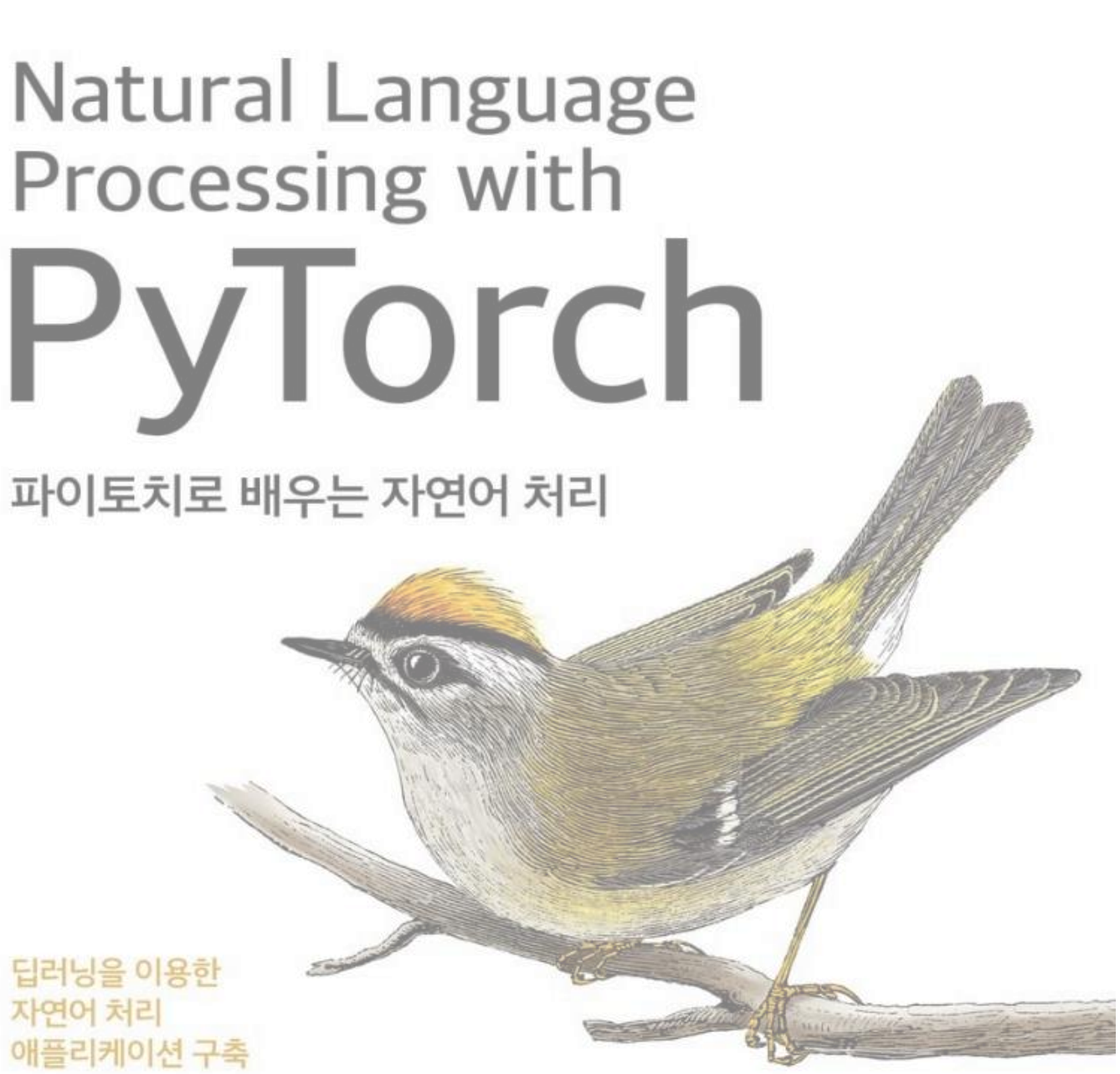

9장 발표

24.08.30
김태균

목차

고전, 최신 모델, 더 배울 것들

1. 전통적인 NLP 주제
2. 다양한 NLP 모델
3. NLP 시스템을 위한 디자인 패턴

고전, 최신 모델, 더 배울 것들

교재 내용 정리

- 파이토치 기초
- NLP 기본 개념
- 딥러닝 기본 개념
- MLP, CNN, RNN
- sequence 모델
- encoder-decoder 모델
- attention 메커니즘

전통적인 NLP 주제

1. 대화 및 상호작용 시스템

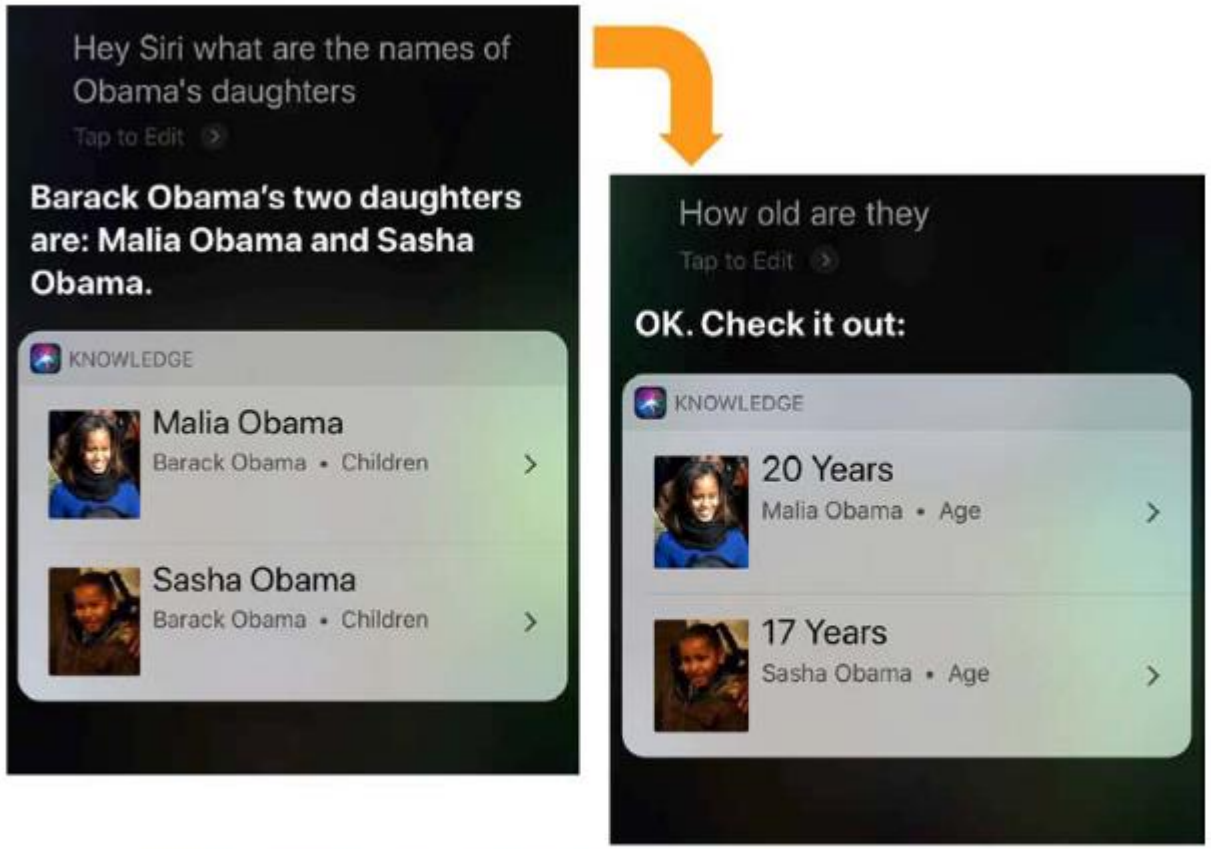

그림 9-1 실제 대화식 시스템(애플 시리). 시스템이 어떻게 문맥을 유지하여 이어지는 질문에 대답하는지 눈여겨보세요. 'they'를 버락 오바마Barack Obama의 딸로 인식합니다.

전통적인 NLP 주제

2. 담화 분석

표 9-1 CoNLL 2015의 단순 담화 분석Shallow Discourse Processing 작업의 예

예시	담화 관계
GM officials want to get their strategy to reduce capacity and the workforce in place before **those talks begin**.	Temporal.Asynchronous.Precedence
But that ghost wouldn't settle for words, he wanted money and people—lots. So **Mr. Carter formed three new Army divisions and gave them to a new bureaucracy in Tampa called the Rapid Deployment Force**.	Contingency.Cause.Result
The Arabs had merely oil. Implicit=while These farmers may have a grip on the world's very heart	Comparison.Contrast

(a) The dog chewed the bone. It was delicious.
(b) The dog chewed the bone. It was a hot day.
(c) Nia drank a tall glass of beer. It was chipped.
(d) Nia drank a tall glass of beer. It was bubbly.

그림 9-2 대용어 복원에서 일어날 수 있는 몇 가지 문제. (a)에서 'It'은 dog 또는 bone을 의미하나요? (b)에서 'It'의 의미는 둘 다 아닙니다. (c)와 (d)에서 'It'은은 각각 glass와 beer를 의미합니다. 유리잔보다 맥주에 거품이 있을 가능성이 높음을 아는 것은 참조 대상을 찾는 데 중요한 역할을 합니다(선택적 선호도selectional preference).

전통적인 NLP 주제

3. 정보 추출과 텍스트 마이닝
   - **정보 추출** : 텍스트에서 구조화된 정보를 자동으로 식별하고 추출하는 과정
   - **텍스트 마이닝** : 대량의 텍스트 데이터에서 유용한 패턴이나 지식을 추출하기 위한 방법론

4. 문서 분석과 문서 추출

ex) 토픽 모델링, 쿼리 파싱, 요약 등

다양한 NLP 모델

- **여러 모델의 조합**

  ex) 단어의 문자에 대한 CNN을 만든 다음 LSTM 연결하고, 마지막에 MLP를 사용해 출력
- **시퀀스를 위한 합성곱**

  ex) 기계 번역
- **어텐션** 메커니즘

  ex) 셀프 어텐션, 멀티헤드 어텐션
- **전이 학습**
- **강화 학습**

NLP 시스템을 위한 디자인 패턴

1. 온라인 vs 오프라인 시스템
   - **온라인 시스템** : 실시간 작업

     ex) 챗봇 서비스
   - **오프라인 시스템** : 효율적인 작업

     ex) 문서 분류, 텍스트 요약

일반적으로 **오프라인 시스템 → 온라인 시스템** 순서로 구축

NLP 시스템을 위한 디자인 패턴

2. 상호작용 vs 비상호작용 시스템

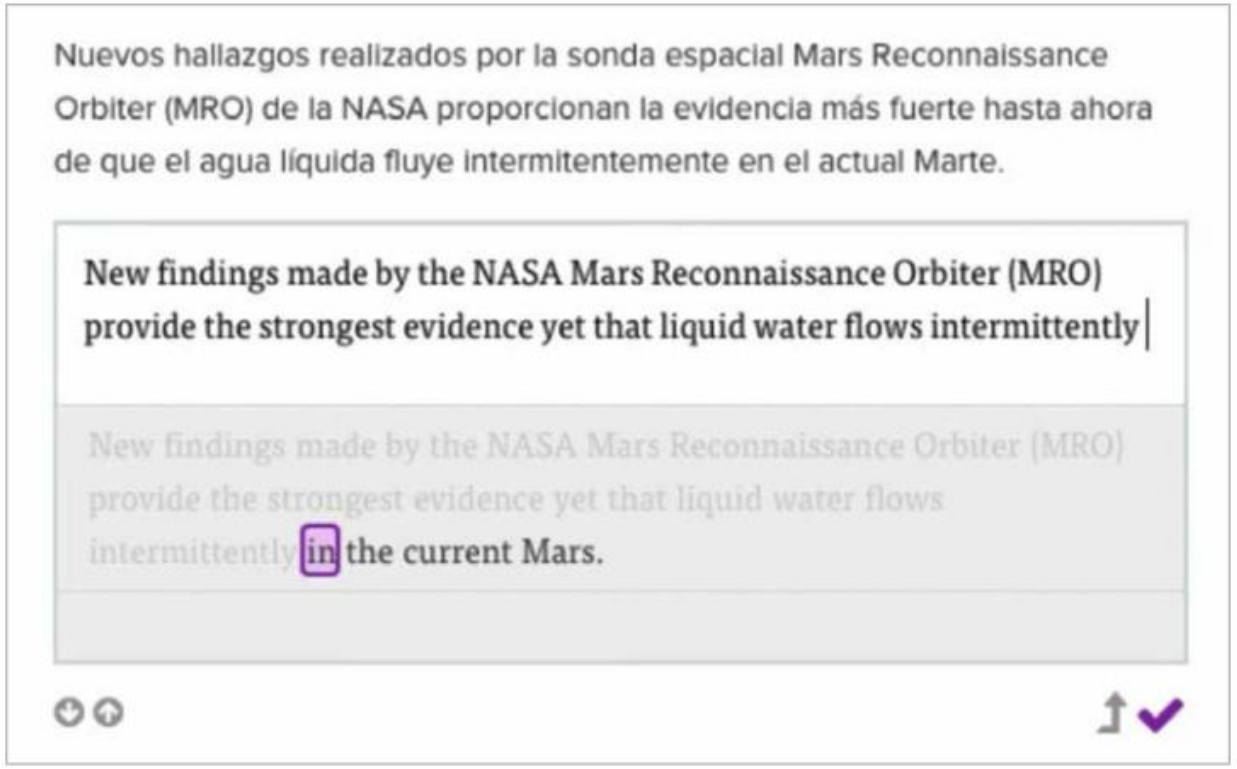

그림 9-4 사람이 참여하는 실제 기계 번역 모델. 사람이 기계 번역 시스템의 제안을 고치거나 수정하여 높은 정확도의 번역을 만들 수 있습니다(출처: Lilt Inc.).

NLP 시스템을 위한 디자인 패턴

3. 유니모달 vs 멀티모달 시스템

- Modality : **정보나 데이터가 표현되는 방식이나 유형**

  ex) 텍스트, 이미지, 음성

	유니모달	멀티모달
장점	단순성, 해석 용이	다양한 정보 반영, 복잡한 문제 해결
단점	제한된 표현력	복잡성, 높은 자원 소모

NLP 시스템을 위한 디자인 패턴

4. 엔드투엔드 vs 분할 시스템

	엔드투엔드	분할
장점	구현, 배포 용이	모듈화
단점	유연성 부족	통합 및 성능 관리가 어려움

NLP 시스템을 위한 디자인 패턴

5. 폐쇄형 도메인 vs 개방형 도메인 시스템
   - **폐쇄형 도메인** : 해당 도메인에서 잘 동작하도록 최적화

     ex) 의학 저널에 최적화 된 기계 번역
   - **개방형 도메인** : 범용적인 목적으로 사용

     ex) 구글 번역기

NLP 시스템을 위한 디자인 패턴

6. 단일 언어 vs 다중 언어 시스템
   - **단일 언어** : 한 언어를 다루는 NLP 시스템
   - **다중 언어** : 여러 언어를 다루는 NLP 시스템

관련 자료

- 파이토치 포럼(https://discuss.pytorch.org)
- 인공지능(https://arxiv.org/list/cs.AI/recent)
- 컨퍼런스
  - ACL(Association of Computational Linguistics)
  - EMNLP(Empirical Methods in Natural Language Processing)
  - NAACL(North American Association for Computational Linguistics)
  - EACL(European chapter of ACL)
  - CoNLL(Conference on Computational Natural Language Learning)

QnA