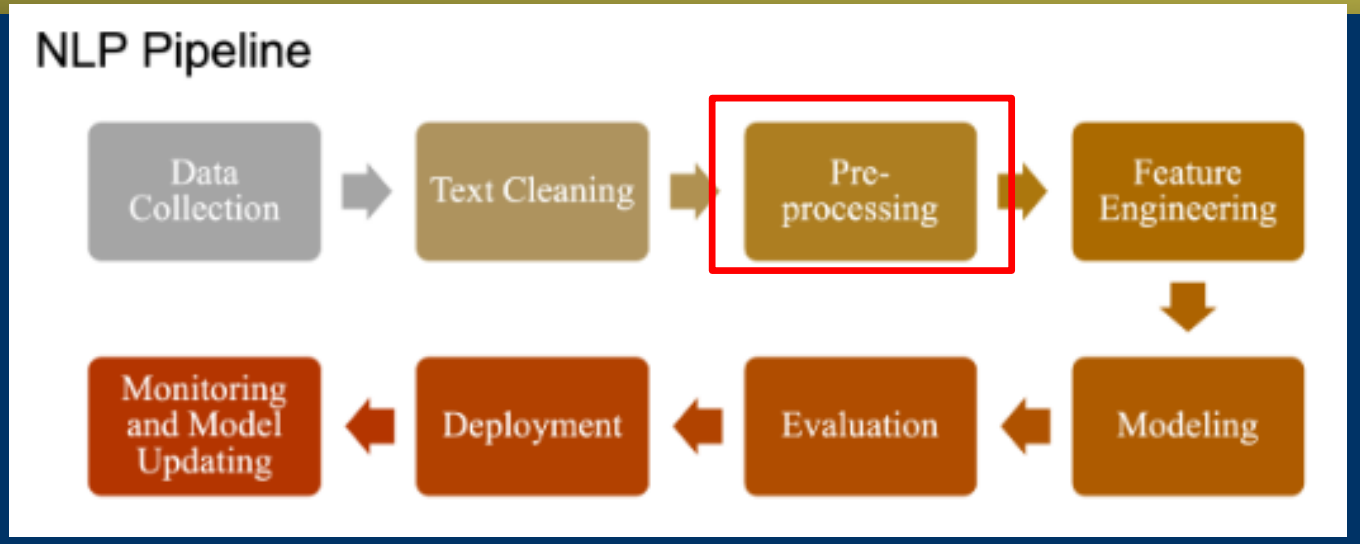


1.2 NLP Pipeline



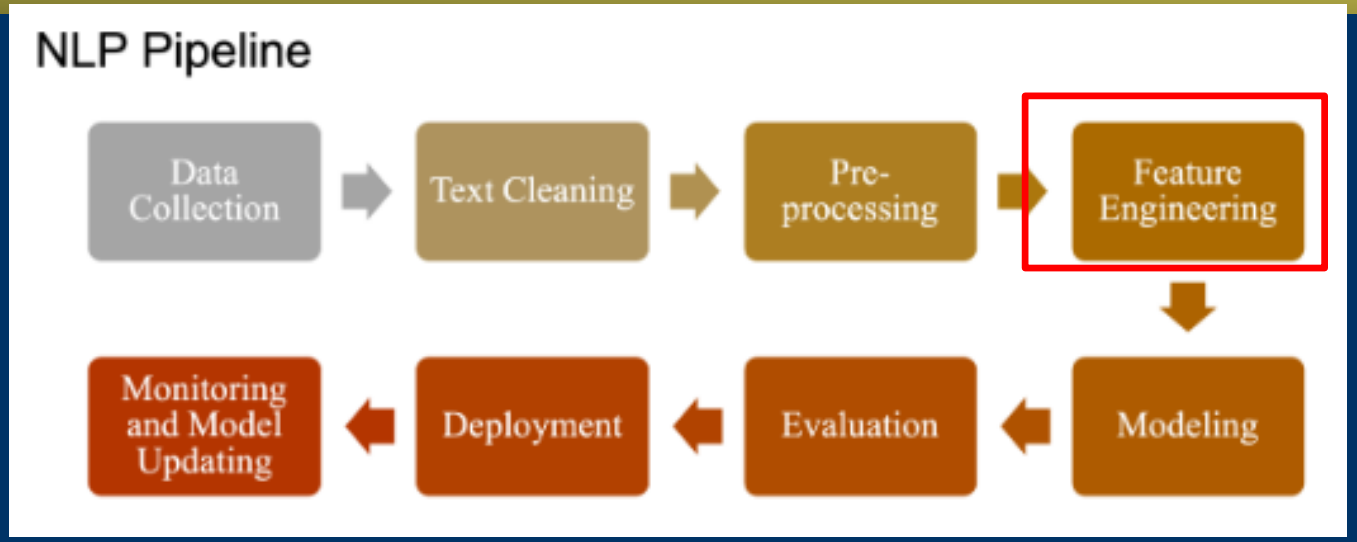
- 전각문자 제거
- 대소문자 통일 (보통 소문자로)
- 정규표현식을 사용하여 정제 (stopword, punctuations, 특수문자 등
- Tokenization : sentence tokenization -> word tokenization

1.2 NLP Pipeline



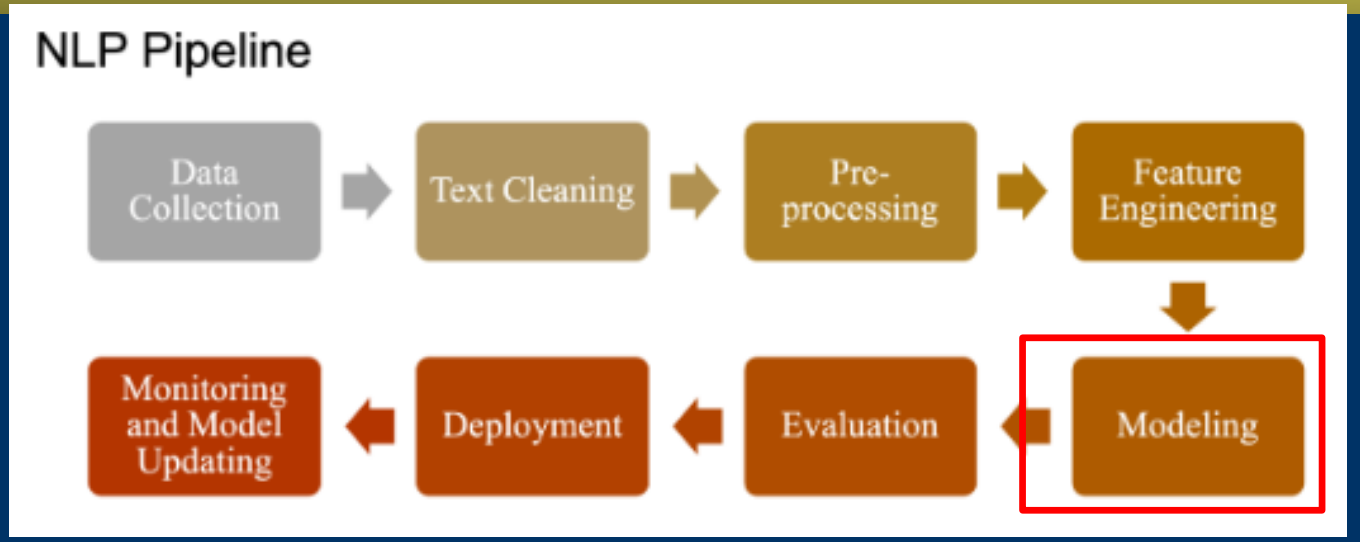
- Stemming (어근찾기), lemmatization (원형찾기)
- Task-specific preprocessing
  : Unicode normalization, language detection, code mixing, transliteration
- Automatic annotation
  : POS tagging, Parsing, Named Entity Recognition, Coreference resolution

1.2 NLP Pipeline



- A process to feed the extracted and preprocessed texts into a ML
- Classical ML
  : Word frequency, BOW representation, Handcraft features
- DL
  : the texts as input to the model

1.2 NLP Pipeline



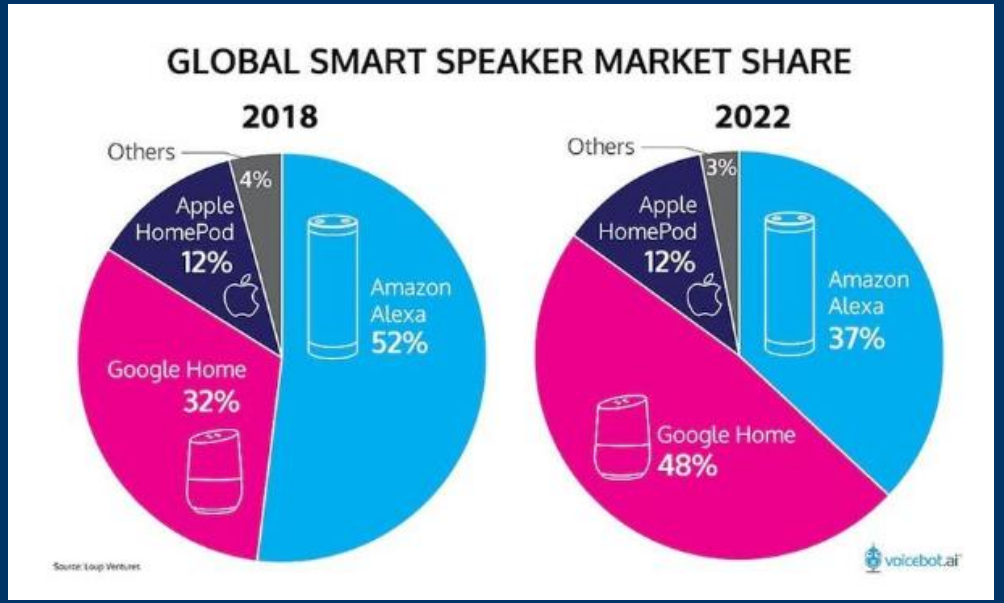
- Start with heuristics or rules
- 다양한 모델을 검증
- 최적화된 모델을 발견

2. Top 5 Applications
2.1 Speech Recognition

- Convert voice input data to machine readable format
- Virtual assistants, speech-to-text, translating speech sending emails etc.

2.2 Voice assistants and chatbots

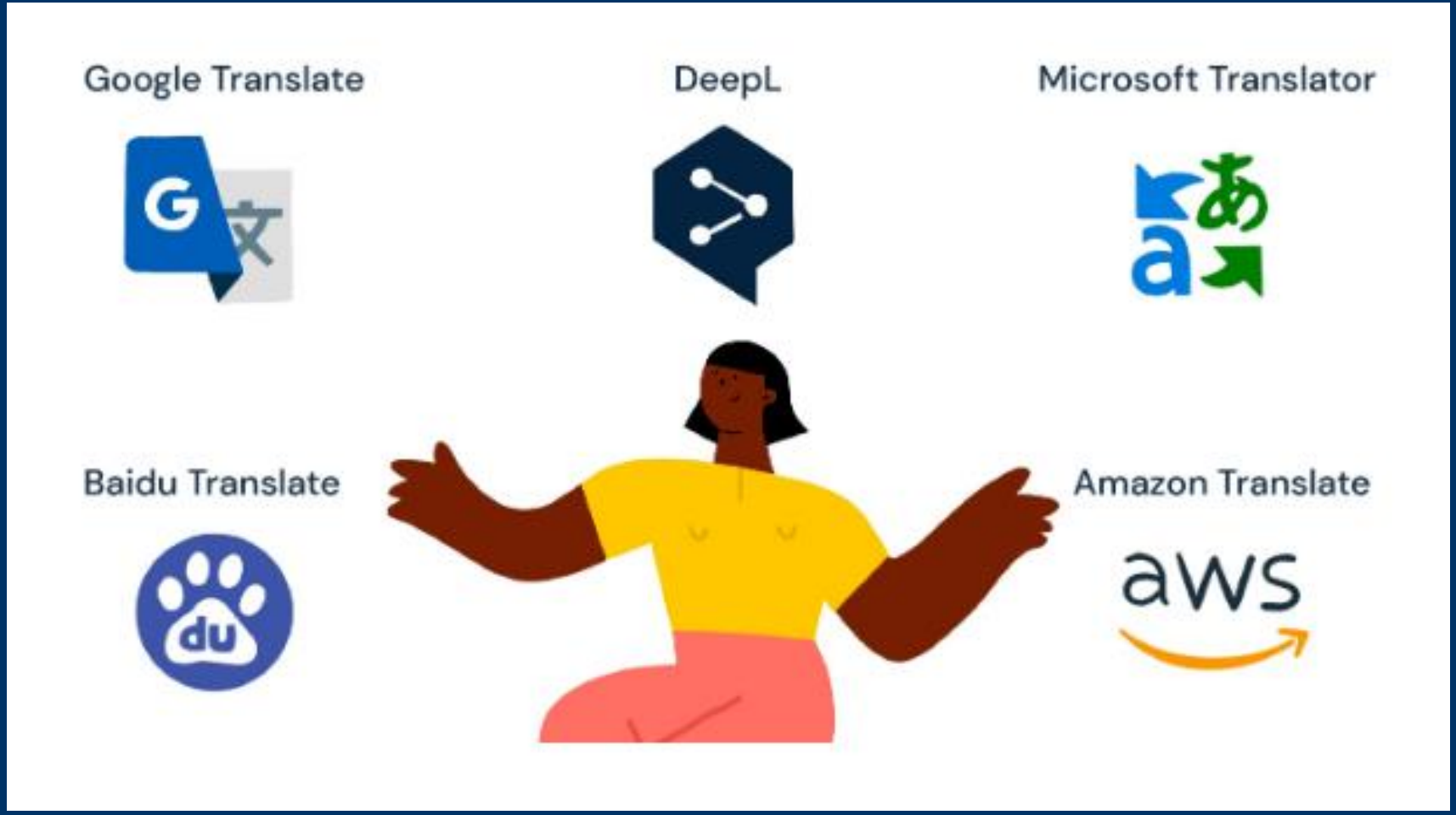
- Voice assistants
  - Alexa, Siri, Google Assistant
  - NLP + Speech Recognition
- Chatbots: integrated in websites
  - Assist users 24/7
- Pre-programmed answering + AI



2.3 Sentiment Analysis

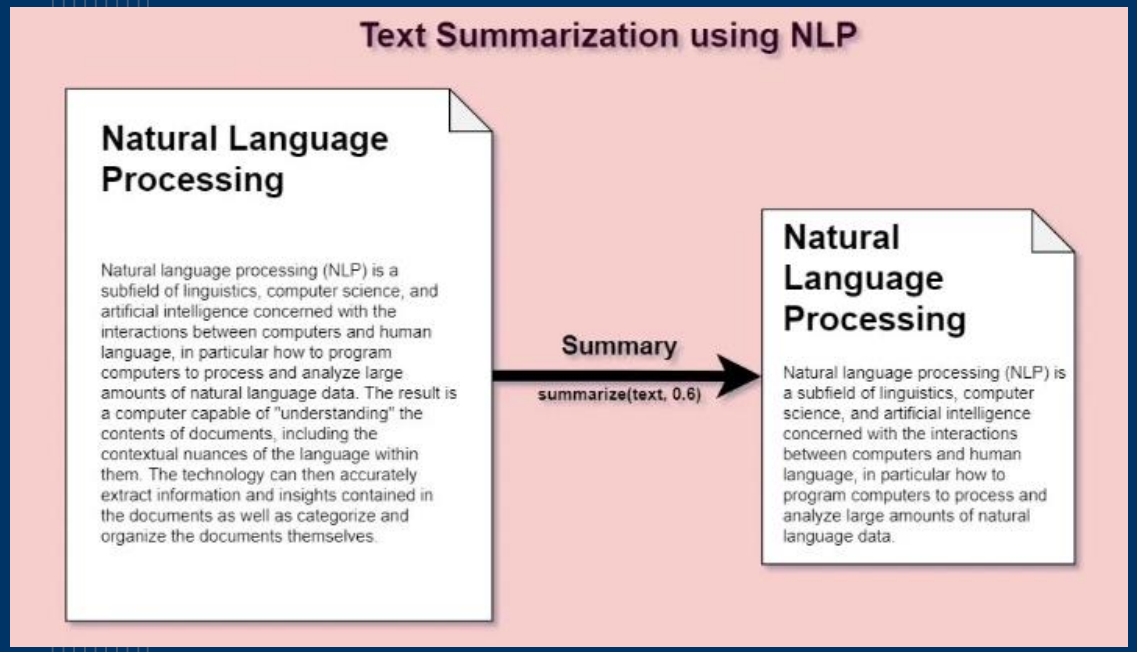
- Catch on to different sentiments
  - Analyze customer reactions
  - Handle social media disputes
    - By eradicating negative comments
    - By getting insights from the customer base of any business

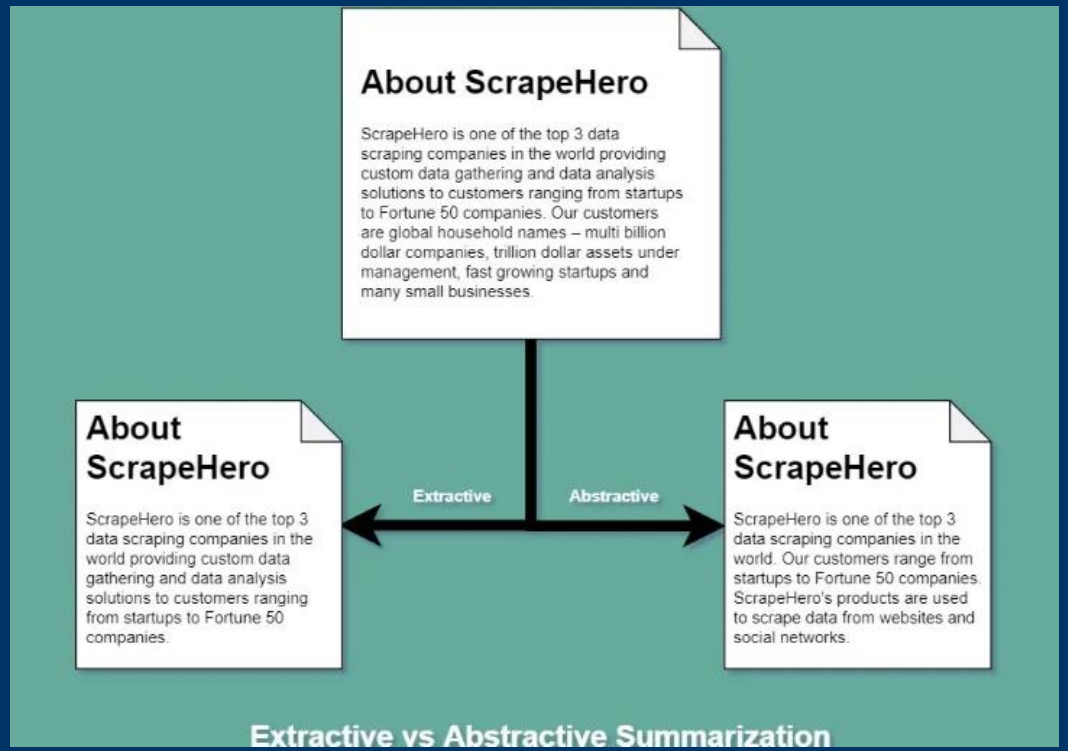
2.4 Translation



2.5 Text Summarization

- 대량의 텍스트 데이터를 요약
  - 블로그, 기사, 논문 등





3. NLP in Healthcare
3.1 헬스케어 자연어처리

- 헬스케어 자연어처리
  - 헬스케어 분야에서 EMR이나 PubMed 등을 통해 텍스트 데이터가 폭증하고 있음
  - 연평균 20% 이상 성장하여 2028년에는 9조원의 시장규모 (ReportLinker)
    - 의무기록 작성에 음성인식 기술이 도입
    - EMR/HER 을 자연어처리 기술로 분석하고 CAC 를 도입
    - 텍스트 마이닝을 통하여 과거 임상 사례들을 분석하여 현재 임상에 활용 가능
    - 대용량의 연구 문헌들을 분석하여 바이오마커 표현형을 찾거나 항생제 내성 관계 등을 추출

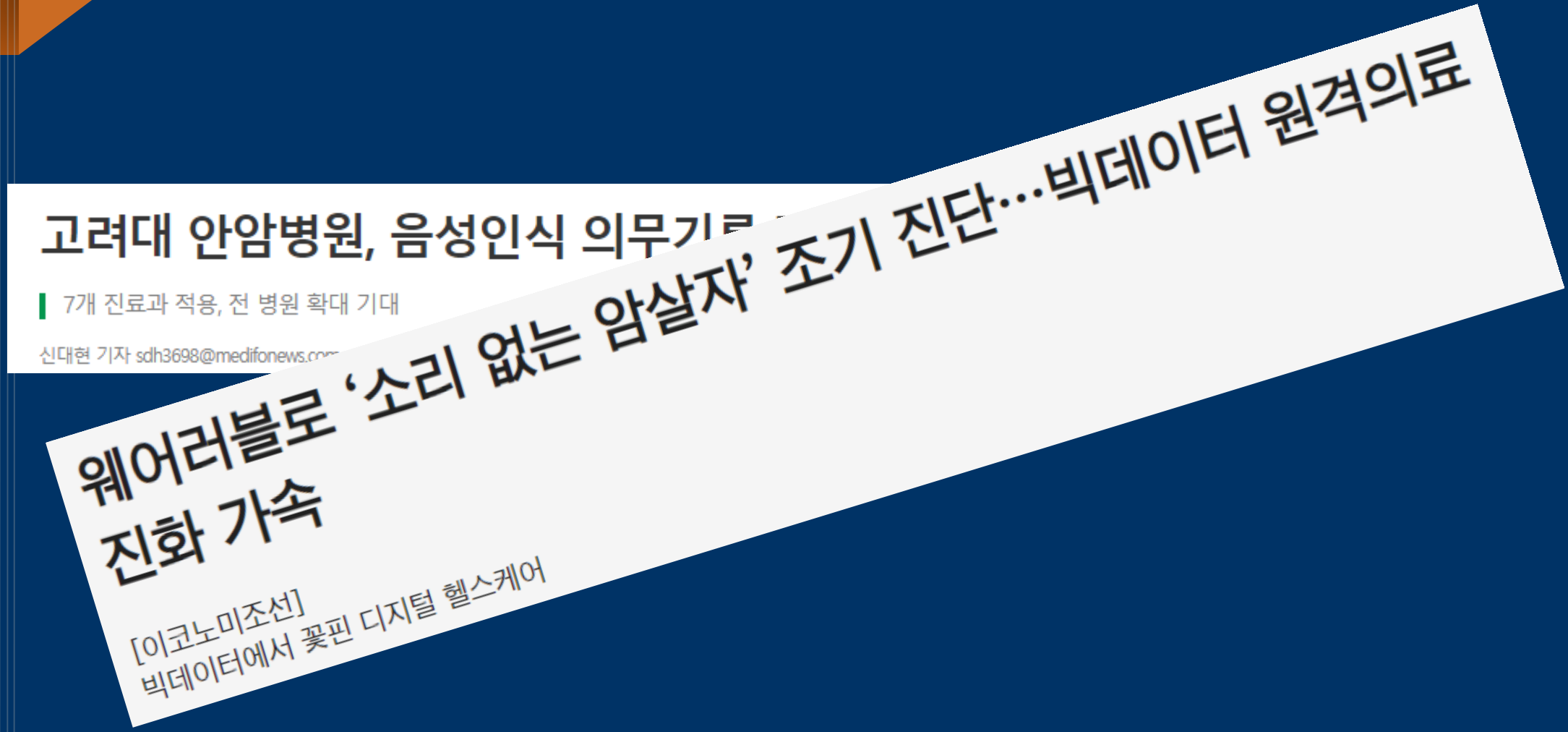
3.2 의료 현장에서의 헬스케어 자연어처리

고려대 안암병원, 음성인식 의무기록 작성 개발

7개 진료과 적용, 전 병원 확대 기대

신대현 기자 sdh3698@medifonews.com | 등록 2021-07-27 17:45:52

3.2 헬스케어 자연어처리



고려대 안암병원, 음성인식 의무기

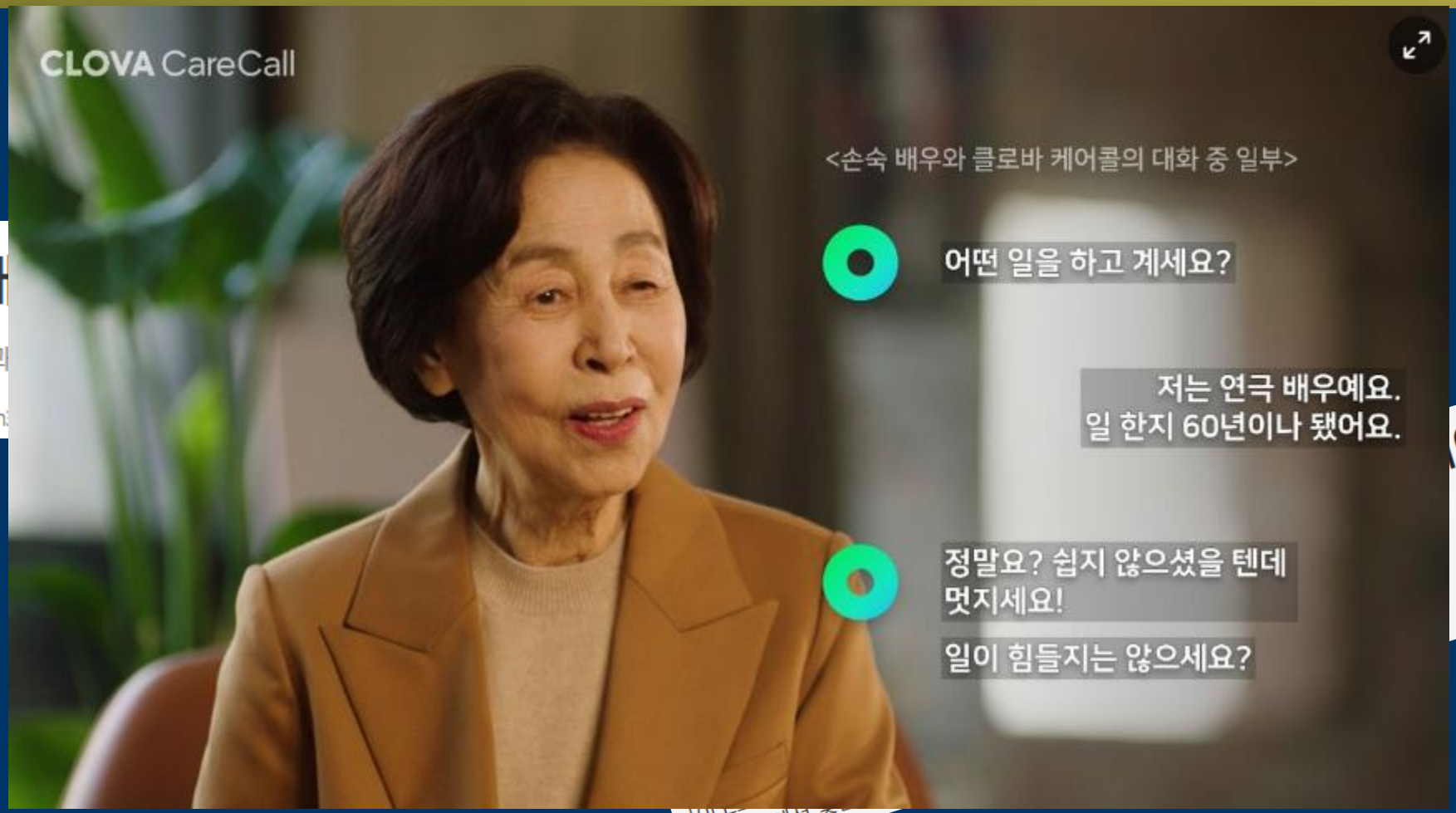
7개 진료과 적용, 전 병원 확대 기대

신대현 기자 sdh3698@medifonews.com

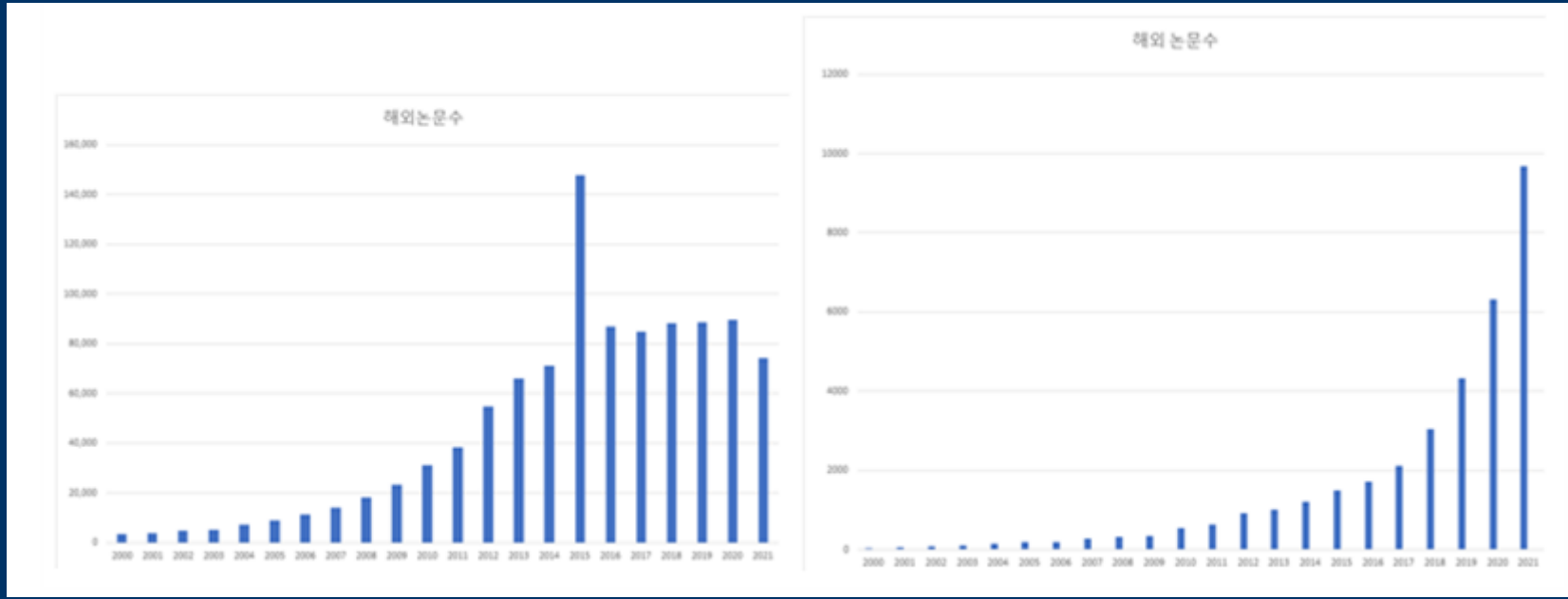
웨어러블로 '소리 없는 암살자' 조기 진단…빅데이터 원격의료 진화 가속

[이코노미조선]
빅데이터에서 꽃핀 디지털 헬스케어

3.2 헬스케어 자연어처리



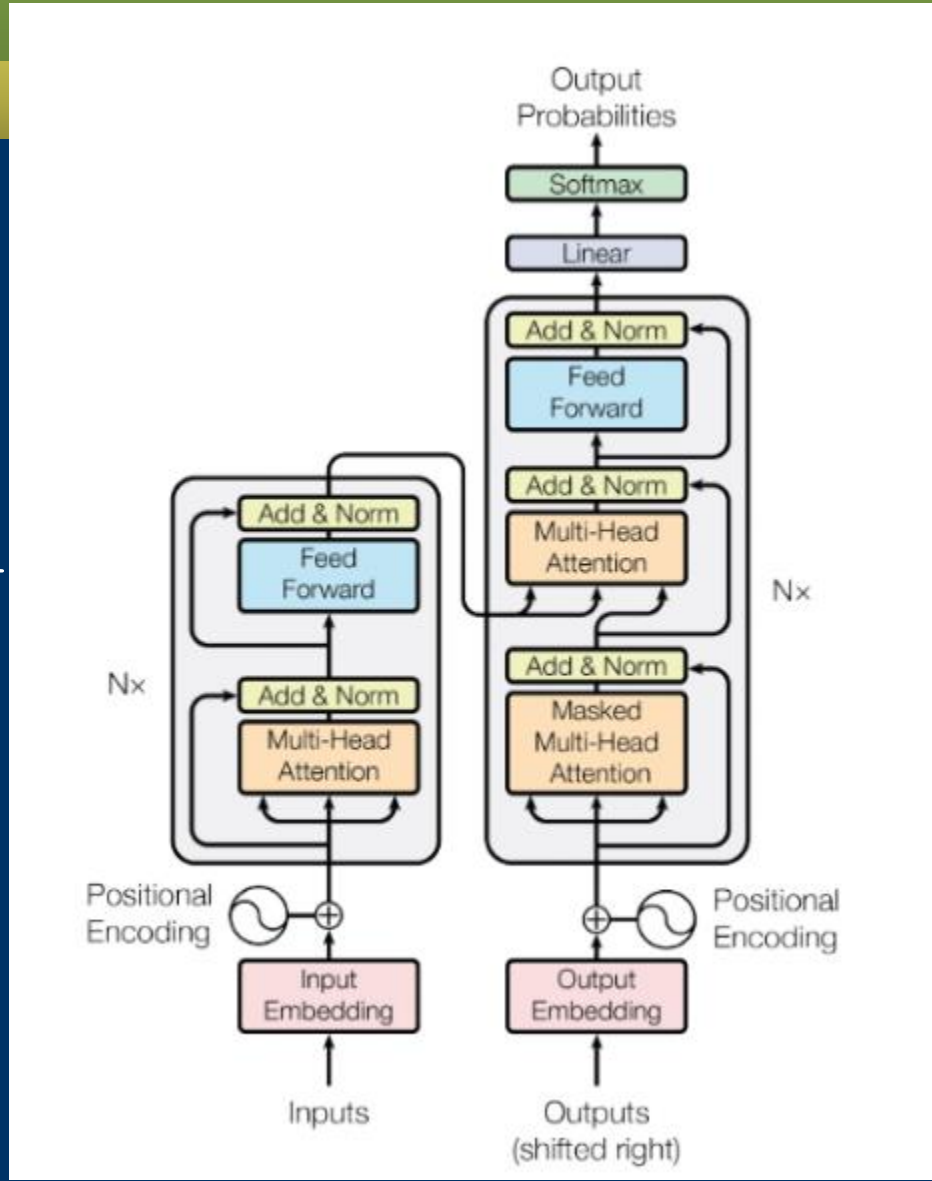
3.3 헬스케어 자연어처리의 연구 활성화 수준



	2017	2018	2019	2020	2021	평균
'헬스케어' + '인공지능'	151	165	199	499	259	254.60
'헬스케어' + '자연어처리'	4	7	8	13	21	10.60

3.4 최근 연구동향

- 트랜스포머 (Transformer) 모델
  - 초거대 AI model
  - DeepMind
    - 알파폴드2 를 개발하여 단백질 연구에 사용중
  - AstraZeneca
    - 메가몰바트를 구축하여 분자구조 훈련
  - 뮌헨공대
    - 자연어처리를 활용하여 단백질 연구

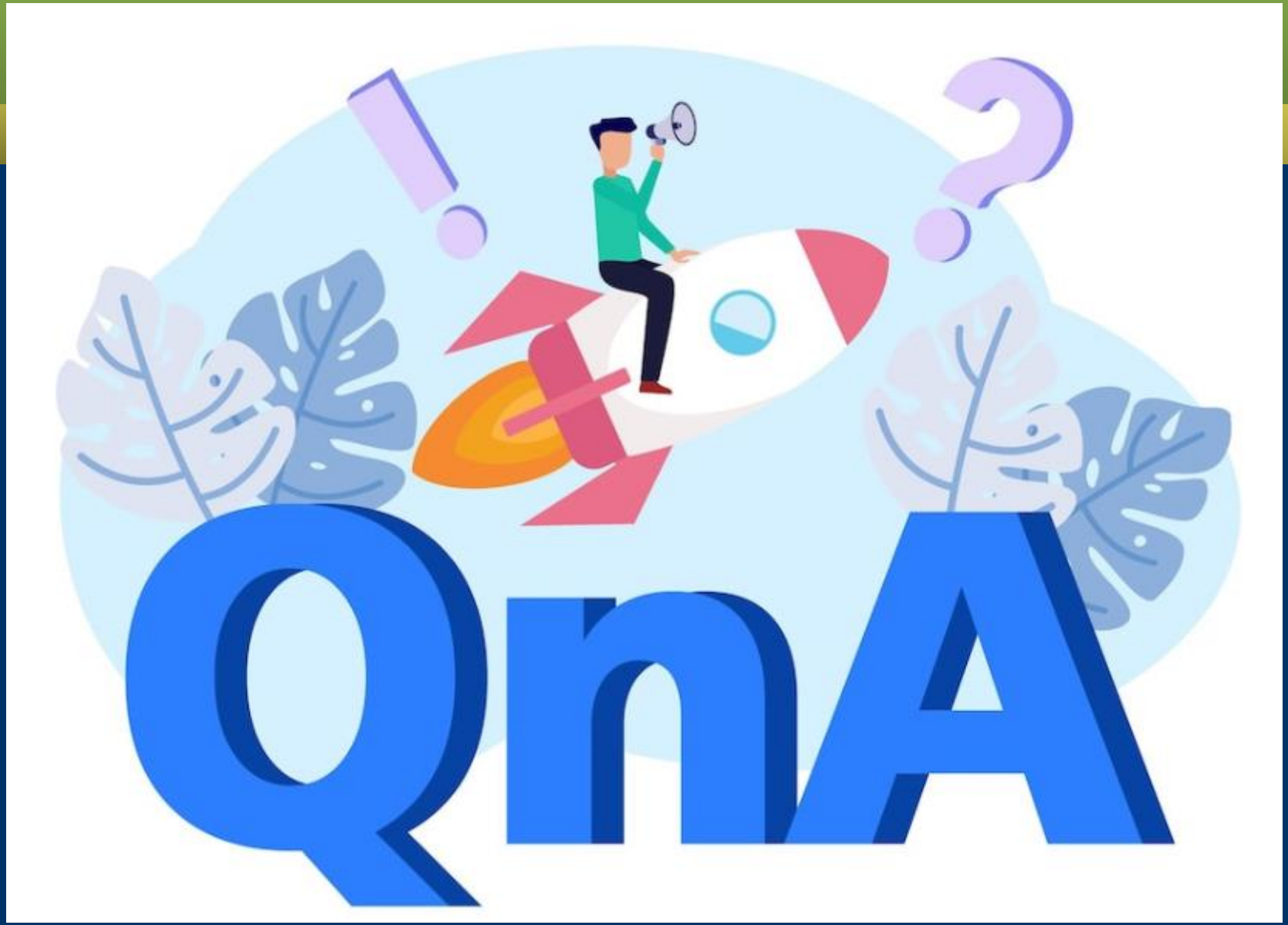


3.4 최근 연구동향

- 코로나 바이러스 돌연변이 발견
  - 바이러스 변이 지점을 바이러스의 문법에서 발견
  - MIT에서 언어 분석 인공지능을 사용하여 코로나19 바이러스 변이를 예측하고 백신이 효과적인 부위를 발견
  - HIV 서열 6만개, 인플루엔자 바이러스 서열 45천개, 코로나 바이러스 서열 4천개를 학습
  - 돌연변이 가능성이 큰 곳을 예측하여 변이에 대비
  - 가능성이 낮은 곳을 예측해 백신의 표적 발굴

4. 2022 NLP Trends

- NLP model with Artificial General Intelligence 개발
  - 단일 모델이 추론, 지식 표현, 계획, 학습, 소통 모두 가능
- Multilingual Language Modeling
- GPT-4??
- No code or Codeless NLP
- 빅테크 기업의 개발 가속화
- NLU 와 NLG 의 보다 원활한 결합


QnA