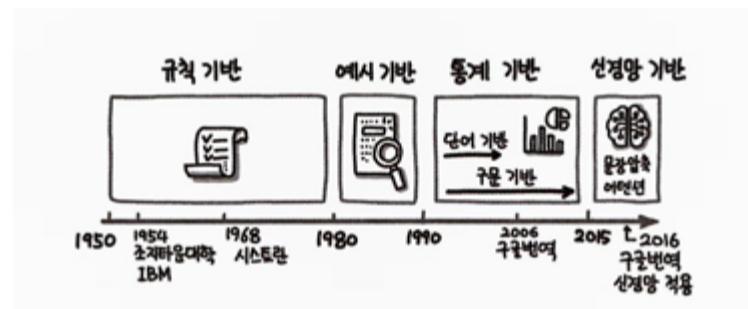


제 6장. 기계번역

인간의 언어가 정말 어려운 이유

- 너무 많은 규칙
 - 인간의 언어는 살아 움직이는 생명체처럼 끊임없이 진화함
 - 일정한 패턴이 아님
 - ex) '담배'라는 단어의 어원은 포르투칼어인 tabaco를 일본에서 담바고라 불렀고 이게 우리나라로 넘어와 담바 등으로 부르다 '담배'가 표준어가 됨
- 너무 많은 오류
 - 일상 대화에 엄청나게 많은 오류가 있음
 - 그럼에도 대화가 가능한건 뇌가 오류를 보정하고 이해하기 때문
- 너무 많은 규칙
 - 같은 발음을 지닌 단어가 여러 뜻을 갖는 경우가 흔함

기계번역의 시작



- 기계번역
 - 인간이 사용하는 언어를 기계를 사용해 다른언어로 번역하는 일

규칙기반, 모든 규칙을 정의하다

있다고 가정해보죠.

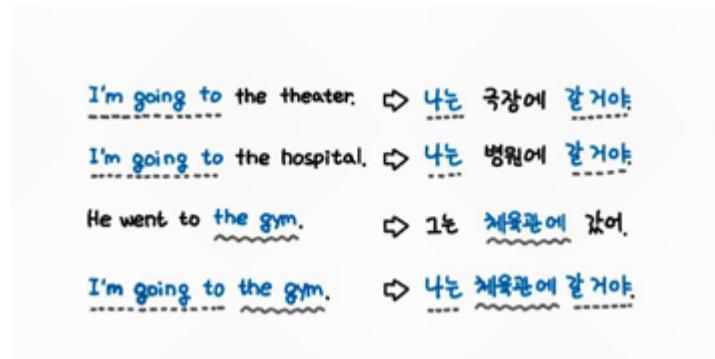
<p>walk ◆ 걷다 walking ◆ 걷고 있다</p> <p>run ◆ 달리다 running ◆ 달리고 있다</p>	<p>여기까지는 규칙이 잘 통하는 것처럼 보입니다. 그렇다면 'fight'는 '싸우다'니까 'fighting!'은 '싸우고 있다!'로 번역하면 될까요? 정말 어색한 번역이죠. '힘내!' 또</p>
--	---

싸우고 있다!

- 기계번역을 대표하는 회사로 '시스템'이라는 회사가 있음
 - 규칙 기반 기계번역을 이용

- 모든 규칙을 정해야 하므로... 반드시 한계가 존재

예시 기반과 통계 기반, 가능성을 보이다



- 예시 기반
 - 교토대학교의 나가오 마코토라는 교수가 예시 기반 기계번역이라는 방식을 제안
 - 사람들이 실제 활용하는 문장 전체의 맥락을 살펴보는 데 주안점을 둠
 - 규칙을 통해 언어를 '이해' 하는게 아닌 경험을 통해 '모방'하는 형태
 - 영어 공부할 때 '숙어'를 암기하고 단어를 알아끼워보며 문장을 번역하는 과정과 유사함

한 번역기 시민 커미노 번역자는 번역률 계산입니다. 예를 들어
'House'를 번역한다면 문장 안에서 '집'으로 번역되는 경우가 가장 많았고, '가정'으로 번역되는 경우가 그다음, '상점'으로 번역되는 경우가 가장 적었을 것입니다. 이렇게 번역될 확률을 계산해 다음과 같이 순서대로 결과를 나열합니다.

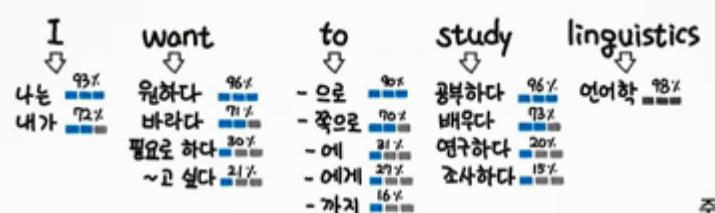
House

집 ■■■ 95%

가정 ■■■ 78%

상점 ■■■ 42%

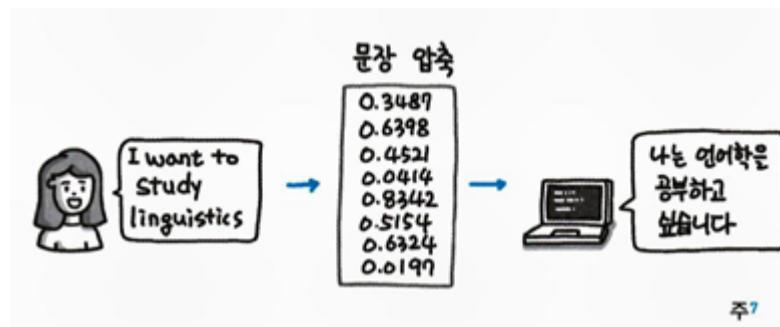
이렇게 문장에 속한 모든 단어의 번역 확률을 나열하고 확률이 높은 순으로 번역된 의미를 조합해 번역문을 만듭니다. 번역 품질을 높



- 통계 기반
 - 문장을 단어 또는 구문 단위로 분할한 다음 이를 번역하고 다시 문장으로 합치는 과정에서 확률적인 방법을 접목

인공 신경망 기반, 마침내 혁신이 시작된다

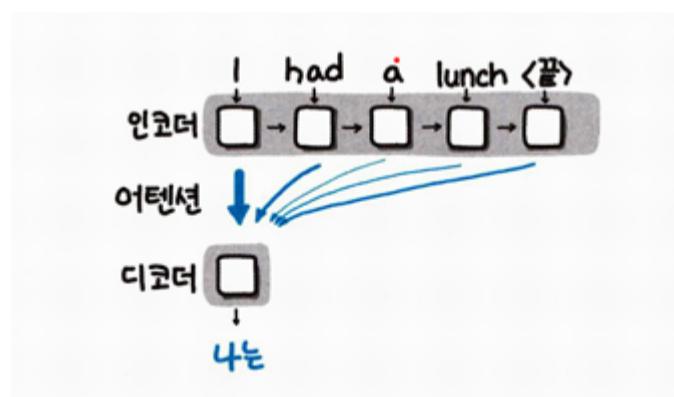
- 문장 전체에 딥러닝을 적용하는걸 신경망 기반 기계번역 이라 함

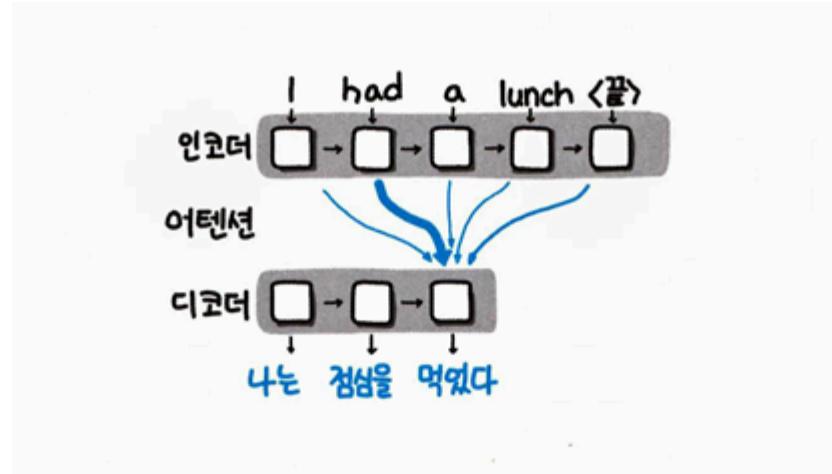


- 신경망이 문장을 통째로 번역하는 과정 예시
 - 문장을 통째로 압축해 숫자로 표현한 벡터 값을 만들어냄 (주스 → 농축)
 - 값을 번역할 언어로 옮긴 다음 풀어서 번역문을 만들어 냄
 - 각각의 숫자에서 가장 확률이 높은 번역문을 찾아냄 (물을 섞어 다시 주스로 만드는 과정)

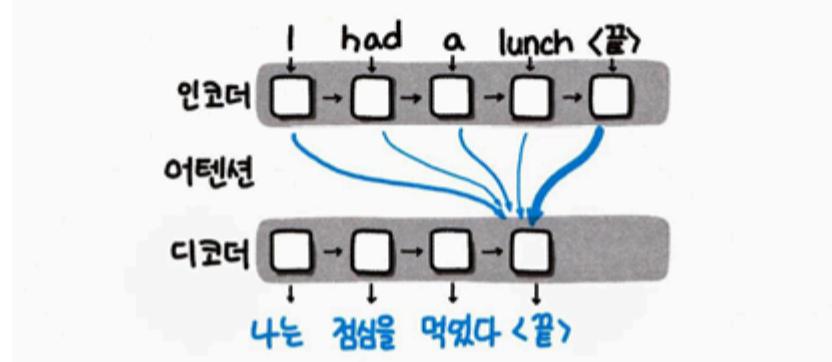
어텐션, 가장 혁신적인 발명

- 어텐션
 - 더 중요한 단어를 강조하는 원리
- 어텐션의 경우 번역문의 단어를 생성할 때마다 출력 문장의 길이에 맞춰 압축 벡터를 생성
 - 이전에는 인코더 디코더를 통해 어떤 분량이든 1줄로 요약했지만 어텐션은 5줄, 10줄 등 자유롭게 요약
- 핵심은 중요한 단어에 별도의 가중치를 부여할 수 있음
 - 이 덕분에 번역 시 어떤 단어를 염두에 두어야 하는지 알 수 있게 되어 번역의 질이 상승
- 예시





다시 번역 과정을 살펴보겠습니다. 그다음에는 'had'에 가중치를 매긴 후 내려보내 이에 대응하는 의미인 '먹었다'로 번역합니다.



- 어텐션만으로 만든 모델인 트랜스포머가 등장하며 연구가 더 활발해짐
 - 다음 챕터에 있는 GPT3도 트랜스포머 모델을 활용
- 현대자동차 얘기도 나오는데... 잘 모르겠음 못들어봤음;