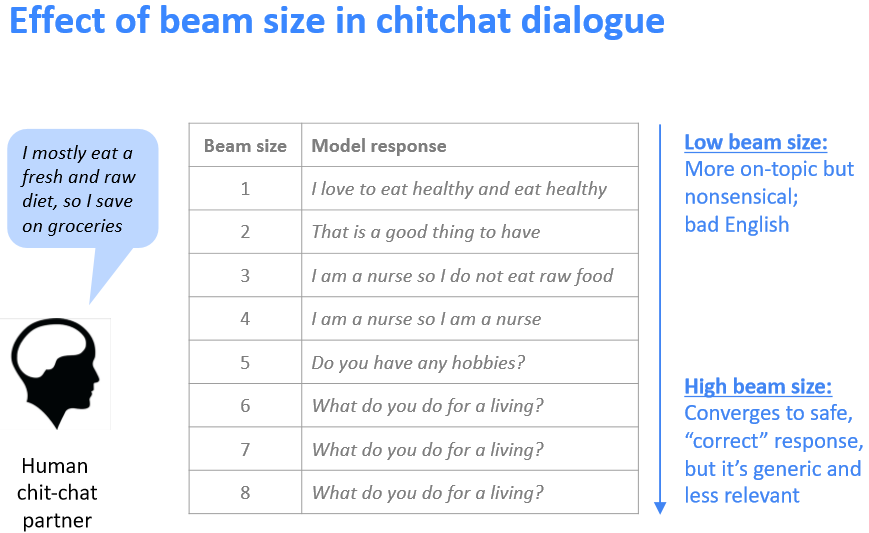


추가적인 Recap으로 우리는 디코딩 시 알고리즘으로 그리디 & 빔서치를 배웠다.

빔 사이즈 k 가 1이면 greedy search와 다를 것이 없고, 크다면 computationally expensive 하고, 너무 크다면 BLEU score를 감소시킨다.(Tu, Koehn 2017)

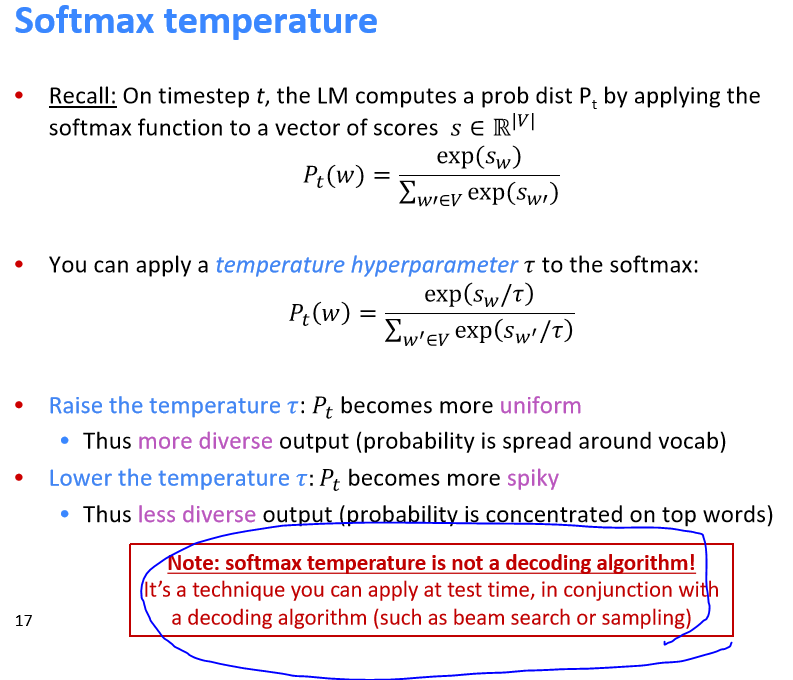
그 이유는 k가 크다면 too short translation을 하려고 하고, 때때로 chit-chat dialogue에서 너무 generic한 답변을 하려하기 때문이다.



따라서 새로운 방법이 제안 됐는데 그것이 Sampling-based decoding 이다.

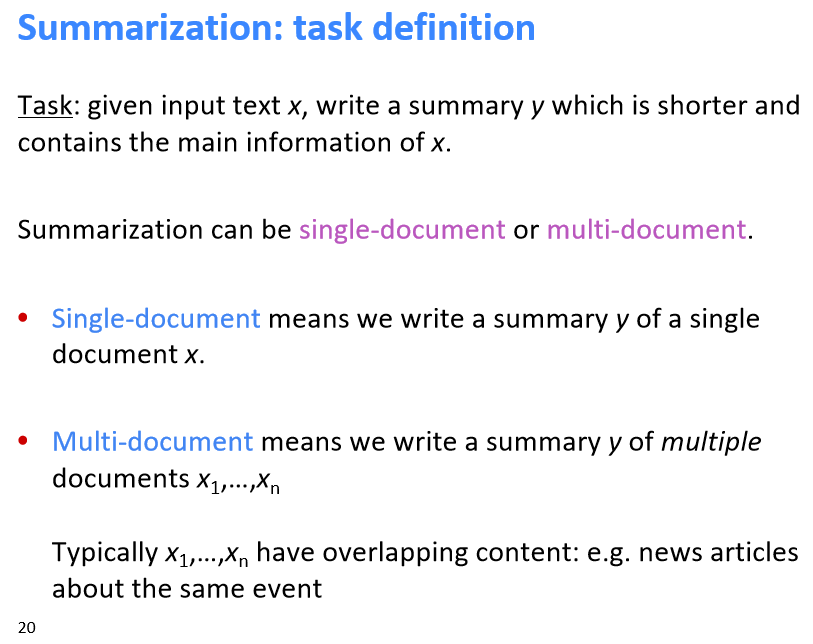
다음 단어를 고를 때 step t에서 확률 분포에서 랜덤하게 1개 고르는 것이다. Argmax의 파라미터가 아닌 sample을 이용하는 방식이다.(그리디한거 같지만 완전 그리디는 아님)

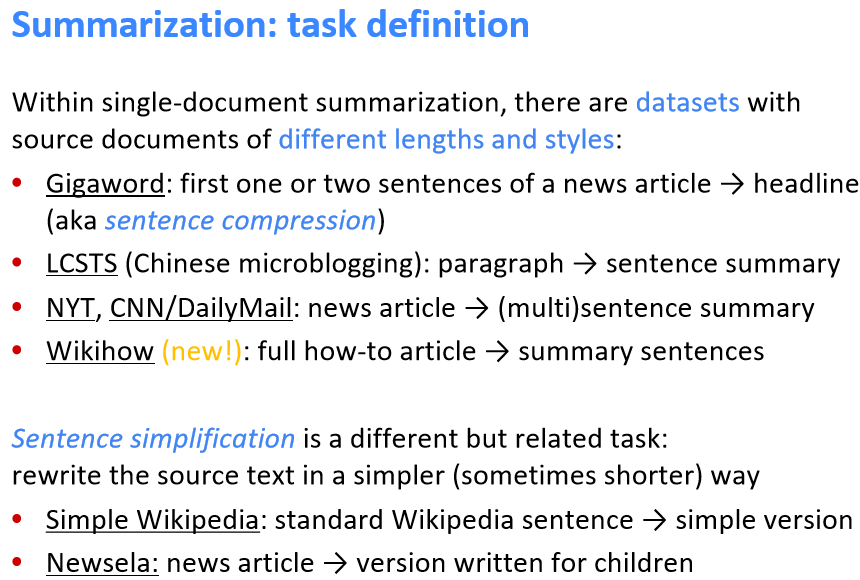
하지만 너무 랜덤하게 진행하면 문제가 될 수 있으니 표본을 top-n으로 줄이는 것이다.(n은 하이퍼 파라미터) 이때 n이 1이면 same greedy 이고 n이 커질수록 diverse하고 risky하며, n이 작을수록 generic하고 safe하다.(V면 pure sample)



Softmax를 변형시켜 사용하는 방법이다.

**Sec2: NLG Tasks & Neural Approaches to them**

라고하는 task이다.



라는 데이터셋들이 있고 비슷한 task인 sentence simplification이 있다.



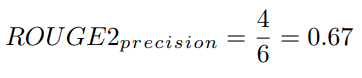
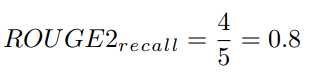
의 두 가지 summarization 방법이 있다.

Pre-Neural Summarization은 mostly extractive 였다.

1. Content selection: 문서에서 문장의 일부를 선택한다. (tf-idf등을 이용해 topic key words가 문장에 존재하는지를 따져서 or 문서에서 문장이 어느 곳에 위치한지를 따져서 // graph based 로 선택할 수도 있다.)
2. Information ordering: 문장마다 순서 매김(multi-documents 일 때 특히 중요)
3. Summary를 위한 마지막 단계: sentence realization

Summarization Evaluation에는 ROUGE가 있다. (Recall-Oriented Understudy for Gisting Evaluation)

BLEU 처럼 n-gram overlap에 기반한다.

“the cat was found under the bed” vs “the cat was under the bed”(gold)  
ROUGE-2를 이용한다면

(bi-gram으로 나누어서 분모, 분자 번갈아 가면서 계산)

ROUGE는 recall에 BLEU는 precision에 기반(둘 다 gold는 사람이 번역한 것)

ROUGE-1,2, L(LCS 기법으로 L을 구함.)이 많이 쓰임

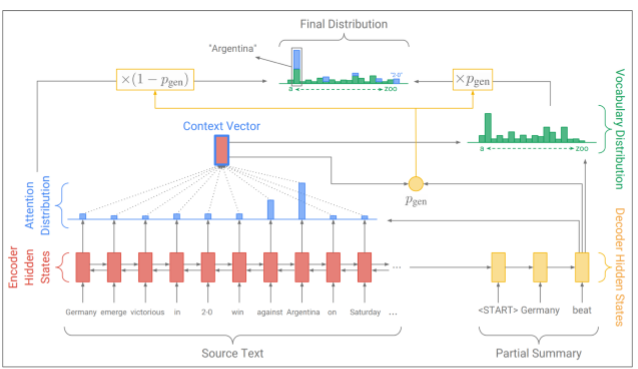
F1 Score로도 많이 쓰임.

Neural Summarization(2015) – seq2seq with attention을 활용한 summarization이 연구되었다. 이후 다양한 연구가 이루어지고 있다.

+취약했다. 따라서 copy mechanism이 seq2seq with attention에 도움을 주었다. (hybrid extractive/abstractive approach의 성능 향상)

아래 그림을 보면 일반 attention 처럼 context vector 생성하면  
1. Attention distribution(encoder states & decoder state) / 2. P-gen(context vector & decoder state) / 3. Vocab distribution(context vector & decoder state)

으로 마지막 그림을 보면 attention 부분은 카피할까? 초록 부분은 새로 생성할까? 의 비율 싸움이다.(그대로 카피가 좋을 경우도 있다.)



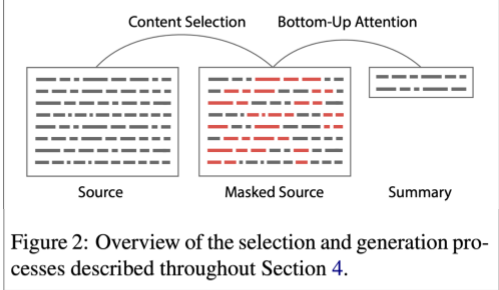
Problems: 긴 source text에 대해 너무 많이 copy 할 수도 있다. 또한 긴 input에 대해 전반적인 content selection(카피 기능 or 생성 기능을 실행하기 위해 일단 인풋 문장을 고르는 것?)이 잘 안된다.

즉, 여기서는 decoder가 surface realization을 attention이 word-level content selection 역할을 해서 summarization이 진행 됐지만 global content selection(문장을 전체적으로 선택 못함.)이 잘 되지 않았다.

따라서 제안된 것이 Bottom-up Summarization이다.

1. Content selection stage: neural sequence-tagging model이 모든 words에 include or don’t-include tag를 한다.
2. Bottom-up attention stage: seq2seq+attention이 don’t-include word는 mask한 뒤에 summary를 한다.

=> 간단한 구조지만 효과적이다: overall content selection이 되며 less copying of long sequences (more abstractive)



이번엔 NLG(Natural Language Generation)의 다른 Task인 Dialogue이다.