1.LSTM 기반 Seq2Seq 모델에서 디코딩할 때 사용하는 Beam Search 동작 방식에 대해서 설명.

- 개념

Greedy Decoding방식과 Exhaustive Search방식의 타협점이 되는 방식으로, 매 타임스텝마다 상위 k개(Beam Size)의 경우의 수만 고려하여 답변을 생성하는 알고리즘을 사용하는 방법이다. 여러 후보(beam width, 보통 k개)를 동시에 추적하며, 각 후보 시퀀스의 누적 확률을 계산하여 다음 시점으로 확장한다.

- 동작원리

1) Decoder 시작(초기 입력(SOS)) 후, 가능한 모든 다음 토큰의 확률을 계산확률이 높은 k개의 시퀀스를 선택하여 beam으로 유지.

2) 이후 각 beam을 확장(각 시퀀스마다 모든 단어를 붙여봄)하고, 확장된 모든 후보 중 다시 가장 높은 확률을 갖는 k개만 남김.

3) EOS까지 도달하거나 사전 정의된 길이까지 반복.

4) 최종적으로 가장 높은 확률을 가진 시퀀스가 최종 결과로 선택됨.

- 특징

1) 생성 속도와 성능의 Trade Off에 대한 타협점

2) 최선의 답변을 보장하는 방식은 아니지만, 그래도 Greedy한 방식보단 더 최선의 답변을 뱉어낼 가능성이 높은 방식임.

3) 시간 복잡도 또한 O(k*V*t)로, 완전 탐색보다 계산량을 훨씬 줄일 수 있음.

2.Seq2Seq with LSTM 모델은 Attention이 없던 시절 제안된 구조임. 기본 Seq2Seq 모델의 한계와 이후 Attention 메커니즘이 이 한계를 어떻게 보완했는지 설명.

- 한계

1) 하나의 고정된 크기의 벡터에 모든 정보를 압축하려고 하여 정보 손실 발생

2) Vanishing Gradient 문제로 입력 문장이 길어지면 품질이 떨어지는 현상 발생

- 보완

Attention : Encoder에서 출력 단어를 예측하는 매 시점마다, encoder에서의 전체 입력 문장을 다시 한 번 참고. 해당 시점에서 예측해야 할 단어와 연관이 있는 입력 단어 부분을 좀 더 집중(attention)해서 보게 됨

1) 문장 길이 제한 완화해 컨텍스트 벡터 대신 유연하게 입력 시퀀스 각 위치의 정보를 실시간으로 조합하여 기존의 정보 손실 감소

2) 디코더가 각 단어를 출력할 때 입력 시퀀스에서 관련 부분에 집중하기에 장기 의존성 개선하고 번역 품질이 향상됨