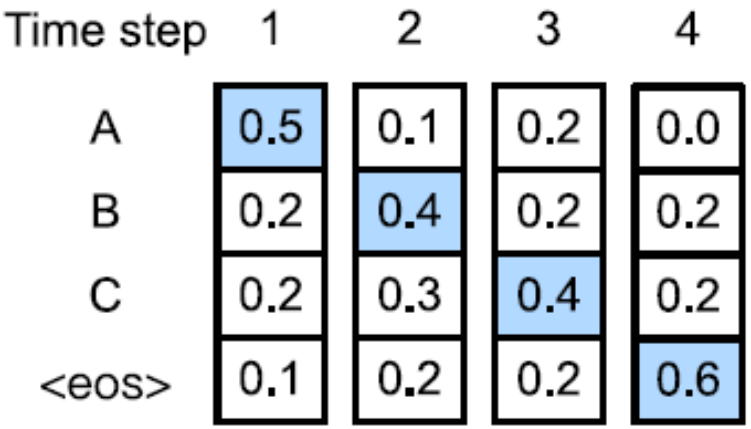
16-B. 한국어 번역 프로그램

16-B 강의 내용

- Beam search: 번역 과정에서 문장을 선택하는 방식
- 영어-한국어 병렬 데이터
- Attention을 적용한 Tensorflow 번역 프로그램

번역 문장의 탐색 방식

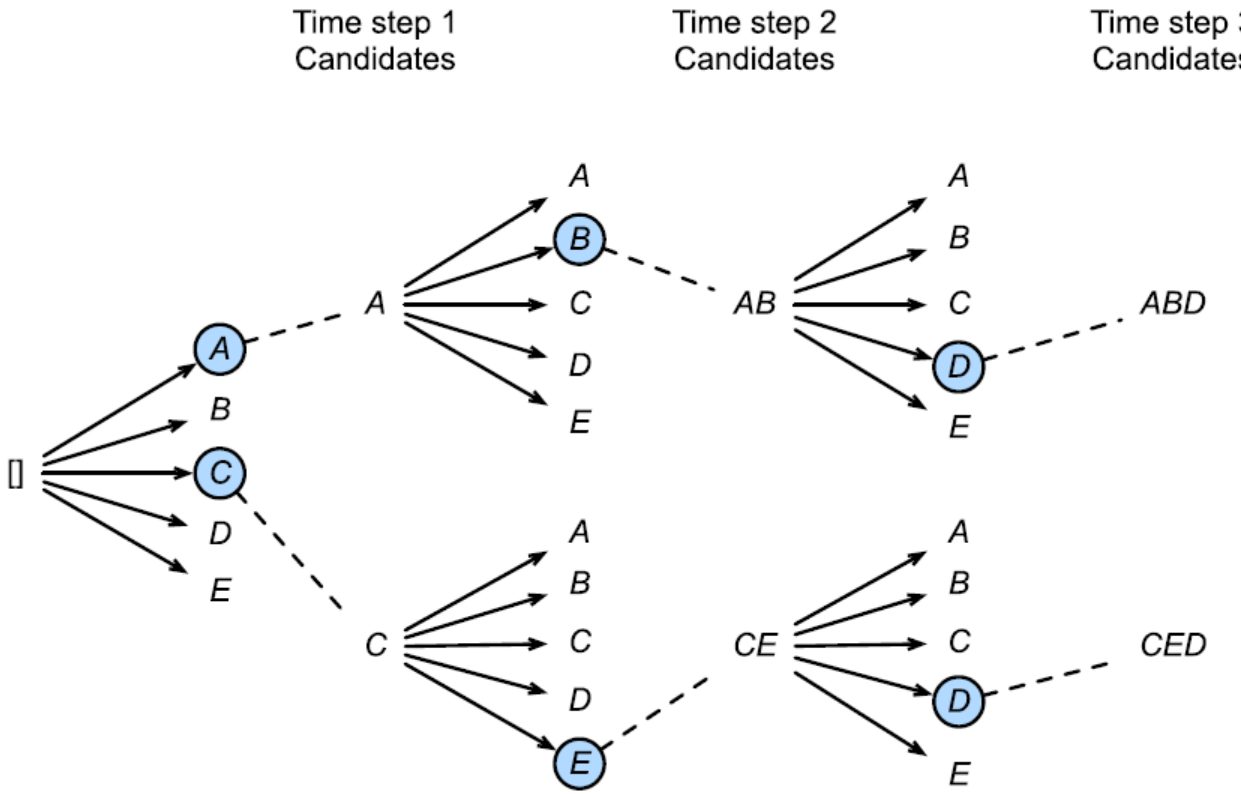
- 현재 상태에서 가장 적합한 번역 문장을 찾아내는 방식을 greedy search 라고 함
- Greedy search 방식은 최적의 문장을 찾지 못할 수 있음
- 모든 가능성을 탐색하는 exhaustive search 방식은 경제적이지 못함



Greedy search

Beam search

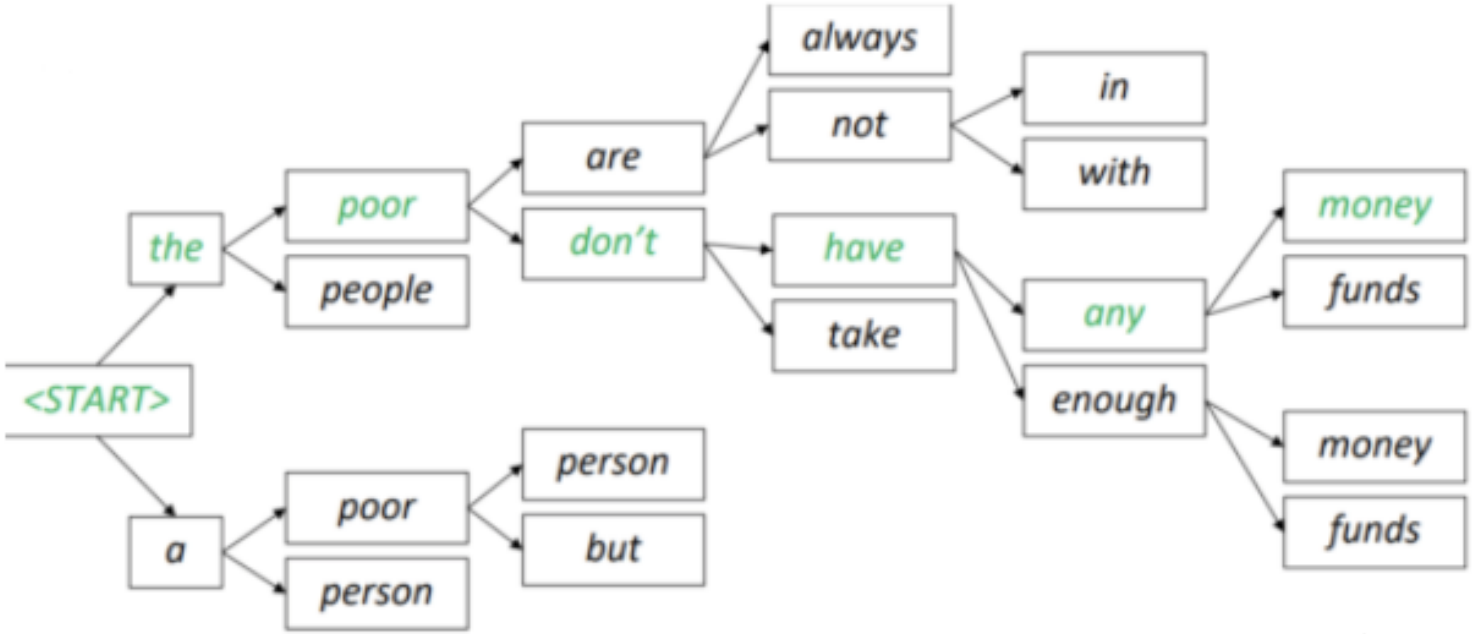
- 각 단계에서 일정한 개수의 문장을 추적하는 beam search 방식이 많이 사용됨
- 탐색하는 문장의 개수를 beam size 라고 함
- 각 beam에 대해 적합성을 게산하고 최종적으로 가장 점수가 높은 문장을 선택



beam size가 2인 사례

Beam search 동작 사례

- 일반적으로 10 이하의 beam size를 적용



영어-한국어 병렬 데이터

- http://opus.nlpl.eu/OpenSubtitles-v2018.php
  - TED 영어 강의와 한글 번역 파일들이 있음
  - 영화 자막을 한글로 번역한 내용도 있음(140만 문장)
  - 이들 문장은 영어와 한글이 다른 파일로 되어 있어서 문장별로 대응시켜야 함
- AI Hub(www.aihub.or.kr) 에서 제공되는 영한 말뭉치가 있음
  - [AI 데이터] – [교육/문화/스포츠] – [한국어-영어 번역 말뭉치]에서 찾을 수 있음

분야	설명	수량
뉴스	뉴스 텍스트	80만 문장
정부 웹사이트/저널	정부/지자체 홈페이지,간행물	10만 문장
법률	행정 규칙,자치 법규	10만 문장
한국문화	한국 역사,문화 콘텐츠	10만 문장
구어체	자연스러운 구어체 문장	40만 문장
대화체	상황/시나리오 기반 대화 세트	10만 문장
합계		160만 문장

AI Hub 데이터

- 각 데이터는 Excel 파일로 되어 있는데, 원문과 번역문을 txt 파일로 저장하여 사용하면 됨: utf-8 포맷으로 저장
- 전체 문장은 160만 개이므로 이중 일부를 사용하더라도 훈련 시간이 길어질 수 있음
- 한글과 영어 단어 구분을 미리 하지 않으면 어휘숫자가 많아서 적절한 결과를 기대하기 어려움

URL	언론사	원문	번역문
http://www.s	서울경제	스키너가 말한 보상은 대부분 눈으로 볼 수 있는 현물이다.	Skinner's reward is mostly eye-watering.
http://www.s	서울경제	심지어 어떤 문제가 발생할 건지도 어느 정도 예측이 가능하다.	Even some problems can be predicted.
http://news.k	국민일보	오직 하나님만이 그 이유를 제대로 알 수 있을 겁니다.	Only God will exactly know why.
http://news.k	국민일보	중국의 논쟁을 보며 간과해선 안 될 게 기업들의 고충이다.	Businesses should not overlook China's dispute.
http://news.k	국민일보	박자가 느린 노래는 오랜 시간이 지나 뜨는 경우가 있다.	Slow-beating songs often float over time.
http://www.h	한겨레	보험 처리가 안 되는 비급여 시술은 엄두도 못 낸다.	I can't even consider uninsured treatments.
http://news.k	국민일보	예수까지 합치면 모두 열세 명이 함께 식사를 하는 것이다.	Including Jesus, thirteen people eat together.
http://www.s	서울경제	인증을 받지 못한 기업은 정부가 만든 플랫폼을 활용해야 한다.	Uncertified companies should use government-created platforms.
http://www.s	서울경제	적어도 누군가 보고 싶은 일이 일어나진 않을 듯 합니다.	At least someone won't be missed.
http://www.h	한겨레	아이들 평가를 해보면 효과가 있다는 것을 알 수 있다.	Children's evaluations show that they work.
http://www.n	내일신문	어떤 학문이든지 일정의 성취를 이루기 위해서는 끊임없는 반복이 필요하다.	Any academic achievement requires constant repetition.
http://news.k	국민일보	정치적 논리가 개입되는 일은 결코 있어선 안 될 것이다.	Political logic should never be involved.
http://news.k	국민일보	슈퍼셀의 다섯 번째 게임이 오는 12일 세상에 모습을 드러낸다.	Supercell's fifth game appears on the 12th.

한국어-영어 번역 프로그램 구현

- 한국어 파일과 영어 파일에 대해 sentencepiece 등을 적용하여 단어숫자를 조정(보통 32,000 단어를 많이 사용)
- Mecab 등의 형태소분석기와 sentencepiece를 같이 적용하면 성능이 가장 좋아짐
- 훈련이 끝난 다음 번역기를 사용할 때도 Tokenizer/detokenizer를 활용해야 함

한국어-영어 문장 전처리

- 프로그램에서 `preprocess_sentence()` 함수를 다음과 같이 변경
- 한국어 문장에서는 숫자, 영어, 한국어만 입력되도록 수정

```
w = re.sub(r"[^0-9a-zA-Z가-힣?.!,¿]+", " ", w)
```

- 영어는 모두 소문자로 변경
- 마침표(`, . ? !`)를 제외한 특수문자는 제거

Tensorflow NMT with Attention

- Tensorflow의 Tutorial 중 하나로 encoder-decoder 구조에서 Attention 기능을 구현했음
- www.tensorflow.org/tutorials/text/nmt_with_attention
- 현재 버전은 영어-스페인어로 되어 있지만 파일을 대체하면 영어-한국어로 바꿀 수 있음
- kor-eng.ipynb 파일은 한국어에 맞게 일부 수정한 버전임
  - AI Hub 데이터에서 가져온 <구어체(1).txt> 파일을 사용

Tensorflow NMT 프로그램

- Encoder와 Decoder는 1,024 셀의 GRU로 구성했음
- 현재의 입력 데이터는 sentencepiece가 적용되지 않은 버전이라서 제대로 동작하지 않을 것임
- encoder.summary()는 다음과 같음

RNN → LSTM GRU

```
Encoder output shape: (batch size, sequence length, units) (64, 59, 1024)
Encoder Hidden state shape: (batch size, units) (64, 1024)
Model: "encoder"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        multiple                  9530368
_________________________________________________________________
gru (GRU)                    multiple                  3938304
=================================================================
Total params: 13,468,672
Trainable params: 13,468,672
Non-trainable params: 0
_________________________________________________________________
```

Tensorflow NMT 프로그램

- 현재 버전에서의 영어와 한글 단어수는 각각 37,228개와 200,928개임
- decoder.summary()는 다음과 같음: 단어숫자가 많아서 embedding과 dense 층의 파라미터 숫자가 너무 높음

```
Decoder output shape: (batch_size, vocab size) (64, 200928)
Model: "decoder"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      multiple                  51437568
_________________________________________________________________
gru_1 (GRU)                  multiple                  7084032
_________________________________________________________________
dense_3 (Dense)              multiple                  205951200
_________________________________________________________________
bahdanau_attention_1 (Bahdan multiple                  2100225
=================================================================
Total params: 266,573,025
Trainable params: 266,573,025
Non-trainable params: 0
_________________________________________________________________
```