기계번역기술은 규칙기반, 통계기반, 인공신경망 기반을 걸쳐 발전했습니다.

**규칙기반 기계번역(RBMT)**은 언어학적, 문법적으로 언어를 번역하는 것으로 번역대상의 언어를 형태소, 구문분석 등의 과정을 거쳐서 목표 언어의 문법적 규칙에 맞게 변환시켜 번역을 진행하는 방식입니다.

**통계기반(SBMT)**은 대용량의 텍스트 말뭉치로부터 학습된 통계정보를 통해 번역을 진행하는, 즉 확률을 사용하는 방식입니다. 예를 들어 한 가지 문장이 여러 가지 표현으로 번역될 경우, 이 중 가장 자주 나타나는 번역 결과를 선택합니다. 또한 어순이 다른 언어를 번역할 때에도 문장을 구, 절 단위로 번역한 뒤 어순을 무작위로 섞어 번역하고 그 중 가장 자연스러운 문장을 고릅니다.

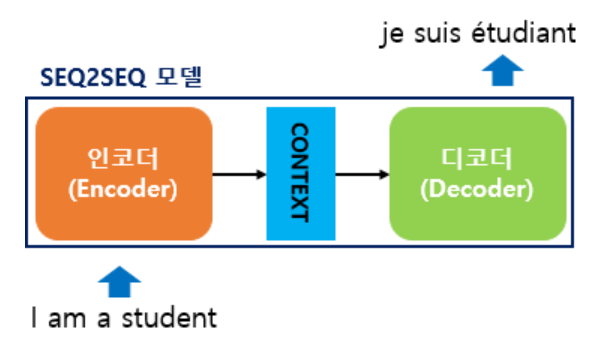
ex) ‘I eat rice’라는 문장을 영어로 번역할 경우, 여기서 ‘rice’ 라는 단어는 밥으로 번역될 수도 있고 쌀으로 번역될 수도 있다. 근데 실제 온라인 상에서 ‘나는 쌀을 먹는다’는 표현은 거의 없고 ‘나는 밥을 먹는다’는 표현이 보편적으로 쓰이기에 컴퓨터는 통계적으로 후자를 택합니다.

ex) ‘I eat an apple’이라는 문장을 번역할 경우 ‘eat’을 먼저 놓기도 해보고, ‘apple’을 먼저 놓기도 해봅니다. 여기서 가장 자연스러운 문장을 고르는 것입니다.

**인공신경망 기계번역(NMT)**는 딥러닝을 이용하여 번역하는 방식입니다. 규칙기반이나 통계기반과 달리 전체 문장을 하나의 번역 단위로 간주해 모든 정보, 네트워크를 통째로 학습하게 된다. 그러면 기계가 한 번도 학습한 적이 없는 문장을 입력하더라도 네트워크를 통과시켜 자동으로 번역해줍니다. 이는 문장 전체의 맥락을 먼저 파악한 후 어순, 의미, 문맥별로 의미 차이 등을 반영해 가장 적합한 문장으로 재배열하는 방식이다

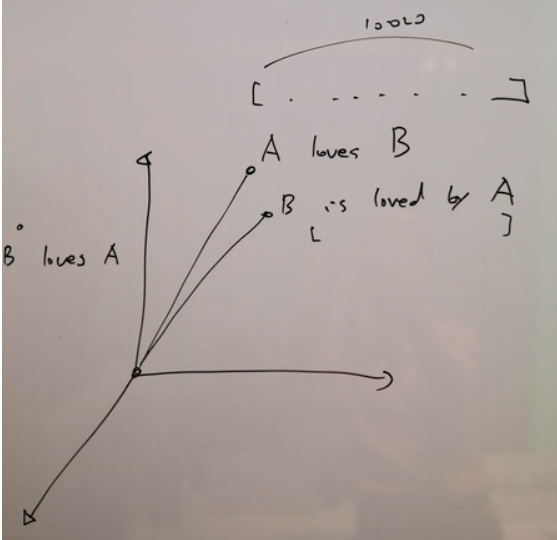
이제 저희가 번역기에 사용할 NMT에 대해 설명하겠습니다. NMT는 seq2seq 모델을 기반으로 하는데 seq2seq는 인코더와 디코더로 이루어집니다. 입력 문장이 들어오면 **Encoder**에서 인간의 언어를 고차원의 Vector 형태로 바꾸어 컴퓨터가 이해할 수 있는 형태로 바꾸어 줍니다. 그러면 **Decoder**에서 다시 vector 형태의 문장을 풀어 출력 문장을 생성합니다.

이렇게 생성된 출력 문장을 올바른 해석 문장과 비교해보면서 오답의 ‘거리’를 측정합니다. 이 때 인코딩과 디코딩을 반복하며 오답을 수정하여 정답과의 거리를 좁혀 나갑니다. 대용량의 데이터로 이 과정을 충분히 반복하다 보면, 두 언어 사이에 문법을 따로 알려주지 않아도 알고리즘이 스스로 규칙을 파악하고 학습합니다.



\*문장정보를 숫자로 바꾼 벡터란?

현재 인공신경망 번역에선 보통 1000차원의 벡터를 씁니다. ‘I love him(나는 그를 사랑한다)’이라는 문장을 번역한다고 가정해 보겠습니다. 이 문장의 정보를 모두 담은 벡터를 1000차원 공간 안의 한 점으로 놓으면 ‘He is by loved me(그는 나에게 사랑을 받는다)’라는 문장 정보를 담은 벡터는 번역기에 입력하면 같은 뜻을 담고 있기 때문에 자동으로 ‘I love him’ 벡터와 아주 가까운 위치에 놓이게 됩니다. 반면 ‘He loves me(그는 나를 사랑한다)’라는 문장 정보를 담은 벡터는 뜻이 다른 만큼 좀 멀리 떨어져 있게 됩니다. 이렇게 단어나 구 등이 공간에서 관계를 맺으며 맵핑됩니다. 또 다른 예로 ‘cat(고양이)’라는 단어 벡터가 1000차원 공간 안에 위치해 있다면, ‘kitten(새끼 고양이)’라는 단어 벡터는 그 옆에 위치한다고 볼 수 있습니다. ‘dog(개)’ 벡터와 ‘puppy(강아지)’ 벡터는 비슷한 위치에 나란히 존재할 것입니다.



인코더와 디코더 작동방식

Nmt 종류 (rnn, transformer, attention)