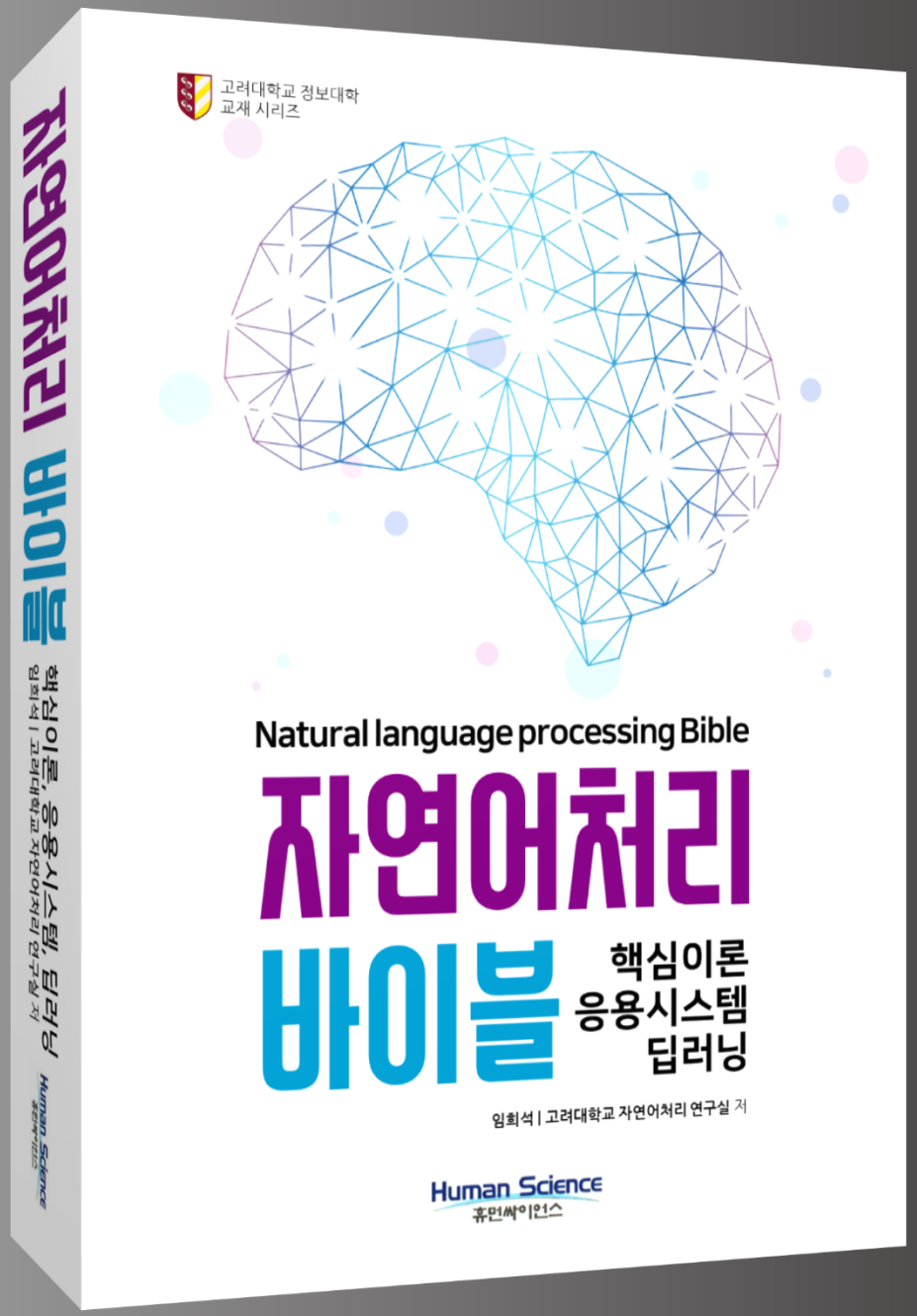


CHAPTER 12
기계 번역
(Machine Translation)

기계 번역이란

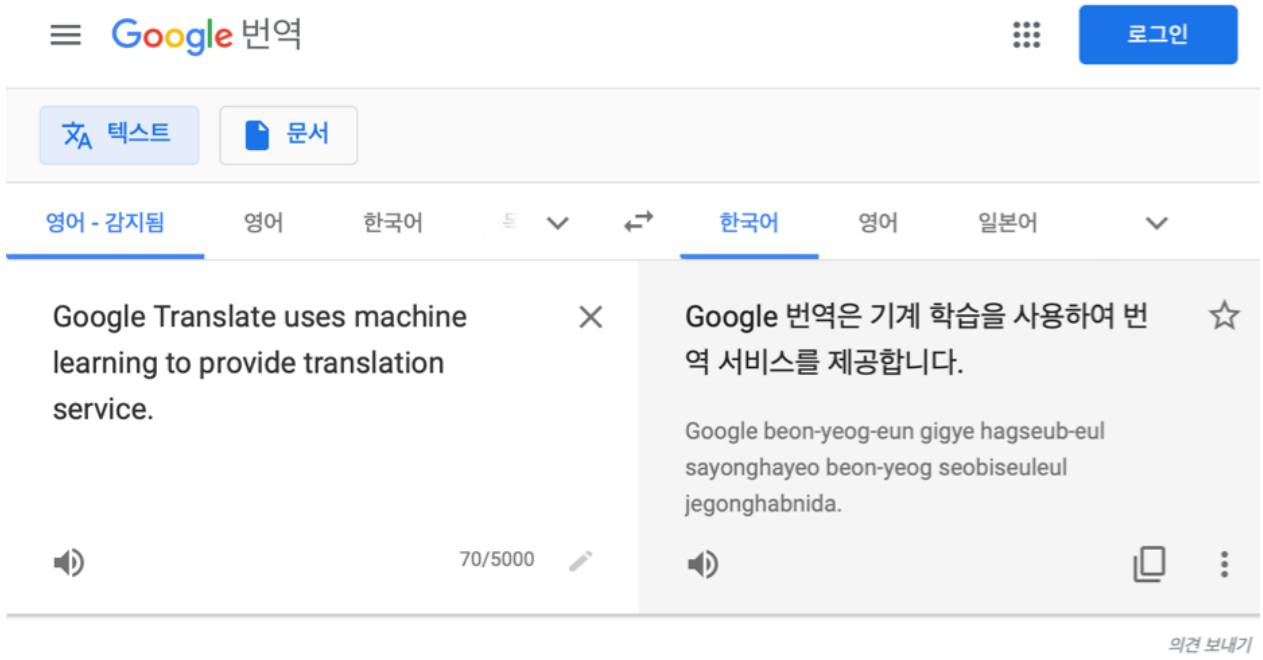
- 번역: 하나의 언어로 쓰인 글을 같은 의미를 나타내는 다른 언어로 변환하는 작업
  - 세계의 다른 나라들과의 손쉬운 교류가 가능해지면서 필수적인 작업이 됨

- 필요성과 중요성에 비해 사람이 직접 할 수 있는 데에는 한계가 있음
  - 번역 작업의 난이도가 매우 높음
  - 필요한 두 언어쌍을 모두 알고 있는 번역가를 구하기 어려움
  - 번역의 속도도 느린 편

- 기계번역: 번역을 컴퓨터가 빠르게 수행하는 것
  - 1995년 알타비스타 사에서 제작한 '바벨피쉬' 번역 서비스: 형편없는 품질로 금방 잊혀짐
  - 현재는 마이크로소프트 빙 번역, 구글 번역 등의 온라인 서비스가 활발
  - 만들어진 번역 시스템을 오프라인으로 옮겨 여행용으로 사용하는 시도도 있음

기계 번역이란

- 구글 사에서 제공하는 번역 서비스의 예시



- 최근 신경망 기반의 번역을 제공하기로 선언한 우 번역의 품질이 향상되었다는 평가를 받음

규칙 기반 기계 번역

- 언어학적, 문법적인 규칙을 양 언어에 쌍으로 적용하여 문장을 번역하는 것
  - 원본 문장을 형태소나 구문 등으로 분석하고 분해된 내용들을 번역한 후, 재조합

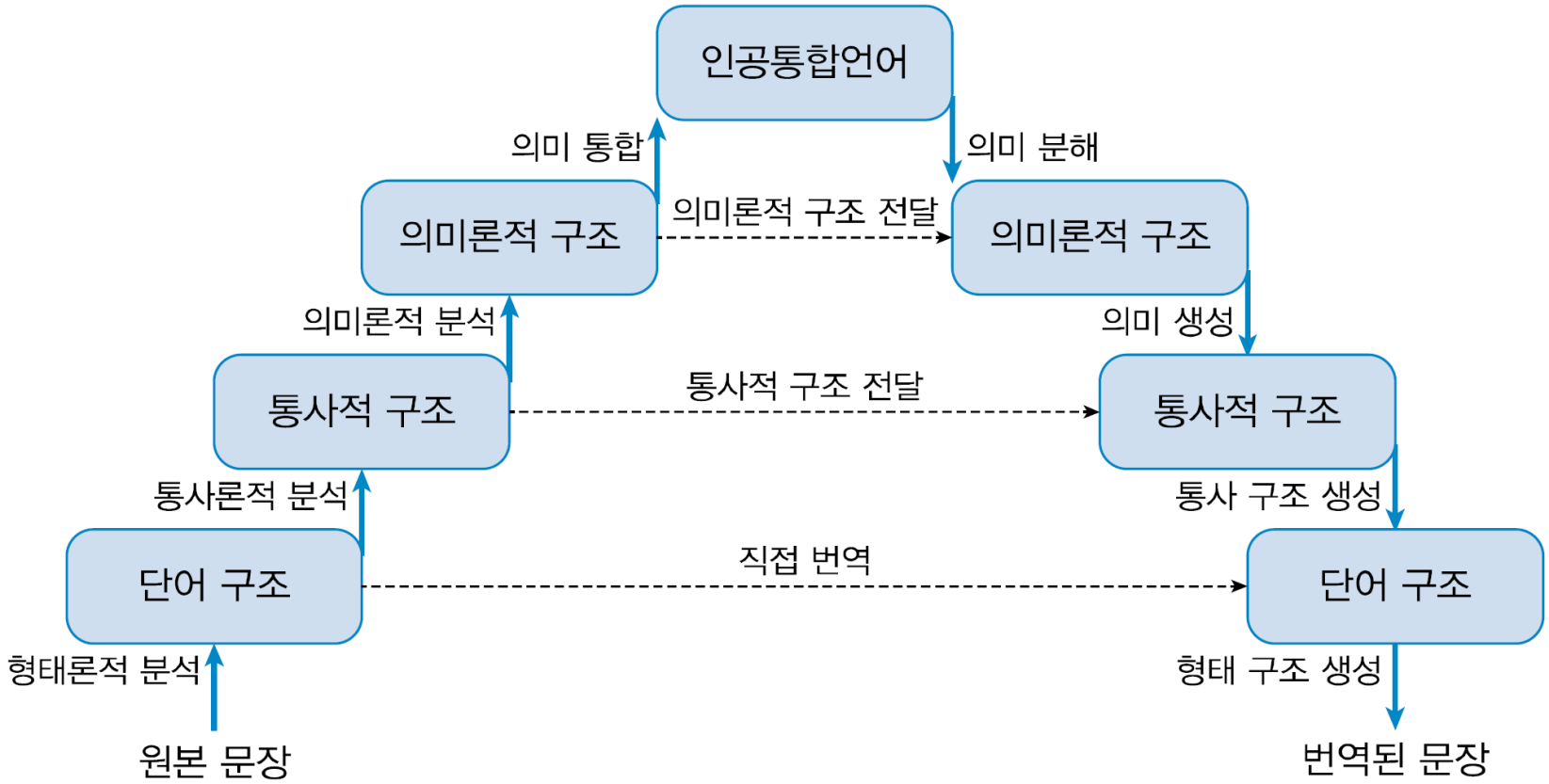


그림 12-2 규칙 기반으로 문장을 번역하는 과정을 설명한 Vauquois Triangle

규칙 기반 기계 번역

- 널리 알려진 문법 규칙을 적용하며 번역하므로 문법적으로 정확한 번역을 기대
- 하지만 규칙을 적용시키기 위한 단위를 분해하는 과정의 정확도가 낮아 발현되지는 않음

- 바벨피쉬 번역기의 한국어→영어 번역기: 잘못된 형태소 분석에 의한 번역 실패의 예시
  - "원어민": 어떤 나라의 말을 모국어로 쓰는 사람 (고려대 한국어대사전)
  - 영어 한 단어로 번역되지 않는 단어
  - 이를 원 + 어민 으로 분해하여 Circle Fishermen으로 번역

- 명확한 한계점을 갖고 있기에 다른 번역 기법이 개발된 후 사장됨
  - 규칙 자체의 순수한 장점을 살리기 위해 딥 러닝 기반에 융합하려는 시도도 존재

통계 기반 기계 번역

- 대량의 문장을 담은 코퍼스를 바탕으로 상관관계를 분석한 통계 모델로 번역을 진행

- 코퍼스의 사용
  - 비교 코퍼스: 같은 주제를 다루고 있는 문장 (같은 사건을 다룬 다른 언어의 기사)
  - 병렬 코퍼스: 문장을 직접 번역한 일대일 쌍으로 구성된 코퍼스
  - 비교 코퍼스가 구축이 쉽지만, 정확도 등에서 불리한 점이 많아 병렬 코퍼스가 주로 사용

한국어	일본어
언어학의 두 흐름으로 이론적 연구와 실증적 연구가 있듯이 컴퓨터를 이용한 언어 정보화에도 두 가지 흐름이 있다.	言語学の二つの流れに理論研究と実証研究があるように、コンピューターを利用した言語情報化にも二つの流れがある。
양자 모두 컴퓨터를 이용한다는 점에서는 일치하지만 실제 결과로서 나타난 언어에 대한 기술 모델은 상이하다.	両者共にコンピューターを利用するという点では一致するが、実際の結果としてあらわれる言語に対する技術モデルは相違する。
일본의 언어 연구의 특징은 실증적 연구 중시에 있다.	日本の言語研究の特徴は、実証的な研究を重視する点にある。

통계 기반 기계 번역

- 코퍼스의 양과 질이 성능을 가장 크게 좌우함

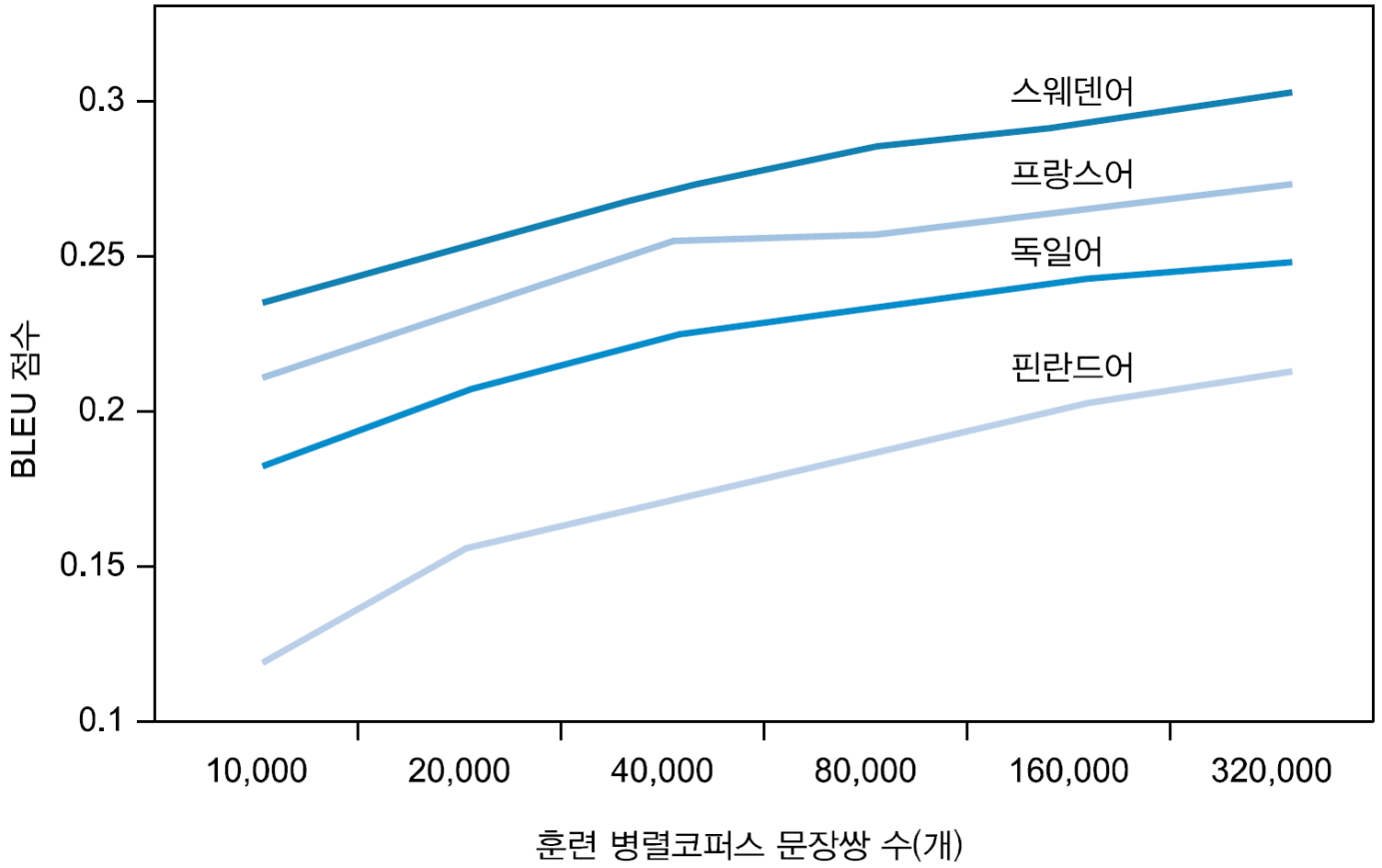


그림 12-3 코퍼스의 크기에 따라 정확도(Bleu 점수)가 상승하는 양상을 보여주는 그래프

통계 기반 번역의 수학적 접근

- 조건부 확률: 입력된 영어 문장 E에 대해, 번역된 프랑스어 문장 F를 찾음

- $P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{count(F,E)}{count(E)}$

- 통계 기반 번역의 기본이 되는 수식이지만, 이것만으로는 번역을 수행할 수 없음
  - 세상의 모든 문장이 코퍼스에 수록되어있는 것이 아니기 때문
  - 문장을 많이 실을수록 코퍼스 구축이 어려워짐
  - 언어가 무한하기 때문에(1장) 모든 문장을 싣는 것 자체가 불가능함
  - 같은 의미를 나타내는 여러 표현 등이 있으므로, 0이 너무 많이 나와 계산이 불가능한 경우가 많음

- 조건부 확률 개념은 그대로 두고, 단어 단위로 분할

- $P(FvertE) = \sqcap_{E_j} \operatorname{argmax}_{F_i} P(F_i \mid E_j)$

번역의 질 지표화

- $P(F|E) = \sqcap_{E_j} \text{argmax}_{F_i} P(F_i \mid E_j)$
- 위 수식만으로 여전히 번역을 진행할 수는 없음
  1. 단어가 항상 한 가지 의미로만 번역되지는 않음(다의어), 같은 글자여도 발음이 다르거나 의미가 여럿일 수 있음 (동음이의어). 비슷한 의미여도 문장 상황에 따라 사용되는 단어가 다름
  2. 언어별로 문법이 다르기 때문에 단어를 직접 변환만 하면 자연스럽게 이어지지 않음
  3. 언어에 따라 어순이 다르기 때문에 순서를 유지한 채 번역만 하면 자연스럽지 않음
- 문장을 고르는 조건을 추가하여 문장 상태를 수학적으로 판단
- 베이즈 정리
  - $F_{best} = \text{argmax}_F P(F|E) = \text{argmax}_F \frac{P(F)P(E|F)}{P(E)} = \text{argmax}_F P(F)P(E|F)$
  - 실제로 확률적인 사건을 구하는 것이 아니므로 P(E|F)와 P(F|E)에는 큰 차이가 없음
  - 베이즈 정리를 통해 형태 변환만으로 P(F)라는 지표를 만들어냄
  - P(F)는 10장의 '언어 모델'과 관련한 것으로, 해당 문장 자체의 자연스러움을 평가
- 입력된 영어 문장이 심하게 훼손된 프랑스어 문장이라고 보는 것: Noisy-channel Model
- 이같은 방법으로 여러 함수를 도입해 번역의 질을 평가: Log-linear Model

구 기반 번역

- Tomorrow I will fly to the conference in France.
- 나는 내일 프랑스의 컨퍼런스에 참가한다(참가하기 위해 비행기를 탄다).

- "To the conference" 가 "컨퍼런스에" 와 같이 단어별이 아닌 구 별로 번역되는 경우가 많음

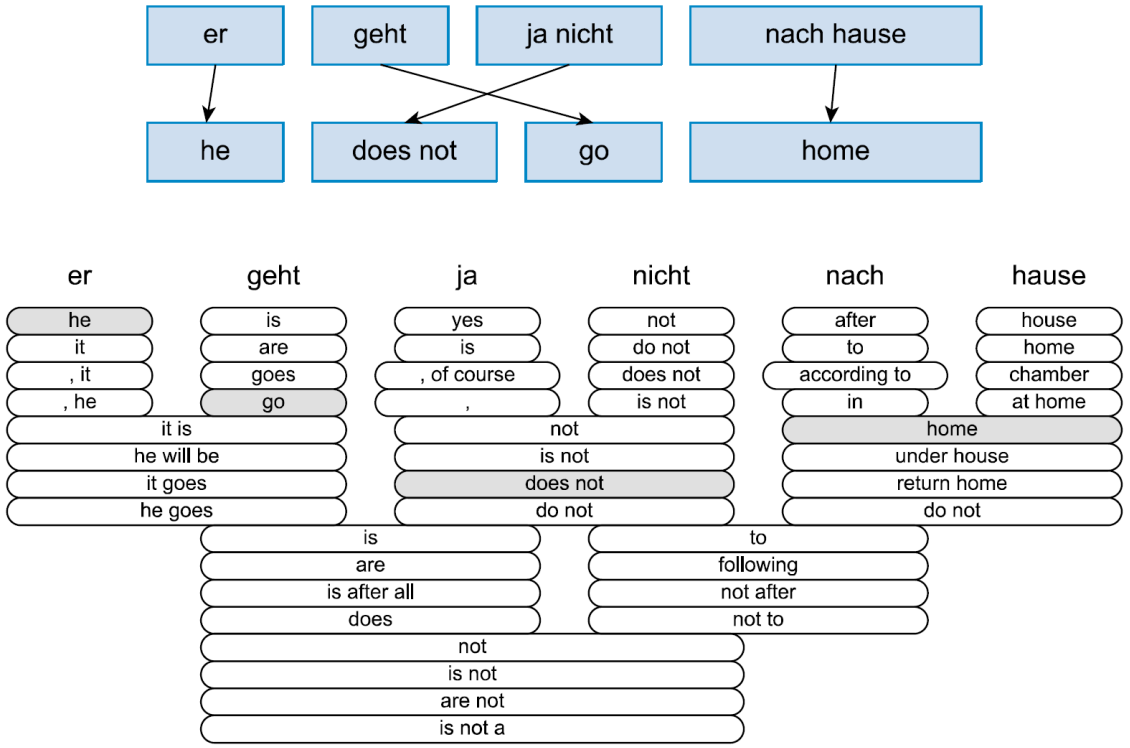


그림 12-4 통계 기반 기계 번역의 수행 과정

딥러닝기반 기계학습

25장에서 학습 예정...