CS236299 Project Segment 4: Semantic Interpretation – Question Answering

August 11, 2023

[ ]:
```
from google.colab import drive
drive.mount('/content/drive')
# my files are in 'labs/lab0-0'
# !cp -r /content/drive/MyDrive/Colab-Notebooks/NLP/project4/* .
!cp -r /content/drive/MyDrive/NLP_labs/project4/* .
!pip install -r requirements.txt
# restart the runtime
import os
os._exit(00)
```

```
Mounted at /content/drive
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages
(from -r requirements.txt (line 1)) (3.8.1)
Requirement already satisfied: cryptography in /usr/local/lib/python3.10/dist-
packages (from -r requirements.txt (line 2)) (41.0.3)
Requirement already satisfied: torch>=1.9.0 in /usr/local/lib/python3.10/dist-
packages (from -r requirements.txt (line 3)) (2.0.1+cu118)
Collecting func_timeout (from -r requirements.txt (line 4))
  Downloading func_timeout-4.3.5.tar.gz (44 kB)
                           44.3/44.3 kB
1.5 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Collecting otter-grader==1.0.0 (from -r requirements.txt (line 5))
  Downloading otter_grader-1.0.0-py3-none-any.whl (163 kB)
                           164.0/164.0

kB 7.1 MB/s eta 0:00:00
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from -r requirements.txt (line 6)) (4.65.0)
Requirement already satisfied: pandas==1.5.3 in /usr/local/lib/python3.10/dist-
packages (from -r requirements.txt (line 7)) (1.5.3)
Collecting transformers==4.26.0 (from -r requirements.txt (line 8))
  Downloading transformers-4.26.0-py3-none-any.whl (6.3 MB)
                           6.3/6.3 MB
75.2 MB/s eta 0:00:00
Collecting tokenizers==0.13.2 (from -r requirements.txt (line 9))
  Downloading
```

```
tokenizers-0.13.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(7.6 MB)
                           7.6/7.6 MB
111.2 MB/s eta 0:00:00
Collecting datasets==2.9.0 (from -r requirements.txt (line 10))
  Downloading datasets-2.9.0-py3-none-any.whl (462 kB)
                          462.8/462.8 kB
33.2 MB/s eta 0:00:00
Collecting wget (from -r requirements.txt (line 11))
  Downloading wget-3.2.zip (10 kB)
  Preparing metadata (setup.py) … done
Requirement already satisfied: pyyaml in /usr/local/lib/python3.10/dist-packages
(from otter-grader==1.0.0->-r requirements.txt (line 5)) (6.0.1)
Requirement already satisfied: nbformat in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (5.9.2)
Requirement already satisfied: ipython in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (7.34.0)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (6.5.4)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (67.7.2)
Requirement already satisfied: tornado in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (6.3.1)
Collecting docker (from otter-grader==1.0.0->-r requirements.txt (line 5))
  Downloading docker-6.1.3-py3-none-any.whl (148 kB)
                          148.1/148.1 kB
17.6 MB/s eta 0:00:00
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-
packages (from otter-grader==1.0.0->-r requirements.txt (line 5)) (3.1.2)
Collecting dill (from otter-grader==1.0.0->-r requirements.txt (line 5))
  Downloading dill-0.3.7-py3-none-any.whl (115 kB)
                          115.3/115.3 kB
15.0 MB/s eta 0:00:00
Collecting pdfkit (from otter-grader==1.0.0->-r requirements.txt (line 5))
  Downloading pdfkit-1.0.0-py3-none-any.whl (12 kB)
Collecting PyPDF2 (from otter-grader==1.0.0->-r requirements.txt (line 5))
  Downloading pypdf2-3.0.1-py3-none-any.whl (232 kB)
                          232.6/232.6 kB
23.9 MB/s eta 0:00:00
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas==1.5.3->-r requirements.txt
(line 7)) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas==1.5.3->-r requirements.txt (line 7)) (2022.7.1)
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-
packages (from pandas==1.5.3->-r requirements.txt (line 7)) (1.23.5)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-
packages (from transformers==4.26.0->-r requirements.txt (line 8)) (3.12.2)
```

```
Collecting huggingface-hub<1.0,>=0.11.0 (from transformers==4.26.0->-r
requirements.txt (line 8))
  Downloading huggingface_hub-0.16.4-py3-none-any.whl (268 kB)
                          268.8/268.8 kB
29.9 MB/s eta 0:00:00
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.26.0->-r
requirements.txt (line 8)) (23.1)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.10/dist-packages (from transformers==4.26.0->-r
requirements.txt (line 8)) (2022.10.31)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from transformers==4.26.0->-r requirements.txt (line 8)) (2.31.0)
Requirement already satisfied: pyarrow>=6.0.0 in /usr/local/lib/python3.10/dist-
packages (from datasets==2.9.0->-r requirements.txt (line 10)) (9.0.0)
Collecting dill (from otter-grader==1.0.0->-r requirements.txt (line 5))
  Downloading dill-0.3.6-py3-none-any.whl (110 kB)
                          110.5/110.5 kB
12.7 MB/s eta 0:00:00
Collecting xxhash (from datasets==2.9.0->-r requirements.txt (line 10))
  Downloading
xxhash-3.3.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                          194.1/194.1 kB
21.1 MB/s eta 0:00:00
Collecting multiprocess (from datasets==2.9.0->-r requirements.txt (line
10))
  Downloading multiprocess-0.70.15-py310-none-any.whl (134 kB)
                          134.8/134.8 kB
16.4 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]>=2021.11.1 in
/usr/local/lib/python3.10/dist-packages (from datasets==2.9.0->-r
requirements.txt (line 10)) (2023.6.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-
packages (from datasets==2.9.0->-r requirements.txt (line 10)) (3.8.5)
Collecting responses<0.19 (from datasets==2.9.0->-r requirements.txt (line 10))
  Downloading responses-0.18.0-py3-none-any.whl (38 kB)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages
(from nltk->-r requirements.txt (line 1)) (8.1.6)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages
(from nltk->-r requirements.txt (line 1)) (1.3.1)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.10/dist-
packages (from cryptography->-r requirements.txt (line 2)) (1.15.1)
Requirement already satisfied: typing-extensions in
/usr/local/lib/python3.10/dist-packages (from torch>=1.9.0->-r requirements.txt
(line 3)) (4.7.1)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages
(from torch>=1.9.0->-r requirements.txt (line 3)) (1.11.1)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-
```

```
packages (from torch>=1.9.0->-r requirements.txt (line 3)) (3.1)
Requirement already satisfied: triton==2.0.0 in /usr/local/lib/python3.10/dist-
packages (from torch>=1.9.0->-r requirements.txt (line 3)) (2.0.0)
Requirement already satisfied: cmake in /usr/local/lib/python3.10/dist-packages
(from triton==2.0.0->torch>=1.9.0->-r requirements.txt (line 3)) (3.25.2)
Requirement already satisfied: lit in /usr/local/lib/python3.10/dist-packages
(from triton==2.0.0->torch>=1.9.0->-r requirements.txt (line 3)) (16.0.6)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-
packages (from cffi>=1.12->cryptography->-r requirements.txt (line 2)) (2.21)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets==2.9.0->-r requirements.txt (line 10)) (23.1.0)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets==2.9.0->-r
requirements.txt (line 10)) (3.2.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets==2.9.0->-r
requirements.txt (line 10)) (6.0.4)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets==2.9.0->-r
requirements.txt (line 10)) (4.0.2)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-
packages (from aiohttp->datasets==2.9.0->-r requirements.txt (line 10)) (1.9.2)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets==2.9.0->-r
requirements.txt (line 10)) (1.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.10/dist-packages (from aiohttp->datasets==2.9.0->-r
requirements.txt (line 10)) (1.3.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.1->pandas==1.5.3->-r requirements.txt (line
7)) (1.16.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->transformers==4.26.0->-r requirements.txt (line 8))
(3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers==4.26.0->-r
requirements.txt (line 8)) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests->transformers==4.26.0->-r
requirements.txt (line 8)) (2023.7.22)
Requirement already satisfied: websocket-client>=0.32.0 in
/usr/local/lib/python3.10/dist-packages (from docker->otter-grader==1.0.0->-r
requirements.txt (line 5)) (1.6.1)
Collecting jedi>=0.16 (from ipython->otter-grader==1.0.0->-r requirements.txt
(line 5))
  Downloading jedi-0.19.0-py2.py3-none-any.whl (1.6 MB)
                           1.6/1.6 MB
69.6 MB/s eta 0:00:00
```

```
Requirement already satisfied: decorator in
/usr/local/lib/python3.10/dist-packages (from ipython->otter-grader==1.0.0->-r
requirements.txt (line 5)) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-
packages (from ipython->otter-grader==1.0.0->-r requirements.txt (line 5))
(0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-
packages (from ipython->otter-grader==1.0.0->-r requirements.txt (line 5))
(5.7.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipython->otter-grader==1.0.0->-r
requirements.txt (line 5)) (3.0.39)
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-
packages (from ipython->otter-grader==1.0.0->-r requirements.txt (line 5))
(2.14.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-
packages (from ipython->otter-grader==1.0.0->-r requirements.txt (line 5))
(0.2.0)
Requirement already satisfied: matplotlib-inline in
/usr/local/lib/python3.10/dist-packages (from ipython->otter-grader==1.0.0->-r
requirements.txt (line 5)) (0.1.6)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-
packages (from ipython->otter-grader==1.0.0->-r requirements.txt (line 5))
(4.8.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->otter-grader==1.0.0->-r
requirements.txt (line 5)) (2.1.3)
INFO: pip is looking at multiple versions of multiprocess to determine which
version is compatible with other requirements. This could take a while.
Collecting multiprocess (from datasets==2.9.0->-r requirements.txt (line 10))
  Downloading multiprocess-0.70.14-py310-none-any.whl (134 kB)
                          134.3/134.3 kB
13.9 MB/s eta 0:00:00
Requirement already satisfied: lxml in /usr/local/lib/python3.10/dist-
packages (from nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5))
(4.9.3)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-
packages (from nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5))
(4.11.2)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages
(from nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5)) (6.0.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.10/dist-
packages (from nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5))
(0.7.1)
Requirement already satisfied: entrypoints>=0.2.2 in
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (0.4)
Requirement already satisfied: jupyter-core>=4.7 in
```

```
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (5.3.1)
Requirement already satisfied: jupyterlab-pygments in
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (0.2.2)
Requirement already satisfied: mistune<2,>=0.8.1 in
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (0.8.4)
Requirement already satisfied: nbclient>=0.5.0 in
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (0.8.0)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.10/dist-packages (from nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (1.5.0)
Requirement already satisfied: tinycss2 in /usr/local/lib/python3.10/dist-
packages (from nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5))
(1.2.1)
Requirement already satisfied: fastjsonschema in /usr/local/lib/python3.10/dist-
packages (from nbformat->otter-grader==1.0.0->-r requirements.txt (line 5))
(2.18.0)
Requirement already satisfied: jsonschema>=2.6 in
/usr/local/lib/python3.10/dist-packages (from nbformat->otter-grader==1.0.0->-r
requirements.txt (line 5)) (4.3.3)
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-
packages (from sympy->torch>=1.9.0->-r requirements.txt (line 3)) (1.3.0)
Requirement already satisfied: parso<0.9.0,>=0.8.3 in
/usr/local/lib/python3.10/dist-packages (from jedi>=0.16->ipython->otter-
grader==1.0.0->-r requirements.txt (line 5)) (0.8.3)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
/usr/local/lib/python3.10/dist-packages (from jsonschema>=2.6->nbformat->otter-
grader==1.0.0->-r requirements.txt (line 5)) (0.19.3)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.10/dist-packages (from jupyter-
core>=4.7->nbconvert->otter-grader==1.0.0->-r requirements.txt (line 5))
(3.10.0)
Requirement already satisfied: jupyter-client>=6.1.12 in
/usr/local/lib/python3.10/dist-packages (from nbclient>=0.5.0->nbconvert->otter-
grader==1.0.0->-r requirements.txt (line 5)) (6.1.12)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.10/dist-packages (from pexpect>4.3->ipython->otter-
grader==1.0.0->-r requirements.txt (line 5)) (0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.10/dist-
packages (from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython->otter-
grader==1.0.0->-r requirements.txt (line 5)) (0.2.6)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-
packages (from beautifulsoup4->nbconvert->otter-grader==1.0.0->-r
requirements.txt (line 5)) (2.4.1)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-
```

```
packages (from bleach->nbconvert->otter-grader==1.0.0->-r requirements.txt (line
5)) (0.5.1)
Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.10/dist-
packages (from jupyter-client>=6.1.12->nbclient>=0.5.0->nbconvert->otter-
grader==1.0.0->-r requirements.txt (line 5)) (23.2.1)
Building wheels for collected packages: func_timeout, wget
  Building wheel for func_timeout (setup.py) … done
  Created wheel for func_timeout: filename=func_timeout-4.3.5-py3-none-any.whl
size=15078
sha256=a89c4eae60c75d3f7b8ec6fc89fe830582681a5caf48b8243a9774a99817c8a2
  Stored in directory: /root/.cache/pip/wheels/3f/83/19/b5552bb9630e353f7c5b15be
44bf10900afe1abbbfcf536afd
  Building wheel for wget (setup.py) … done
  Created wheel for wget: filename=wget-3.2-py3-none-any.whl size=9656
sha256=f64c2874da33f0855c5620b589a3b16178019e842df1c9836e49caf616f7e5c8
  Stored in directory: /root/.cache/pip/wheels/8b/f1/7f/5c94f0a7a505ca1c81cd1d92
08ae2064675d97582078e6c769
Successfully built func_timeout wget
Installing collected packages: wget, tokenizers, pdfkit, func_timeout, xxhash,
PyPDF2, jedi, dill, responses, multiprocess, huggingface-hub, docker,
transformers, datasets, otter-grader
Successfully installed PyPDF2-3.0.1 datasets-2.9.0 dill-0.3.6 docker-6.1.3
func_timeout-4.3.5 huggingface-hub-0.16.4 jedi-0.19.0 multiprocess-0.70.14
otter-grader-1.0.0 pdfkit-1.0.0 responses-0.18.0 tokenizers-0.13.2
transformers-4.26.0 wget-3.2 xxhash-3.3.0
```

[1]:

```
# Please do not change this cell because some hidden tests might depend on it.
import os

# Otter grader does not handle ! commands well, so we define and use our
# own function to execute shell commands.
def shell(commands, warn=True):
    """Executes the string `commands` as a sequence of shell commands.

       Prints the result to stdout and returns the exit status.
       Provides a printed warning on non-zero exit status unless `warn`
       flag is unset.
    """
    file = os.popen(commands)
    print (file.read().rstrip('\n'))
    exit_status = file.close()
    if warn and exit_status != None:
        print(f"Completed with errors. Exit status: {exit_status}\n")
    return exit_status

shell("""
ls requirements.txt >/dev/null 2>&1
```

```
if [ ! $? = 0 ]; then
 rm -rf .tmp
 git clone https://github.com/cs236299-2023-spring/project4.git .tmp
 mv .tmp/requirements.txt ./
 rm -rf .tmp
fi
pip install -q -r requirements.txt
""")
```

[2]:
```
# Initialize Otter
import otter
grader = otter.Notebook()
```

1 236299 - Introduction to Natural Language Processing

1.1 Project 4: Semantic Interpretation – Question Answering

The goal of semantic parsing is to convert natural language utterances to a meaning representation such as a *logical form* expression or a *SQL query*. In the previous project segment, you built a parsing system to reconstruct parse trees from the natural-language queries in the ATIS dataset. However, that only solves an intermediary task, not the end-user task of obtaining answers to the queries.

In this final project segment, you will go further, building a semantic parsing system to convert English queries to SQL queries, so that by consulting a database you will be able to answer those questions. You will implement both a rule-based approach and an end-to-end sequence-to-sequence (seq2seq) approach. Both algorithms come with their pros and cons, and by the end of this segment you should have a basic understanding of the characteristics of the two approaches.

1.2 Goals

1. Build a semantic parsing algorithm to convert text to SQL queries based on the syntactic parse trees from the last project.
2. Build an attention-based end-to-end seq2seq system to convert text to SQL.
3. Improve the attention-based end-to-end seq2seq system with self-attention to convert text to SQL.
4. Discuss the pros and cons of the rule-based system and the end-to-end system.
5. (Optional) Use the state-of-the-art pretrained transformers for text-to-SQL conversion.

This will be an extremely challenging project, so we recommend that you start early.

2 Setup

[3]:
```python
import copy
import datetime
import math
import re
import sys
import warnings

import wget
import nltk
import sqlite3
import csv
import torch
import torch.nn as nn
import datasets

from datasets import load_dataset
from tokenizers import Tokenizer
from tokenizers import Regex
from tokenizers.pre_tokenizers import WhitespaceSplit, Split
from tokenizers.processors import TemplateProcessing
from tokenizers import normalizers
from tokenizers.models import WordLevel
from tokenizers.trainers import WordLevelTrainer
from transformers import PreTrainedTokenizerFast

from cryptography.fernet import Fernet
from func_timeout import func_set_timeout
from torch.nn.utils.rnn import pack_padded_sequence as pack
from torch.nn.utils.rnn import pad_packed_sequence as unpack
from tqdm import tqdm
from transformers import BartTokenizer, BartForConditionalGeneration
```

[4]:
```python
# Set random seeds
seed = 1234
torch.manual_seed(seed)
# Set timeout for executing SQL
TIMEOUT = 3 # seconds

# GPU check: Set runtime type to use GPU where available
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
print (device)
```

```
cuda
```

[5]:
```python
## Download needed scripts and data
def download_if_needed(source, dest, filename):
    os.path.exists(f"./{dest}{filename}") or wget.download(source + filename, out=dest)

os.makedirs('data', exist_ok=True)
os.makedirs('scripts', exist_ok=True)
source_url = "https://raw.githubusercontent.com/nlp-236299/data/master"

# Grammar to augment for this segment
if not os.path.isfile('data/grammar'):
  download_if_needed(source_url, "data/", "/ATIS/grammar_distrib4.crypt")

  # Decrypt the grammar file
  key = b'bfksTY2BJ5VKKK9xZb1PDDLaGkdu7KCDFYfVePSEfGY='
  fernet = Fernet(key)
  with open('./data/grammar_distrib4.crypt', 'rb') as f:
    restored = Fernet(key).decrypt(f.read())
  with open('./data/grammar', 'wb') as f:
    f.write(restored)

# Download scripts and ATIS database
download_if_needed(source_url, "scripts/", "/scripts/trees/transform.py")
download_if_needed(source_url, "data/", "/ATIS/atis_sqlite.db")
```

[6]:
```python
# Import downloaded scripts for parsing augmented grammars
sys.path.insert(1, './scripts')
import transform as xform
```

3 Semantically augmented grammars

In the first part of this project segment, you'll be implementing a rule-based system for semantic interpretation of sentences. Before jumping into using such a system on the ATIS dataset – we'll get to that soon enough – let's first work with some trivial examples to get things going.

The fundamental idea of rule-based semantic interpretation is the rule of compositionality, that *the meaning of a constituent is a function of the meanings of its immediate subconstituents and the syntactic rule that combined them.* This leads to an infrastructure for specifying semantic interpretation in which each syntactic rule in a grammar (in our case, a context-free grammar) is associated with a semantic rule that applies to the meanings associated with the elements on the right-hand side of the rule.

3.1 Example: arithmetic expressions

As a first example, let's consider an augmented grammar for arithmetic expressions, familiar from lab 3-1. We again use the function `xform.parse_augmented_grammar` to parse the augmented grammar. You can read more about it in the file `scripts/transform.py`.

[10]:
```
arithmetic_grammar, arithmetic_augmentations = xform.parse_augmented_grammar(
    """
    ## Sample grammar for arithmetic expressions

    S -> NUM                                : lambda Num: Num
       | S OP S                             : lambda S1, Op, S2: Op(S1, S2)

    OP -> ADD                               : lambda Op: Op
        | SUB
        | MULT
        | DIV

    NUM -> 'zero'                           : lambda: 0
         | 'one'                            : lambda: 1
         | 'two'                            : lambda: 2
         | 'three'                          : lambda: 3
         | 'four'                           : lambda: 4
         | 'five'                           : lambda: 5
         | 'six'                            : lambda: 6
         | 'seven'                          : lambda: 7
         | 'eight'                          : lambda: 8
         | 'nine'                           : lambda: 9
         | 'ten'                            : lambda: 10

    ADD -> 'plus' | 'added' 'to'            : lambda: lambda x, y: x + y
    SUB -> 'minus'                          : lambda: lambda x, y: x - y
    MULT -> 'times' | 'multiplied' 'by'     : lambda: lambda x, y: x * y
    DIV -> 'divided' 'by'                   : lambda: lambda x, y: x / y
    """
)
```

Recall that in this grammar specification format, rules that are not explicitly provided with an augmentation (like all the `OP` rules after the first `OP -> ADD`) are associated with the textually most recent one (`lambda Op: Op`).

The `parse_augmented_grammar` function returns both an NLTK grammar and a dictionary that maps from productions in the grammar to their associated augmentations. Let's examine the returned grammar.

[11]:
```
for production in arithmetic_grammar.productions():
  print(f"{repr(production):25}     {arithmetic_augmentations[production]}")
```

```
S -> NUM                     <function <lambda> at 0x7a57f15a6f80>
S -> S OP S                  <function <lambda> at 0x7a57f15a6ef0>
OP -> ADD                    <function <lambda> at 0x7a57f15a6d40>
OP -> SUB                    <function <lambda> at 0x7a57f15a63b0>
OP -> MULT                   <function <lambda> at 0x7a57f15a7e20>
OP -> DIV                    <function <lambda> at 0x7a57f15a6c20>
```

```
NUM -> 'zero'                   <function <lambda> at 0x7a57f15a5ab0>
NUM -> 'one'                    <function <lambda> at 0x7a57f15a7eb0>
NUM -> 'two'                    <function <lambda> at 0x7a57f15a5fc0>
NUM -> 'three'                  <function <lambda> at 0x7a57f15a6cb0>
NUM -> 'four'                   <function <lambda> at 0x7a57f15a6320>
NUM -> 'five'                   <function <lambda> at 0x7a57f15a5b40>
NUM -> 'six'                    <function <lambda> at 0x7a57f15a5d80>
NUM -> 'seven'                  <function <lambda> at 0x7a57f15a5bd0>
NUM -> 'eight'                  <function <lambda> at 0x7a57f15a7f40>
NUM -> 'nine'                   <function <lambda> at 0x7a593034cc10>
NUM -> 'ten'                    <function <lambda> at 0x7a593034c940>
ADD -> 'plus'                   <function <lambda> at 0x7a59202d1120>
ADD -> 'added' 'to'             <function <lambda> at 0x7a5930858550>
SUB -> 'minus'                  <function <lambda> at 0x7a5930858940>
MULT -> 'times'                 <function <lambda> at 0x7a5930317c70>
MULT -> 'multiplied' 'by'       <function <lambda> at 0x7a5930317f40>
DIV -> 'divided' 'by'           <function <lambda> at 0x7a5930317a30>
```

We can parse with the grammar using one of the built-in NLTK parsers.

[12]:
```python
arithmetic_parser = nltk.parse.BottomUpChartParser(arithmetic_grammar)
parses = [p for p in arithmetic_parser.parse('three plus one times four'.
 ↪split())]
for parse in parses:
  parse.pretty_print()
```

```
            S
        ____|_________
       S         |    |
   ____|____     |    |
  S    OP   S    OP   S
  |    |    |    |    |
 NUM  ADD  NUM  MULT NUM
  |    |    |    |    |
three plus one times four

            S
   _________|____
  |    |         S
  |    |     ____|____
  S    OP   S    OP   S
  |    |    |    |    |
 NUM  ADD  NUM  MULT NUM
  |    |    |    |    |
three plus one times four
```

Now let's turn to the augmentations. They can be arbitrary Python functions applied to the semantic representations associated with the right-hand-side nonterminals, returning the semantic representation of the left-hand side. To interpret the semantic representation of the entire sentence (at the root of the parse tree), we can use the following pseudo-code:

```
to interpret a tree:
  interpret each of the nonterminal-rooted subtrees
  find the augmentation associated with the root production of the tree
    (it should be a function of as many arguments as there are nonterminals on the right-hand s
  return the result of applying the augmentation to the subtree values
```

(The base case of this recursion occurs when the number of nonterminal-rooted subtrees is zero, that is, a rule all of whose right-hand side elements are terminals.)

Suppose we had such a function, call it `interpret`. How would it operate on, for instance, the tree `(S (S (NUM three)) (OP (ADD plus)) (S (NUM one)))`?

```
interpret (S (S (NUM three)) (OP (ADD plus)) (S (NUM one)))
    |->interpret (S (NUM three))
    |      |->interpret (NUM three)
    |      |      |->(no subconstituents to evaluate)
    |      |      |->apply the augmentation for the rule NUM -> three to the empty set of valu
    |      |      |      (lambda: 3) () ==> 3
    |      |      \==> 3
    |      |->apply the augmentation for the rule S -> NUM to the value 3
    |      |      (lambda NUM: NUM)(3) ==> 3
    |      \==> 3
    |->interpret (OP (ADD plus))
    |      |...
    |      \==> lambda x, y: x + y
    |->interpret (S (NUM one))
    |      |...
    |      \==> 1
    |->apply the augmentation for the rule S -> S OP S to the values 3, (lambda x, y: x + y), a
    |      (lambda S1, Op, S2: Op(S1, S2))(3, (lambda x, y: x + y), 1) ==> 4
    \==> 4
```

Thus, the string "three plus one" is semantically interpreted as the value 4.

We provide the `interpret` function to carry out this recursive process, copied over from lab 4-2:

[13]:
```python
def interpret(tree, augmentations):
  syntactic_rule = tree.productions()[0]
  semantic_rule = augmentations[syntactic_rule]
  child_meanings = [interpret(child, augmentations)
                    for child in tree
                    if isinstance(child, nltk.Tree)]
  return semantic_rule(*child_meanings)
```

Now we should be able to evaluate the arithmetic example from above.

[14]:
```python
interpret(parses[0], arithmetic_augmentations)
```

[14]: 16

And we can even write a function that parses and interprets a string. We'll have it evaluate each of the possible parses and print the results.

[15]:
```
def parse_and_interpret(string, grammar, augmentations):
  parser = nltk.parse.BottomUpChartParser(grammar)
  parses = parser.parse(string.split())
  for parse in parses:
    parse.pretty_print()
    print(parse, "==>", interpret(parse, augmentations))
```

[16]:
```
parse_and_interpret("three plus one times four", arithmetic_grammar,␣
↪arithmetic_augmentations)
```

```
            S
        ____|_________
       S         |    |
   ____|____     |    |
  S    OP   S    OP   S
  |    |    |    |    |
 NUM  ADD  NUM  MULT NUM
  |    |    |    |    |
three plus one times four

(S
  (S (S (NUM three)) (OP (ADD plus)) (S (NUM one)))
  (OP (MULT times))
  (S (NUM four))) ==> 16
            S
   _________|____
  |    |         S
  |    |     ____|____
  S    OP   S    OP   S
  |    |    |    |    |
 NUM  ADD  NUM  MULT NUM
  |    |    |    |    |
three plus one times four

(S
  (S (NUM three))
  (OP (ADD plus))
  (S (S (NUM one)) (OP (MULT times)) (S (NUM four)))) ==> 7
```

Since the string is syntactically ambiguous according to the grammar, it is semantically ambiguous as well.

3.2 Some grammar specification conveniences

Before going on, it will be useful to have a few more conveniences in writing augmentations for rules. First, since the augmentations are arbitrary Python expressions, they can be built from and make use of other functions. For instance, you'll notice that many of the augmentations at the leaves of the tree took no arguments and returned a constant. We can define a function `constant` that returns a function that ignores its arguments and returns a particular value.

[17]:
```python
def constant(value):
  """Return `value`, ignoring any arguments"""
  return lambda *args: value
```

Similarly, several of the augmentations are functions that just return their first argument. Again, we can define a generic form `first` of such a function:

[18]:
```python
def first(*args):
  """Return the value of the first (and perhaps only) subconstituent,
     ignoring any others"""
  return args[0]
```

We can now rewrite the grammar above to take advantage of these shortcuts.

> In the call to `parse_augmented_grammar` below, we pass in the global environment, extracted via a `globals()` function call, via the named argument `globals`. This allows the `parse_augmented_grammar` function to make use of the global bindings for `constant`, `first`, and the like when evaluating the augmentation expressions to their values. You can check out the code in `transform.py` to see how the passed in `globals` bindings are used. To help understand what's going on, see what happens if you don't include the `globals=globals()`.

[19]:
```python
arithmetic_grammar_2, arithmetic_augmentations_2 = xform.
 ↪parse_augmented_grammar(
    """
    ## Sample grammar for arithmetic expressions

    S -> NUM                                : first
       | S OP S                             : lambda S1, Op, S2: Op(S1, S2)

    OP -> ADD                               : first
       | SUB
       | MULT
       | DIV

    NUM -> 'zero'                           : constant(0)
         | 'one'                            : constant(1)
         | 'two'                            : constant(2)
         | 'three'                          : constant(3)
         | 'four'                           : constant(4)
         | 'five'                           : constant(5)
```

```
        | 'six'                           : constant(6)
        | 'seven'                         : constant(7)
        | 'eight'                         : constant(8)
        | 'nine'                          : constant(9)
        | 'ten'                           : constant(10)

    ADD -> 'plus' | 'added' 'to'          : constant(lambda x, y: x + y)
    SUB -> 'minus'                        : constant(lambda x, y: x - y)
    MULT -> 'times' | 'multiplied' 'by'   : constant(lambda x, y: x * y)
    DIV -> 'divided' 'by'                 : constant(lambda x, y: x / y)
    """,
    globals=globals())
```

Finally, it might make our lives easier to write a template of augmentations whose instantiation depends on the right-hand side of the rule.

We use a reserved keyword `_RHS` to denote the right-hand side of the syntactic rule, which will be replaced by a **list** of the right-hand-side strings. For example, an augmentation `numeric_template(_RHS)` would be as if written as `numeric_template(['zero'])` when the rule is `NUM -> 'zero'`, and `numeric_template(['one'])` when the rule is `NUM -> 'one'`. The details of how this works can be found at `scripts/transform.py`.

This would allow us to use a single template function, for example,

[20]:
```
def numeric_template(rhs):
  """Ignore the subphrase meanings and lookup the first right-hand-side symbol
     as a number"""
  return constant({'zero':0, 'one':1, 'two':2, 'three':3, 'four':4, 'five':5,
         'six':6, 'seven':7, 'eight':8, 'nine':9, 'ten':10}[rhs[0]])
```

and then further simplify the grammar specification:

[21]:
```
arithmetic_grammar_3, arithmetic_augmentations_3 = xform.
 ↪parse_augmented_grammar(
    """
    ## Sample grammar for arithmetic expressions

    S -> NUM                              : first
       | S OP S                           : lambda S1, Op, S2: Op(S1, S2)

    OP -> ADD                             : first
       | SUB
       | MULT
       | DIV

    NUM -> 'zero'  | 'one'   | 'two'      : numeric_template(_RHS)
         | 'three' | 'four'  | 'five'
         | 'six'   | 'seven' | 'eight'
```

```
        | 'nine'  | 'ten'

    ADD -> 'plus' | 'added' 'to'            : constant(lambda x, y: x + y)
    SUB -> 'minus'                          : constant(lambda x, y: x - y)
    MULT -> 'times' | 'multiplied' 'by'     : constant(lambda x, y: x * y)
    DIV -> 'divided' 'by'                   : constant(lambda x, y: x / y)
    """,
    globals=globals())
```

[22]:
```
parse_and_interpret("six divided by three", arithmetic_grammar_3,␣
  ↪arithmetic_augmentations_3)
```

```
       S
  _____|______________
 S           OP       S
 |           |        |
NUM         DIV      NUM
 |      _____|___     |
six divided      by three

(S (S (NUM six)) (OP (DIV divided by)) (S (NUM three))) ==> 2.0
```

3.3 Example: *Green Eggs and Ham* revisited

This stuff is tricky, so it's useful to see more examples before jumping in the deep end. In this simple GEaH fragment grammar, we use a larger set of auxiliary functions to build the augmentations.

[23]:
```
def forward(F, A):
  """Forward application: Return the application of the first
     argument to the second"""
  return F(A)

def backward(A, F):
  """Backward application: Return the application of the second
     argument to the first"""
  return F(A)

def second(*args):
  """Return the value of the second subconstituent, ignoring any others"""
  return args[1]

def ignore(*args):
  """Return `None`, ignoring everything about the constituent. (Good as a
     placeholder until a better augmentation can be devised.)"""
  return None
```

Using these, we can build and test the grammar.

[24]:
```
geah_grammar_spec = """
  ## Productions
  S -> NP VP           : backward
  VP -> V NP           : forward

  ## Lexicon
  V -> 'likes'         : constant(lambda Object: lambda Subject:␣
↪f"like({Subject}, {Object})")
  NP -> 'Sam' | 'sam'  : constant(_RHS[0])
  NP -> 'ham'
  NP -> 'eggs'
"""
```

[25]:
```
geah_grammar, geah_augmentations = xform.
↪parse_augmented_grammar(geah_grammar_spec,                                  ␣
↪globals=globals())
```

[26]:
```
parse_and_interpret("Sam likes ham", geah_grammar, geah_augmentations)
```

```
      S
  ____|____
 |        VP
 |     ____|___
NP   V        NP
 |   |        |
Sam likes    ham

(S (NP Sam) (VP (V likes) (NP ham))) ==> like(Sam, ham)
```

4 Semantics of ATIS queries

Now you're in a good position to understand and add augmentations to a more comprehensive grammar, say, one that parses ATIS queries and generates SQL queries.

In preparation for that, we need to load the ATIS data, both NL and SQL queries.

4.1 Loading and preprocessing the corpus

To simplify things a bit, we'll only consider ATIS queries whose question type (remember that from project segment 1?) is `flight_id`. We download training, development, and test splits for this subset of the ATIS corpus, including corresponding SQL queries.

[27]:
```
# Acquire the datasets - training, development, and test splits of the
# ATIS queries and corresponding SQL queries
download_if_needed(source_url, "data/", "/ATIS/test_flightid.nl")
download_if_needed(source_url, "data/", "/ATIS/test_flightid.sql")
download_if_needed(source_url, "data/", "/ATIS/dev_flightid.nl")
```

```
download_if_needed(source_url, "data/", "/ATIS/dev_flightid.sql")
download_if_needed(source_url, "data/", "/ATIS/train_flightid.nl")
download_if_needed(source_url, "data/", "/ATIS/train_flightid.sql")
```

[28]:
```
# Process data
for split in ['train', 'dev', 'test']:
    src_in_file = f'data/{split}_flightid.nl'
    tgt_in_file = f'data/{split}_flightid.sql'
    out_file = f'data/{split}.csv'

    with open(src_in_file, 'r') as f_src_in, open(tgt_in_file, 'r') as f_tgt_in:
        with open(out_file, 'w') as f_out:
            src, tgt= [], []
            writer = csv.writer(f_out)
            writer.writerow(('src','tgt'))
            for src_line, tgt_line in zip(f_src_in, f_tgt_in):
                writer.writerow((src_line.strip(), tgt_line.strip()))
```

Let's take a look at what the data file looks like.

[29]:
```
shell("head -2 data/dev.csv")
```

```
src,tgt
what flights are available tomorrow from denver to philadelphia,"SELECT DISTINCT
flight_1.flight_id FROM flight flight_1 , airport_service airport_service_1 ,
city city_1 , airport_service airport_service_2 , city city_2 , days days_1 ,
date_day date_day_1 WHERE flight_1.from_airport = airport_service_1.airport_code
AND airport_service_1.city_code = city_1.city_code AND city_1.city_name =
'DENVER' AND ( flight_1.to_airport = airport_service_2.airport_code AND
airport_service_2.city_code = city_2.city_code AND city_2.city_name =
'PHILADELPHIA' AND flight_1.flight_days = days_1.days_code AND days_1.day_name =
date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 1
AND date_day_1.day_number = 20 )"
```

4.2 Corpus preprocessing

We'll use `tokenizers` and `datasets` to process the data. We'll use the NLTK tokenizer from project segment 3.

[30]:
```
## NLTK Tokenizer
tokenizer_pattern = '\d+|st\.|[\w-]+|\$[\d\.]+|\S+'
nltk_tokenizer = nltk.tokenize.RegexpTokenizer(tokenizer_pattern)
def tokenize_nltk(string):
  return nltk_tokenizer.tokenize(string.lower())

## Demonstrating the tokenizer
## Note especially the handling of `"11pm"` and hyphenated words.
```

```
print(tokenize_nltk("Are there any first-class flights from St. Louis at 11pm␣
↪for less than $3.50?"))
```

```
['are', 'there', 'any', 'first-class', 'flights', 'from', 'st.', 'louis', 'at',
'11', 'pm', 'for', 'less', 'than', '$3.50', '?']
```

[31]:

```
dataset = load_dataset('csv', data_files={'train':f'data/train.csv', \
                                          'val': f'data/dev.csv', \
                                          'test': f'data/test.csv'})
dataset
```

```
WARNING:datasets.builder:Using custom data configuration
default-8ca8d8a59fc02b6c
```

```
Downloading and preparing dataset csv/default to /root/.cache/huggingface/datase
ts/csv/default-8ca8d8a59fc02b6c/0.0.0/6b34fb8fcf56f7c8ba51dc895bfa2bfbe43546f190
a60fcf74bb5e8afdcc2317…
```

```
Downloading data files:   0%|          | 0/3 [00:00<?, ?it/s]
```

```
Extracting data files:   0%|          | 0/3 [00:00<?, ?it/s]
```

```
Generating train split: 0 examples [00:00, ? examples/s]
```

```
/usr/local/lib/python3.10/dist-
packages/datasets/download/streaming_download_manager.py:776: FutureWarning: the
'mangle_dupe_cols' keyword is deprecated and will be removed in a future
version. Please take steps to stop the use of 'mangle_dupe_cols'
  return pd.read_csv(xopen(filepath_or_buffer, "rb",
use_auth_token=use_auth_token), **kwargs)
```

```
Generating val split: 0 examples [00:00, ? examples/s]
```

```
/usr/local/lib/python3.10/dist-
packages/datasets/download/streaming_download_manager.py:776: FutureWarning: the
'mangle_dupe_cols' keyword is deprecated and will be removed in a future
version. Please take steps to stop the use of 'mangle_dupe_cols'
  return pd.read_csv(xopen(filepath_or_buffer, "rb",
use_auth_token=use_auth_token), **kwargs)
```

```
Generating test split: 0 examples [00:00, ? examples/s]
```

```
Dataset csv downloaded and prepared to /root/.cache/huggingface/datasets/csv/def
ault-8ca8d8a59fc02b6c/0.0.0/6b34fb8fcf56f7c8ba51dc895bfa2bfbe43546f190a60fcf74bb
5e8afdcc2317. Subsequent calls will reuse this data.
```

```
/usr/local/lib/python3.10/dist-
packages/datasets/download/streaming_download_manager.py:776: FutureWarning: the
'mangle_dupe_cols' keyword is deprecated and will be removed in a future
version. Please take steps to stop the use of 'mangle_dupe_cols'
  return pd.read_csv(xopen(filepath_or_buffer, "rb",
use_auth_token=use_auth_token), **kwargs)
```

```
0%|          | 0/3 [00:00<?, ?it/s]
```

[31]:
```
DatasetDict({
    train: Dataset({
        features: ['src', 'tgt'],
        num_rows: 3651
    })
    val: Dataset({
        features: ['src', 'tgt'],
        num_rows: 398
    })
    test: Dataset({
        features: ['src', 'tgt'],
        num_rows: 332
    })
})
```

[32]:
```
train_data = dataset['train']
val_data = dataset['val']
test_data = dataset['test']
```

[33]:
```
MIN_FREQ = 3
unk_token = '[UNK]'
pad_token = '[PAD]'
bos_token = '<bos>'
eos_token = '<eos>'

src_tokenizer = Tokenizer(WordLevel(unk_token=unk_token))
src_tokenizer.normalizer = normalizers.Lowercase()
src_tokenizer.pre_tokenizer = Split(Regex(tokenizer_pattern),
 ↪behavior='removed', invert=True)

src_trainer = WordLevelTrainer(min_frequency=MIN_FREQ,
 ↪special_tokens=[pad_token, unk_token])
src_tokenizer.train_from_iterator(train_data['src'], trainer=src_trainer)

tgt_tokenizer = Tokenizer(WordLevel(unk_token=unk_token))
tgt_tokenizer.pre_tokenizer = WhitespaceSplit()

tgt_trainer = WordLevelTrainer(min_frequency=MIN_FREQ,
 ↪special_tokens=[pad_token, unk_token, bos_token, eos_token])

tgt_tokenizer.train_from_iterator(train_data['tgt'], trainer=tgt_trainer)

tgt_tokenizer.post_processor = TemplateProcessing(single=f"{bos_token} $A
 ↪{eos_token}", special_tokens=[(bos_token, tgt_tokenizer.
 ↪token_to_id(bos_token)), (eos_token, tgt_tokenizer.token_to_id(eos_token))])
```

Note that we prepended `<bos>` and appended `<eos>` to target sentences.

We use `datasets.Dataset.map` to convert text into word ids. As shown in lab 1-5, first we need to wrap `tokenizer` with the `transformers.PreTrainedTokenizerFast` class to be compatible with the `datasets` library.

[34]:
```
hf_src_tokenizer = PreTrainedTokenizerFast(tokenizer_object=src_tokenizer,␣
↪pad_token=pad_token)
hf_tgt_tokenizer = PreTrainedTokenizerFast(tokenizer_object=tgt_tokenizer,␣
↪pad_token=pad_token)

def encode(example):
    example['src_ids'] = hf_src_tokenizer(example['src']).input_ids
    example['tgt_ids'] = hf_tgt_tokenizer(example['tgt']).input_ids
    return example

train_data = train_data.map(encode)
val_data = val_data.map(encode)
test_data = test_data.map(encode)
```

```
0%|          | 0/3651 [00:00<?, ?ex/s]

0%|          | 0/398 [00:00<?, ?ex/s]

0%|          | 0/332 [00:00<?, ?ex/s]
```

[35]:
```
# Compute size of vocabulary
src_vocab = src_tokenizer.get_vocab()
tgt_vocab = tgt_tokenizer.get_vocab()

print(f"Size of English vocab: {len(src_vocab)}")
print(f"Size of SQL vocab: {len(tgt_vocab)}")
print(f"Index for src padding: {src_vocab[pad_token]}")
print(f"Index for tgt padding: {tgt_vocab[pad_token]}")
print(f"Index for start of sequence token: {tgt_vocab[bos_token]}")
print(f"Index for end of sequence token: {tgt_vocab[eos_token]}")
```

```
Size of English vocab: 421
Size of SQL vocab: 392
Index for src padding: 0
Index for tgt padding: 0
Index for start of sequence token: 2
Index for end of sequence token: 3
```

Next, we batch our data to facilitate processing on a GPU. Batching is a bit tricky because the source and target will typically be of different lengths and have to padd the sequences to the same length. Since there is padding, we need to handle them with `pack` and `unpack` later on in the seq2seq part (as in lab 4-5).

```
[36]: BATCH_SIZE = 16     # batch size for training and validation
TEST_BATCH_SIZE = 1 # batch size for test; we use 1 to make implementation␣
 ↪easier

# Defines how to batch a list of examples together
def collate_fn(examples):
    batch = {}
    bsz = len(examples)
    src_ids, tgt_ids = [], []
    for example in examples:
        src_ids.append(example['src_ids'])
        tgt_ids.append(example['tgt_ids'])

    src_len = torch.LongTensor([len(word_ids) for word_ids in src_ids]).
 ↪to(device)
    src_max_length = max(src_len)
    tgt_max_length = max([len(word_ids) for word_ids in tgt_ids])

    src_batch = torch.zeros(bsz, src_max_length).long().
 ↪fill_(src_vocab[pad_token]).to(device)
    tgt_batch = torch.zeros(bsz, tgt_max_length).long().
 ↪fill_(tgt_vocab[pad_token]).to(device)
    for b in range(bsz):
        src_batch[b][:len(src_ids[b])] = torch.LongTensor(src_ids[b]).to(device)
        tgt_batch[b][:len(tgt_ids[b])] = torch.LongTensor(tgt_ids[b]).to(device)

    batch['src_lengths'] = src_len
    batch['src_ids'] = src_batch
    batch['tgt_ids'] = tgt_batch
    return batch

train_iter = torch.utils.data.DataLoader(train_data,
                                         batch_size=BATCH_SIZE,
                                         shuffle=True,
                                         collate_fn=collate_fn)
val_iter = torch.utils.data.DataLoader(val_data,
                                       batch_size=BATCH_SIZE,
                                       shuffle=False,
                                       collate_fn=collate_fn)
test_iter = torch.utils.data.DataLoader(test_data,
                                        batch_size=TEST_BATCH_SIZE,
                                        shuffle=False,
                                        collate_fn=collate_fn)
```

Let’s look at a single batch from one of these iterators.

```python
[37]: batch = next(iter(train_iter))
src_ids = batch['src_ids']
src_example = src_ids[2]
print (f"Size of text batch: {src_ids.size()}")
print (f"Third sentence in batch: {src_example}")
print (f"Length of the third sentence in batch: {len(src_example)}")
print (f"Converted back to string: {hf_src_tokenizer.decode(src_example)}")

tgt_ids = batch['tgt_ids']
tgt_example = tgt_ids[2]
print (f"Size of sql batch: {tgt_ids.size()}")
print (f"Third sql in batch: {tgt_example}")
print (f"Converted back to string: {hf_tgt_tokenizer.decode(tgt_example)}")
```

```
Size of text batch: torch.Size([16, 30])
Third sentence in batch: tensor([ 9,  7,  4,  3, 13, 16,  2, 11,  6, 69,  0,  0,
0,  0,  0,  0,  0,  0,
         0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0], device='cuda:0')
Length of the third sentence in batch: 30
Converted back to string: show me flights from san francisco to boston on
thursday [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
Size of sql batch: torch.Size([16, 153])
Third sql in batch: tensor([  2,  14,  31,  11,  13,  12,  16,   6,   7,  22,
6,   8,  23,   6,
          7,  29,   6,   8,  30,   6,  33,  40,   6,  38,  46,  15,  21,   4,
         18,   5,  19,   4,  17,   5,  20,   4,  54,  56,   5,   9,  24,   4,
         25,   5,  26,   4,  27,   5,  28,   4,  52,   5,  34,   4,  36,   5,
         37,   4,  41,   5,  44,   4,  35,   5,  43,   4, 103,   5,  42,   4,
        126,  10,   3,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,
          0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0,   0],
       device='cuda:0')
Converted back to string: <bos> SELECT DISTINCT flight_1.flight_id FROM flight
flight_1, airport_service airport_service_1, city city_1, airport_service
airport_service_2, city city_2, days days_1, date_day date_day_1 WHERE
flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'SAN
FRANCISCO' AND ( flight_1.to_airport = airport_service_2.airport_code AND
airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'BOSTON'
AND flight_1.flight_days = days_1.days_code AND days_1.day_name =
date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 5
AND date_day_1.day_number = 24 ) <eos> [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
```

```
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
[PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
```

Alternatively, we can directly iterate over the raw examples:

[38]:
```
for _, example in zip(range(1), train_data):
  train_text_1 = example['src'] # detokenized question
  train_sql_1 = example['tgt']  # detokenized sql
  print (f"Question: {train_text_1}\n")
  print (f"SQL: {train_sql_1}")
```

```
Question: list all the flights that arrive at general mitchell international
from various cities

SQL: SELECT DISTINCT flight_1.flight_id FROM flight flight_1 , airport airport_1
, airport_service airport_service_1 , city city_1 WHERE flight_1.to_airport =
airport_1.airport_code AND airport_1.airport_code = 'MKE' AND
flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND 1 = 1
```

4.3 Establishing a SQL database for evaluating ATIS queries

The output of our systems will be SQL queries. How should we determine if the generated queries are correct? We can't merely compare against the gold SQL queries, since there are many ways to implement a SQL query that answers any given NL query.

Instead, we will execute the queries – both the predicted SQL query and the gold SQL query – on an actual database, and verify that the returned responses are the same. For that purpose, we need a SQL database server to use. We'll set one up here, using the Python `sqlite3` module.

[39]:
```
@func_set_timeout(TIMEOUT)
def execute_sql(sql):
  conn = sqlite3.connect('data/atis_sqlite.db')  # establish the DB based on␣
 ↪the downloaded data
  c = conn.cursor()                              # build a "cursor"
  c.execute(sql)
  results = list(c.fetchall())
  c.close()
  conn.close()
  return results
```

To run a query, we use the cursor's `execute` function, and retrieve the results with `fetchall`. Let's get all the flights that arrive at General Mitchell International – the query `train_sql_1` above. There's a lot, so we'll just print out the first few.

[40]:
```
predicted_ret = execute_sql(train_sql_1)

print(f"""
Executing: {train_sql_1}

Result: {len(predicted_ret)} entries starting with

{predicted_ret[:10]}
""")
```

```
Executing: SELECT DISTINCT flight_1.flight_id FROM flight flight_1 , airport
airport_1 , airport_service airport_service_1 , city city_1 WHERE
flight_1.to_airport = airport_1.airport_code AND airport_1.airport_code = 'MKE'
AND flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND 1 = 1

Result: 534 entries starting with

[(107929,), (107930,), (107931,), (107932,), (107933,), (107934,), (107935,),
(107936,), (107937,), (107938,)]
```

For your reference, the SQL database we are using has a database schema described at https://github.com/jkkummerfeld/text2sql-data/blob/master/data/atis-schema.csv, and is consistent with the SQL queries provided in the various `.sql` files loaded above.

5 Rule-based parsing and interpretation of ATIS queries

First, you will implement a rule-based semantic parser using a grammar like the one you completed in the third project segment. We've placed an initial grammar in the file `data/grammar`. In addition to the helper functions defined above (`constant`, `first`, etc.), it makes use of some other simple functions. We've included those below, but you can (and almost certainly should) augment this set with others that you define as you build out the full set of augmentations.

[41]:
```
def upper(term):
  return '"' + term.upper() + '"'

def weekday(day):
  return f"flight.flight_days IN (SELECT days.days_code FROM days WHERE days.
↪day_name = '{day.upper()}')"

def month_name(month):
  return {'JANUARY' : 1,
          'FEBRUARY' : 2,
          'MARCH' : 3,
          'APRIL' : 4,
```

```
            'MAY' : 5,
            'JUNE' : 6,
            'JULY' : 7,
            'AUGUST' : 8,
            'SEPTEMBER' : 9,
            'OCTOBER' : 10,
            'NOVEMBER' : 11,
            'DECEMBER' : 12}[month.upper()]

def airports_from_airport_name(airport_name):
  return f"(SELECT airport.airport_code FROM airport WHERE airport.airport_name␣
 ↪= {upper(airport_name)})"

def airports_from_city(city):
  return f"""
    (SELECT airport_service.airport_code FROM airport_service WHERE␣
 ↪airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = {upper(city)}))
  """

def null_condition(*args, **kwargs):
  return 1

def depart_around(time):
  return f"""
    flight.departure_time >= {add_delta(miltime(time), -15).strftime('%H%M')}
    AND flight.departure_time <= {add_delta(miltime(time), 15).strftime('%H%M')}
    """.strip()

def add_delta(tme, delta):
    # transform to a full datetime first
    return (datetime.datetime.combine(datetime.date.today(), tme) +
            datetime.timedelta(minutes=delta)).time()

def miltime(minutes):
  return datetime.time(hour=int(minutes/100), minute=(minutes % 100))
```

[42]:
```
#auxiliary functions#
def flight_dest(dest):
  return f"flight.to_airport IN {dest}"

def flight_origin(origin):
  return f"flight.from_airport IN {origin}"

def a_and_b(a, b):
  return f"{a} AND {b}"
```

```python
def SQL_start(np):
  return f"SELECT DISTINCT flight.flight_id FROM flight WHERE {np}"

def airline_code(code):
  return f"flight.airline_code = '{code}'"

def arrive_before(t):
  return f"flight.arrival_time < {t}"

def arrive_after(t):
  return f"flight.arrival_time > {t}"

def depart_before(t):
  return f"flight.departure_time < {t}"

def depart_after(t):
   return f"flight.departure_time > {t}"

def depart_around(time):
  return f"""
    flight.arrival_time >= {add_delta(miltime(time), -15).strftime('%H%M')}
    AND flight.arrival_time <= {add_delta(miltime(time), 15).strftime('%H%M')}
    """.strip()

def depart_on(t):
  return f"flight.departure_time = {t}"

def arrive_on(t):
  return f"flight.arrival_time = {t}"

# def date_MDY(month = None, day = None, year = None):
#   date_constraint = []
#   if day is not None:
#     date_constraint.append(f"DAY_NUMBER = {day}")
#   if month is not None:
#     date_constraint.append(f"MONTH_NUMBER = {month}")
#   if year is not None:
#     date_constraint.append(f"YEAR = {year}")

#   return f"flight.flight_days IN (SELECT days.days_code FROM  WHERE ()"
```

We can build a parser with the augmented grammar:

[43]:
```python
atis_grammar, atis_augmentations = xform.read_augmented_grammar('data/grammar',␣
↪globals=globals())
atis_parser = nltk.parse.BottomUpChartParser(atis_grammar)
```

We'll define a function to return a parse tree for a string according to the ATIS grammar (if available).

[44]:
```
def parse_tree(sentence):
  """Parse a sentence and return the parse tree, or None if failure."""
  try:
    parses = list(atis_parser.parse(tokenize_nltk(sentence)))
    if len(parses) == 0:
      return None
    else:
      return parses[0]
  except:
    return None
```

We can check the overall coverage of this grammar on the training set by using the `parse_tree` function to determine if a parse is available. The grammar that we provide should get about a 44% coverage of the training set.

[45]:
```
# Check coverage on training set
parsed = 0
with open("data/train_flightid.nl") as train:
  examples = train.readlines()[:]
for sentence in tqdm(examples):
  if parse_tree(sentence):
    parsed += 1
  else:
    next

print(f"\nParsed {parsed} of {len(examples)} ({parsed*100/(len(examples)):.
↪2f}%)")
```

```
100%|      | 3651/3651 [00:15<00:00, 228.20it/s]


Parsed 1609 of 3651 (44.07%)
```

5.1 Goal 1: Construct SQL queries from a parse tree and evaluate the results

It's time to turn to the first major part of this project segment, implementing a rule-based semantic parsing system to answer flight-ID-type ATIS queries.

Recall that in rule-based semantic parsing, each syntactic rule is associated with a semantic composition rule. The grammar we've provided has semantic augmentations for some of the low-level phrases – cities, airports, times, airlines – but not the higher level syntactic types. You'll be adding those.

In the ATIS grammar that we provide, as with the earlier toy grammars, the augmentation for a rule with n nonterminals and m terminals on the right-hand side is assumed to be called with n positional arguments (the values for the corresponding children). The `interpret` function you've already defined should therefore work well with this grammar.

Let's run through one way that a semantic derivation might proceed, for the sample query "flights to boston":

[46]:
```
sample_query = "flights to boston"
print(tokenize_nltk(sample_query))
sample_tree = parse_tree(sample_query)
sample_tree.pretty_print()
```

```
['flights', 'to', 'boston']
                S
                |
            NP_FLIGHT
                |
            NOM_FLIGHT
                |
             N_FLIGHT
      __________|_________
     |                    PP
     |                    |
     |                 PP_PLACE
     |           _________|_________
  N_FLIGHT      |                N_PLACE
     |          |                   |
TERM_FLIGHT  P_PLACE            TERM_PLACE
     |          |                   |
  flights       to                boston
```

Given a sentence, we first construct its parse tree using the syntactic rules, then compose the corresponding semantic rules bottom-up, until eventually we arrive at the root node with a finished SQL statement. For this query, we will go through what the possible meaning representations for the subconstituents of "flights to boston" might be. But this is just one way of doing things; other ways are possible, and you should feel free to experiment.

Working from bottom up:

1. The `TERM_PLACE` phrase "boston" uses the composition function template `constant(airports_from_city(' '.join(_RHS)))`, which will be instantiated as `constant(airports_from_city(' '.join(['boston'])))` (recall that `_RHS` is replaced by the right-hand side of the rule). The meaning of `TERM_PLACE` will be the SQL snippet

```
SELECT airport_service.airport_code
FROM airport_service
WHERE airport_service.city_code IN
    (SELECT city.city_code
     FROM city
     WHERE city.city_name = "BOSTON")
```

(This query generates a list of all of the airports in Boston.)

2. The `N_PLACE` phrase "boston" can have the same meaning as the `TERM_PLACE`.

3. The `P_PLACE` phrase "to" might be associated with a function that maps a SQL query for a list of airports to a SQL condition that holds of flights that go to one of those airports, i.e., `flight.to_airport IN (...)`.
4. The `PP_PLACE` phrase "to boston" might apply the `P_PLACE` meaning to the `TERM_PLACE` meaning, thus generating a SQL condition that holds of flights that go to one of the Boston airports:

```
flight.to_airport IN
    (SELECT airport_service.airport_code
     FROM airport_service
     WHERE airport_service.city_code IN
         (SELECT city.city_code
          FROM city
          WHERE city.city_name = "BOSTON"))
```

5. The `PP` phrase "to Boston" can again get its meaning from the `PP_PLACE`.
6. The `TERM_FLIGHT` phrase "flights" might also return a condition on flights, this time the "null condition", represented by the SQL truth value `1`. Ditto for the `N_FLIGHT` phrase "flights".
7. The `N_FLIGHT` phrase "flights to boston" can conjoin the two conditions, yielding the SQL condition

```
flight.to_airport IN
    (SELECT airport_service.airport_code
     FROM airport_service
     WHERE airport_service.city_code IN
         (SELECT city.city_code
          FROM city
          WHERE city.city_name = "BOSTON"))
AND 1
```

which can be inherited by the `NOM_FLIGHT` and `NP_FLIGHT` phrases.

8. The `S` phrase "flights to boston" can use the condition provided by the `NP_FLIGHT` phrase to select all flights satisfying the condition with a SQL query like

```
SELECT DISTINCT flight.flight_id
FROM flight
WHERE flight.to_airport IN
        (SELECT airport_service.airport_code
         FROM airport_service
         WHERE airport_service.city_code IN
             (SELECT city.city_code
              FROM city
              WHERE city.city_name = "BOSTON"))
      AND 1
```

This SQL query is then taken to be a representation of the meaning for the NL query "flights to boston", and can be executed against the ATIS database to retrieve the requested flights.

Now, it's your turn to add augmentations to `data/grammar` to make this example work. The augmentations that we have provided for the grammar make use of a set of auxiliary functions that we defined above. You should feel free to add your own auxiliary functions that you make use of in the grammar.

[47]:
```
#TODO: add augmentations to `data/grammar` to make this example work
atis_grammar, atis_augmentations = xform.read_augmented_grammar('data/grammar',␣
↪globals=globals())
atis_parser = nltk.parse.BottomUpChartParser(atis_grammar)
predicted_sql = interpret(sample_tree, atis_augmentations)
print("Predicted SQL:\n\n", predicted_sql, "\n")
```

```
Predicted SQL:

 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "BOSTON"))
   AND 1
```

Verification on some examples With a rule-based semantic parsing system, we can generate SQL queries given questions, and then execute those queries on a SQL database to answer the given questions. To evaluate the performance of the system, we compare the returned results against the results of executing the ground truth queries.

We provide a function `verify` to compare the results from our generated SQL to the ground truth SQL. It should be useful for testing individual queries.

[48]:
```
def verify(predicted_sql, gold_sql, silent=True):
  """
  Compare the correctness of the generated SQL by executing on the
  ATIS database and comparing the returned results.
  Arguments:
      predicted_sql: the predicted SQL query
      gold_sql: the reference SQL query to compare against
      silent: print outputs or not
  Returns: True if the returned results are the same, otherwise False
  """
  # Execute predicted SQL
  try:
    predicted_result = execute_sql(predicted_sql)
  except BaseException as e:
    if not silent:
      print(f"predicted sql exec failed: {e}")
    return False
  if not silent:
    print("Predicted DB result:\n\n", predicted_result[:10], "\n")
```

```
  # Execute gold SQL
  try:
    gold_result = execute_sql(gold_sql)
  except BaseException as e:
    if not silent:
      print(f"gold sql exec failed: {e}")
    return False
  if not silent:
    print("Gold DB result:\n\n", gold_result[:10], "\n")

  # Verify correctness
  if gold_result == predicted_result:
    return True
```

Let's try this methodology on a simple example: "flights from phoenix to milwaukee". we provide it along with the gold SQL query.

[49]:
```
def rule_based_trial(sentence, gold_sql):
  print("Sentence: ", sentence, "\n")
  tree = parse_tree(sentence)
  print("Parse:\n\n")
  tree.pretty_print()

  predicted_sql = interpret(tree, atis_augmentations)
  print("Predicted SQL:\n\n", predicted_sql, "\n")

  if verify(predicted_sql, gold_sql, silent=False):
    print ('Correct!')
  else:
    print ('Incorrect!')
```

[50]:
```
# Run this cell to reload augmentations after you make changes to `data/grammar`
atis_grammar, atis_augmentations = xform.read_augmented_grammar('data/grammar',␣
↪globals=globals())
atis_parser = nltk.parse.BottomUpChartParser(atis_grammar)
```

[51]:
```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 1
example_1 = 'flights from phoenix to milwaukee'
gold_sql_1 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2
  WHERE flight_1.from_airport = airport_service_1.airport_code
```

```
        AND airport_service_1.city_code = city_1.city_code
        AND city_1.city_name = 'PHOENIX'
        AND flight_1.to_airport = airport_service_2.airport_code
        AND airport_service_2.city_code = city_2.city_code
        AND city_2.city_name = 'MILWAUKEE'
  """

rule_based_trial(example_1, gold_sql_1)
```

```
Sentence:  flights from phoenix to milwaukee

Parse:


                                  S
                                  |
                              NP_FLIGHT
                                  |
                              NOM_FLIGHT
                                  |
                               N_FLIGHT
                 _________________|_________________
            N_FLIGHT                                  |
      _________|________                              |
     |                  PP                            PP
     |                  |                             |
     |              PP_PLACE                      PP_PLACE
     |           ________|_________           _______|_________
  N_FLIGHT      |                N_PLACE     |                 N_PLACE
     |          |                  |         |                    |
TERM_FLIGHT P_PLACE          TERM_PLACE P_PLACE           TERM_PLACE
     |          |                  |         |                    |
  flights     from              phoenix     to                milwaukee

Predicted SQL:

 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "MILWAUKEE"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "PHOENIX"))
   AND 1

Predicted DB result:
```

```
 [(108086,), (108087,), (301763,), (301764,), (301765,), (301766,), (302323,),
(304881,), (310619,), (310620,)]

Gold DB result:

 [(108086,), (108087,), (301763,), (301764,), (301765,), (301766,), (302323,),
(304881,), (310619,), (310620,)]

Correct!
```

To make development faster, we recommend starting with a few examples before running the full evaluation script. We've taken some examples from the ATIS dataset including the gold SQL queries that they provided. Of course, yours (and those of the project segment solution set) may differ.

[52]:
```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 2
example_2 = 'i would like a united flight'
gold_sql_2 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1
  WHERE flight_1.airline_code = 'UA'
  """

rule_based_trial(example_2, gold_sql_2)
```

```
Sentence:  i would like a united flight

Parse:


                                                S
____________|___________________________________________________
                                    |
NP_FLIGHT
                                    |
|
                                PREIGNORE
NOM_FLIGHT
        ____________________________|____________
___________|___________
       |                                     PREIGNORE
ADJ                        |
       |                 _________________________|____________
|                        |
       |               |                                   PREIGNORE
ADJ_AIRLINE              NOM_FLIGHT
```

```
       |               |                          ____________|____________
|                      |
       |               |                         |                    PREIGNORE
TERM_AIRLINE              N_FLIGHT
       |               |                         |                         |
|                      |
PREIGNORESYMBOL PREIGNORESYMBOL           PREIGNORESYMBOL
PREIGNORESYMBOL TERM_AIRBRAND            TERM_FLIGHT
       |               |                         |                         |
|                      |
       i             would                      like                       a
united                  flight
```

```
Predicted SQL:

 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.airline_code = 'UA'
AND 1

Predicted DB result:

 [(100094,), (100099,), (100145,), (100158,), (100164,), (100167,), (100169,),
(100203,), (100204,), (100296,)]

Gold DB result:

 [(100094,), (100099,), (100145,), (100158,), (100164,), (100167,), (100169,),
(100203,), (100204,), (100296,)]

Correct!
```

[53]:

```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 3
example_3 = 'i would like a flight between boston and dallas'
gold_sql_3 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2
  WHERE flight_1.from_airport = airport_service_1.airport_code
        AND airport_service_1.city_code = city_1.city_code
        AND city_1.city_name = 'BOSTON'
        AND flight_1.to_airport = airport_service_2.airport_code
        AND airport_service_2.city_code = city_2.city_code
        AND city_2.city_name = 'DALLAS'
  """
```

```
# Note that the parse tree might appear wrong: instead of
# `PP_PLACE -> 'between' N_PLACE 'and' N_PLACE`, the tree appears to be
# `PP_PLACE -> 'between' 'and' N_PLACE N_PLACE`. But it's only a visualization
# error of tree.pretty_print() and you should assume that the production is
# `PP_PLACE -> 'between' N_PLACE 'and' N_PLACE` (you can verify by printing out
# all productions).
rule_based_trial(example_3, gold_sql_3)
```

```
Sentence:  i would like a flight between boston and dallas

Parse:


         S
____________________________________________________|__________
                                    |
NP_FLIGHT
                                    |
|
                                    |
NOM_FLIGHT
                                    |
|
                               PREIGNORE
N_FLIGHT
        ____________________________|____________
__________|_________
       |                                PREIGNORE
|                    PP
       |                ___________________________|____________
|                    |
       |                |                             PREIGNORE
|                 PP_PLACE
       |                |                         ____________|____________
|           _________|_____________________
       |                |                         |                     PREIGNORE
N_FLIGHT       |          |       N_PLACE    N_PLACE
       |                |                         |                         |
|          |          |          |          |
PREIGNORESYMBOL PREIGNORESYMBOL          PREIGNORESYMBOL
PREIGNORESYMBOL TERM_FLIGHT     |          |     TERM_PLACE TERM_PLACE
       |                |                         |                         |
|          |          |          |          |
       i              would                       like                       a
flight    between      and      boston     dallas
```

```
Predicted SQL:

 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "BOSTON"))
   AND flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "DALLAS"))
   AND 1

Predicted DB result:

 [(103171,), (103172,), (103173,), (103174,), (103175,), (103176,), (103177,),
(103178,), (103179,), (103180,)]

Gold DB result:

 [(103171,), (103172,), (103173,), (103174,), (103175,), (103176,), (103177,),
(103178,), (103179,), (103180,)]

Correct!
```

[54]:

```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 4
example_4 = 'show me the united flights from denver to baltimore'
gold_sql_4 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2
  WHERE flight_1.airline_code = 'UA'
        AND ( flight_1.from_airport = airport_service_1.airport_code
              AND airport_service_1.city_code = city_1.city_code
              AND city_1.city_name = 'DENVER'
              AND flight_1.to_airport = airport_service_2.airport_code
              AND airport_service_2.city_code = city_2.city_code
              AND city_2.city_name = 'BALTIMORE' )

  """

rule_based_trial(example_4, gold_sql_4)
```

```
Sentence:  show me the united flights from denver to baltimore
```

Parse:

```
                                                                              S
_____________________________________________________|__________
                       |
NP_FLIGHT
                       |
|
                       |
NOM_FLIGHT
                       |
_______________________|___________________
                       |                                        |
NOM_FLIGHT
                       |                                        |
|
                       |                                        |
N_FLIGHT
                       |                                        |
___________________|_________________
                       |                                        |
N_FLIGHT                                  |
                       |                                        |
__________|_________                            |
                   PREIGNORE                                   ADJ            |
PP                          PP
        _______________|____________                          |             |
|                           |
       |                    PREIGNORE                    ADJ_AIRLINE        |
PP_PLACE                    PP_PLACE
       |               ____________|____________               |             |
_________|_________          _______|_________
       |               |                     PREIGNORE     TERM_AIRLINE
N_FLIGHT       |               N_PLACE      |               N_PLACE
       |               |                       |              |             |
|                  |         |                 |
PREIGNORESYMBOL PREIGNORESYMBOL           PREIGNORESYMBOL TERM_AIRBRAND
TERM_FLIGHT  P_PLACE            TERM_PLACE P_PLACE          TERM_PLACE
       |               |                       |              |             |
|                  |         |                 |
      show             me                     the           united
flights      from               denver      to            baltimore
```

Predicted SQL:

```
 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.airline_code = 'UA'
AND flight.to_airport IN
```

```
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "BALTIMORE"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "DENVER"))
   AND 1

Predicted DB result:

 [(101231,), (101233,), (305983,)]

Gold DB result:

 [(101231,), (101233,), (305983,)]

Correct!
```

[55]:
```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 5
example_5 = 'show flights from cleveland to miami that arrive before 4pm'
gold_sql_5 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2
  WHERE flight_1.from_airport = airport_service_1.airport_code
        AND airport_service_1.city_code = city_1.city_code
        AND city_1.city_name = 'CLEVELAND'
        AND ( flight_1.to_airport = airport_service_2.airport_code
              AND airport_service_2.city_code = city_2.city_code
              AND city_2.city_name = 'MIAMI'
              AND flight_1.arrival_time < 1600 )
  """

rule_based_trial(example_5, gold_sql_5)
```

```
Sentence:  show flights from cleveland to miami that arrive before 4pm

Parse:


                               S
        _______________________|______________________________________________
       |
```

```
NP_FLIGHT
       |                                                                       |
       |
NOM_FLIGHT
       |                                                                       |
       |
N_FLIGHT
       |
___________________________|_____________________
       |                                       N_FLIGHT
|
       |                        __________________|________________
|
       |                    N_FLIGHT                              |
PP
       |              _________|________                          |
|
       |             |                 PP                          PP
PP_TIME
       |             |                 |                           |
_______|__________________
       |             |              PP_PLACE                    PP_PLACE
|                        NP_TIME
       |             |          ________|_________           _______|_________
|                            |
   PREIGNORE      N_FLIGHT     |              N_PLACE      |
N_PLACE           |                     TERM_TIME
       |             |         |                  |         |                 |
|                 _________|__________
PREIGNORESYMBOL TERM_FLIGHT P_PLACE          TERM_PLACE P_PLACE
TERM_PLACE      P_TIME         TERM_TIME          TERM_TIMEMOD
       |             |         |                  |         |                 |
_____|_______          |                    |
      show        flights     from           cleveland     to
miami    that arrive  before     4                  pm
```

Predicted SQL:

```
 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.arrival_time < 1600
AND flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "MIAMI"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "CLEVELAND"))
   AND 1
```

```
Predicted DB result:

 [(107698,), (301117,)]

Gold DB result:

 [(107698,), (301117,)]

Correct!
```

[56]:

```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 6
example_6 = 'okay how about a flight on sunday from tampa to charlotte'
gold_sql_6 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2 ,
       days days_1 ,
       date_day date_day_1
  WHERE flight_1.from_airport = airport_service_1.airport_code
        AND airport_service_1.city_code = city_1.city_code
        AND city_1.city_name = 'TAMPA'
        AND ( flight_1.to_airport = airport_service_2.airport_code
              AND airport_service_2.city_code = city_2.city_code
              AND city_2.city_name = 'CHARLOTTE'
              AND flight_1.flight_days = days_1.days_code
              AND days_1.day_name = date_day_1.day_name
              AND date_day_1.year = 1991
              AND date_day_1.month_number = 8
              AND date_day_1.day_number = 27 )
  """

# You might notice that the gold answer above used the exact date, which is
# not easily implementable. A more implementable way (generated by the project
# segment 4 solution code) is:
gold_sql_6b = """
  SELECT DISTINCT flight.flight_id
  FROM flight
  WHERE ((((1
            AND flight.flight_days IN (SELECT days.days_code
                                       FROM days
                                       WHERE days.day_name = 'SUNDAY')
            )
```

```
           AND flight.from_airport IN (SELECT airport_service.airport_code
                                      FROM airport_service
                                      WHERE airport_service.city_code IN␣
↪(SELECT city.city_code
                                                                             FROM␣
↪city
                                                                            ␣
↪WHERE city.city_name = "TAMPA")))
          AND flight.to_airport IN (SELECT airport_service.airport_code
                                   FROM airport_service
                                   WHERE airport_service.city_code IN (SELECT␣
↪city.city_code
                                                                          FROM␣
↪city
                                                                          WHERE␣
↪city.city_name = "CHARLOTTE"))))
"""

rule_based_trial(example_6, gold_sql_6b)
```

```
Sentence:  okay how about a flight on sunday from tampa to charlotte

Parse:


                   S
                                     ...........................................
___________________|____________________________
                                    |
NP_FLIGHT
                                    |
|
                                    |
NOM_FLIGHT
                                    |
|
                                    |
N_FLIGHT
                                    |
___________|_________________________________
                                    |
N_FLIGHT                                             |
                                    |
__________________|____________________                          |
                                PREIGNORE
N_FLIGHT                                    |                           |
        ____________________________|___________
```

```
_________|________                                        |                            |
       |                                     PREIGNORE
|                  PP                              PP
PP
       |                _________________________|___________
|                  |                              |                            |
       |               |                                   PREIGNORE
|               PP_DATE                        PP_PLACE
PP_PLACE
       |               |                       ____________|___________
|          ________|_________             _________|_________
_______|_________
       |               |                          |                    PREIGNORE
N_FLIGHT     |              NP_DATE        |              N_PLACE      |
N_PLACE
       |               |                          |                         |
|         |                 |           |                  |         |
|
PREIGNORESYMBOL PREIGNORESYMBOL           PREIGNORESYMBOL
PREIGNORESYMBOL TERM_FLIGHT  P_DATE          TERM_WEEKDAY  P_PLACE
TERM_PLACE P_PLACE          TERM_PLACE
       |               |                          |                         |
|         |                 |           |                  |         |
|
      okay            how                      about                      a
flight      on              sunday       from              tampa       to
charlotte
```

Predicted SQL:

```
 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "CHARLOTTE"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "TAMPA"))
   AND flight.flight_days IN (SELECT days.days_code FROM days WHERE
days.day_name = 'SUNDAY') AND 1
```

Predicted DB result:

```
 [(101860,), (101861,), (101862,), (101863,), (101864,), (101865,), (305231,)]
```

Gold DB result:

```
 [(101860,), (101861,), (101862,), (101863,), (101864,), (101865,), (305231,)]
```

```
Correct!
```

[57]:
```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 7
example_7 = 'list all flights going from boston to atlanta that leaves before 7␣
↪am on thursday'
gold_sql_7 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2 ,
       days days_1 ,
       date_day date_day_1
  WHERE flight_1.from_airport = airport_service_1.airport_code
        AND airport_service_1.city_code = city_1.city_code
        AND city_1.city_name = 'BOSTON'
        AND ( flight_1.to_airport = airport_service_2.airport_code
              AND airport_service_2.city_code = city_2.city_code
              AND city_2.city_name = 'ATLANTA'
              AND ( flight_1.flight_days = days_1.days_code
                    AND days_1.day_name = date_day_1.day_name
                    AND date_day_1.year = 1991
                    AND date_day_1.month_number = 5
                    AND date_day_1.day_number = 24
                    AND flight_1.departure_time < 700 ) )
  """

# Again, the gold answer above used the exact date, as opposed to the
# following approach:
gold_sql_7b = """
  SELECT DISTINCT flight.flight_id
  FROM flight
  WHERE ((1
          AND ((((1
                  AND flight.from_airport IN (SELECT airport_service.
↪airport_code
                                              FROM airport_service
                                              WHERE airport_service.city_code␣
↪IN (SELECT city.city_code
                                                                              ␣
↪  FROM city
                                                                              ␣
↪  WHERE city.city_name = "BOSTON")))
                 AND flight.to_airport IN (SELECT airport_service.airport_code
```

```
                                          FROM airport_service
                                          WHERE airport_service.city_code IN␣
↪(SELECT city.city_code
                                                                                                ␣
↪FROM city
                                                                                                ␣
↪WHERE city.city_name = "ATLANTA")))
                 AND flight.departure_time <= 0700)
                AND flight.flight_days IN (SELECT days.days_code
                                          FROM days
                                          WHERE days.day_name = 'THURSDAY'))))
 """

rule_based_trial(example_7, gold_sql_7b)
```

```
Sentence:  list all flights going from boston to atlanta that leaves before 7 am
on thursday

Parse:


                                  S
__________________________|_____________________________________________________
       |
NP_FLIGHT
       |          _____________________________________________________________
________|_________
       |         |
NOM_FLIGHT
       |         |
|
       |         |
N_FLIGHT
       |         |
__________|______________________________________________________________
       |         |
N_FLIGHT                                                                       |
       |         |
_____________________________________|___________________________
|
       |         |                             N_FLIGHT
|                                                |
       |         |                          ________|___________________________
|                                                |
       |         |                    N_FLIGHT                                |
PP                                               |
       |         |          _______________|________                          |
```

```
|                                               |
       |         |       |                      PP                           PP
PP_TIME                                          PP
       |         |       |                       |                           |
-------|-----------------                                 |
       |         |       |                   PP_PLACE
PP_PLACE                         |                      NP_TIME
PP_DATE
       |         |       |                _______|________
_______|________                     |                          |
_______|________
   PREIGNORE     |    N_FLIGHT           |              N_PLACE      |
N_PLACE                |                     TERM_TIME                |
NP_DATE
       |         |       |               |                 |         |
|                   |                ________|_________          |
|
PREIGNORESYMBOL DET TERM_FLIGHT       P_PLACE          TERM_PLACE P_PLACE
TERM_PLACE            P_TIME        TERM_TIME          TERM_TIMEMOD P_DATE
TERM_WEEKDAY
       |         |       |          ______|________          |         |
|           ________|_______          |                    |         |
|
      list      all   flights   going            from     boston      to
atlanta      that    leaves  before     7                   am        on
thursday
```

Predicted SQL:

```
 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.flight_days IN
(SELECT days.days_code FROM days WHERE days.day_name = 'THURSDAY') AND
flight.departure_time < 700 AND flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "ATLANTA"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "BOSTON"))
   AND 1
```

Predicted DB result:

```
 [(100014,)]
```

Gold DB result:

```
 [(100014,)]
```

Correct!

[58]:
```
#TODO: add augmentations to `data/grammar` to make this example work
# Example 8
example_8 = 'list the flights from dallas to san francisco on american airlines'
gold_sql_8 = """
  SELECT DISTINCT flight_1.flight_id
  FROM flight flight_1 ,
       airport_service airport_service_1 ,
       city city_1 ,
       airport_service airport_service_2 ,
       city city_2
  WHERE flight_1.airline_code = 'AA'
        AND ( flight_1.from_airport = airport_service_1.airport_code
              AND airport_service_1.city_code = city_1.city_code
              AND city_1.city_name = 'DALLAS'
              AND flight_1.to_airport = airport_service_2.airport_code
              AND airport_service_2.city_code = city_2.city_code
              AND city_2.city_name = 'SAN FRANCISCO' )
  """

rule_based_trial(example_8, gold_sql_8)
```

```
Sentence:  list the flights from dallas to san francisco on american airlines

Parse:


      S
_________________________________________________________________|________
                    |
NP_FLIGHT
                    |
|
                    |
NOM_FLIGHT
                    |
|
                    |
N_FLIGHT
                    |
__________________|__________________________________________
                    |
N_FLIGHT                                                                 |
                    |
__________________|__________________
|
```

```
                    |                                N_FLIGHT
|                                           |
                    |                           _________|________
|                                           |
                    |                          |                 PP
PP                                         PP
                    |                          |                 |
|                                           |
               PREIGNORE                       |              PP_PLACE
PP_PLACE                                  PP_AIRLINE
        ____________|____________              |         ________|_________
________|__________                      ___________|____________
       |                     PREIGNORE      N_FLIGHT     |               N_PLACE
|               N_PLACE                 |                  TERM_AIRLINE
       |                         |             |         |                  |
|                   |                   |
____________|______________
PREIGNORESYMBOL           PREIGNORESYMBOL TERM_FLIGHT P_PLACE
TERM_PLACE P_PLACE            TERM_PLACE           P_AIRLINE TERM_AIRBRAND
TERM_AIRBRANDTYP
       |                         |             |         |                  |
|                   |                   |           |
E
       |                         |             |         |                  |
|         __________|__________         |           |
|
      list                     the         flights     from              dallas
to      san                francisco     on       american
airlines
```

Predicted SQL:

```
 SELECT DISTINCT flight.flight_id FROM flight WHERE flight.airline_code = 'AA'
AND flight.to_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "SAN FRANCISCO"))
   AND flight.from_airport IN
    (SELECT airport_service.airport_code FROM airport_service WHERE
airport_service.city_code IN
      (SELECT city.city_code FROM city WHERE city.city_name = "DALLAS"))
   AND 1
```

Predicted DB result:

```
 [(108452,), (108454,), (108456,), (111083,), (111085,), (111086,), (111090,),
(111091,), (111092,), (111094,)]
```

```
Gold DB result:

 [(108452,), (108454,), (108456,), (111083,), (111085,), (111086,), (111090,),
(111091,), (111092,), (111094,)]

Correct!
```

5.1.1 Systematic evaluation on a test set

We can perform a more systematic evaluation by checking the accuracy of the queries on an entire test set for which we have gold queries. The `evaluate` function below does just this, calculating precision, recall, and F1 metrics for the test set. It takes as argument a "predictor" function, which maps token sequences to predicted SQL queries. We've provided a predictor function for the rule-based model in the next cell (and a predictor for the seq2seq system below when we get to that system).

The rule-based system does not generate predictions for all queries; many queries won't parse. The precision and recall metrics take this into account in measuring the efficacy of the method. The recall metric captures what proportion of *all of the test examples* for which the system generates a correct query. The precision metric captures what proportion of *all of the test examples for which a prediction is generated* for which the system generates a correct query. (Recall that F1 is just the geometric mean of precision and recall.)

Once you've made some progress on adding augmentations to the grammar, you can evaluate your progress by seeing if the precision and recall have improved. For reference, the solution code achieves precision of about 66% and recall of about 28% for an F1 of 39%.

[59]:
```python
def evaluate(predictor, dataset, num_examples=0, silent=True):
  """Evaluate accuracy of `predictor` by executing predictions on a
  SQL database and comparing returned results against those of gold queries.

  Arguments:
      predictor:    a function that maps a token sequence
                    to a predicted SQL query string
      dataset:      the dataset of token sequences and gold SQL queries
      num_examples: number of examples from `dataset` to use; all of
                    them if 0
      silent: if set to False, will print out logs
  Returns: precision, recall, and F1 score
  """
  # Prepare to count results
  if num_examples <= 0:
    num_examples = len(dataset)
  example_count = 0
  predicted_count = 0
  correct = 0
  incorrect = 0

  # Process the examples from the dataset
```

```
  for _, example in tqdm(zip(range(num_examples), dataset)):
    example_count += 1
    # obtain query SQL
    predicted_sql = predictor(example['src'])
    if predicted_sql == None:
      continue
    predicted_count += 1
    # obtain gold SQL
    gold_sql = example['tgt']

    # check that they're compatible
    if verify(predicted_sql, gold_sql):
      correct += 1
    else:
      incorrect += 1

  # Compute and return precision, recall, F1
  precision = correct / predicted_count if predicted_count > 0 else 0
  recall = correct / example_count
  f1 = (2 * precision * recall) / (precision + recall) if precision + recall >␣
↪0 else 0
  return precision, recall, f1
```

[60]:
```
def rule_based_predictor(query):
  tree = parse_tree(query)
  if tree is None:
    return None
  try:
    predicted_sql = interpret(tree, atis_augmentations)
  except Exception as err:
    return None
  return predicted_sql
```

[61]:
```
precision, recall, f1 = evaluate(rule_based_predictor, test_data,␣
↪num_examples=0)
print(f"precision: {precision:3.2f}")
print(f"recall:    {recall:3.2f}")
print(f"F1:        {f1:3.2f}")
```

```
332it [00:02, 124.58it/s]
```

```
precision: 0.68
recall:    0.28
F1:        0.39
```

6 End-to-End Seq2Seq Model

In this part, you will implement a seq2seq model **with attention mechanism** to directly learn the translation from NL query to SQL. You might find labs 4-4 and 4-5 particularly helpful, as the primary difference here is that we are using a different dataset.

Note: We recommend using GPUs to train the model in this part (one way to get GPUs is to use Google Colab and clicking Menu -> Runtime -> Change runtime type -> GPU), as we need to use a very large model to solve the task well. For development we recommend starting with a smaller model and training for only 1 epoch.

6.1 Goal 2: Implement a seq2seq model (with attention)

In lab 4-5, you implemented a neural encoder-decoder model with attention. That model was used to convert English number phrases to numbers, but one of the biggest advantages of neural models is that we can easily apply them to different tasks (such as machine translation and document summarization) by using different training datasets.

Implement the class `AttnEncoderDecoder` to convert natural language queries into SQL statements. You may find that you can reuse most of the code you wrote for lab 4-5. A reasonable way to proceed is to implement the following methods:

- **Model**
  1. `__init__`: an initializer where you create network modules.
  2. `forward`: given source word ids of size `(batch_size, max_src_len)`, source lengths of size `(batch_size)` and decoder input target word ids `(batch_size, max_tgt_len)`, returns logits `(batch_size, max_tgt_len, V_tgt)`. For better modularity you might want to implement it by implementing two functions `forward_encoder` and `forward_decoder`.
- **Optimization**
  3. `train_all`: compute loss on training data, compute gradients, and update model parameters to minimize the loss.
  4. `evaluate_ppl`: evaluate the current model's perplexity on a given dataset iterator, we use the perplexity value on the validation set to select the best model.
- **Decoding**
  5. `predict`: Generates the target sequence given a list of source tokens using beam search decoding. Note that here you can assume the batch size to be 1 for simplicity.

[64]:
```
#TODO - implement the `AttnEncoderDecoder` class.

#################### BeamSearch ####################
MAX_T = 15
class Beam():
  """
  Helper class for storing a hypothesis, its score and its decoder hidden state.
  """
```

```
  def __init__(self, decoder_state, tokens, score):
    self.decoder_state = decoder_state
    self.tokens = tokens
    self.score = score

class BeamSearcher():
  """
  Main class for beam search.
  """
  def __init__(self, model):
    self.model = model
    self.bos_id = model.bos_id
    self.eos_id = model.eos_id
    self.padding_id_src = model.padding_id_src
    self.V = model.V_tgt


  def beam_search(self, src, src_lengths, K, max_T=MAX_T):
    """
    Performs beam search decoding.
    Arguments:
        src: src batch of size (1, max_src_len)
        src_lengths: src lengths of size (1)
        K: beam size
        max_T: max possible target length considered
    Returns:
        a list of token ids and a list of attentions
    """
    finished = []
    all_attns = []
    # Initialize the beam
    self.model.eval()
    # fill in `memory_bank`, `encoder_final_state`, and `init_beam`
    (memory_bank, encoder_final_state) = self.model.forward_encoder(src,␣
↪src_lengths)
    bos = torch.tensor([[self.model.bos_id]], device=device)
    init_beam = Beam(encoder_final_state, bos, 0)
    beams = [init_beam]

    with torch.no_grad():
      for t in range(max_T): # main body of search over time steps

        # Expand each beam by all possible tokens y_{t+1}
        all_total_scores = []
        for beam in beams:
          y_1_to_t, score, decoder_state = beam.tokens, beam.score, beam.
↪decoder_state
```

```
        y_t = y_1_to_t[-1]
        src_mask = src.ne(self.padding_id_src)
        logits, decoder_state, attn = self.model.
↪forward_decoder_incrementally(decoder_state, y_t, memory_bank, src_mask)
        total_scores = logits
        all_total_scores.append(total_scores)
        all_attns.append(attn) # keep attentions for visualization
        beam.decoder_state = decoder_state # update decoder state in the beam
      all_total_scores = torch.stack(all_total_scores) # (K, V) when t>0, (1,␣
↪V) when t=0

      # Find K best next beams
      # The code below has the same functionality as line 6-12, but is more␣
↪efficient
      all_scores_flattened = all_total_scores.view(-1) # K*V when t>0, 1*V␣
↪when t=0
      topk_scores, topk_ids = all_scores_flattened.topk(K, 0)
      beam_ids = topk_ids.div(self.V, rounding_mode='floor')
      next_tokens = topk_ids - beam_ids * self.V
      new_beams = []
      for k in range(K):
        beam_id = beam_ids[k]       # which beam it comes from
        y_t_plus_1 = next_tokens[k] # which y_{t+1}
        score = topk_scores[k]
        beam = beams[beam_id]
        decoder_state = beam.decoder_state
        y_1_to_t = beam.tokens
        y_t = torch.cat((y_1_to_t, y_t_plus_1.view(1,1)))
        new_beam = Beam(decoder_state, y_t, beam.score+score)
        new_beams.append(new_beam)
      beams = new_beams

      # Set aside completed beams
      # move completed beams to `finished` (and remove them from `beams`)
      new_beams = []
      for beam in beams:
          y_1_to_t, score, decoder_state = beam.tokens, beam.score, beam.
↪decoder_state
          y_t = y_1_to_t[-1]
          if y_t.view(1) == self.model.eos_id:
              finished.append(beam)
          else:
              new_beams.append(beam)
      beams = new_beams

      # Break the loop if everything is completed
      if len(beams) == 0:
```

```python
            break

    # Return the best hypothesis
    if len(finished) > 0:
      finished = sorted(finished, key=lambda beam: -beam.score)
      return [token.item() for token in finished[0].tokens], all_attns
    else: # when nothing is finished, return an unfinished hypothesis
      return [token.item() for token in beams[0].tokens], all_attns

##################### attention #####################
def attention(batched_Q, batched_K, batched_V, mask=None):
  """
  Performs the attention operation and returns the attention matrix
  `batched_A` and the context matrix `batched_C` using queries
  `batched_Q`, keys `batched_K`, and values `batched_V`.

  Arguments:
      batched_Q: (bsz, q_len, D)
      batched_K: (bsz, k_len, D)
      batched_V: (bsz, k_len, D)
      mask: (bsz, q_len, k_len). An optional boolean mask *disallowing*
            attentions where the mask value is *`False`*.
  Returns:
      batched_A: the normalized attention scores (bsz, q_len, k_len)
      batched_C: a tensor of size (bsz, q_len, D).
  """
  # Check sizes
  D = batched_Q.size(-1)
  bsz = batched_Q.size(0)
  q_len = batched_Q.size(1)
  k_len = batched_K.size(1)
  assert batched_K.size(-1) == D and batched_V.size(-1) == D
  assert batched_K.size(0) == bsz and batched_V.size(0) == bsz
  assert batched_V.size(1) == k_len
  if mask is not None:
    assert mask.size() == torch.Size([bsz, q_len, k_len])

  K_T = torch.transpose(batched_K, dim0=1, dim1=2)
  temp = torch.bmm(batched_Q, K_T)
  if(mask is not None):
    mask = mask == False
    temp = temp.masked_fill_(mask, -float("Inf"))
  batched_A = torch.softmax(temp, dim=-1)
  batched_C = torch.bmm(batched_A, batched_V)

  # Verify that things sum up to one properly.
  assert torch.all(torch.isclose(batched_A.sum(-1),
```

```
                                  torch.ones(bsz, q_len).to(device)))
  return batched_A, batched_C


##################### AttnEncoderDecoder #####################
class AttnEncoderDecoder(nn.Module):
  def __init__(self, hf_src_tokenizer, hf_tgt_tokenizer, hidden_size=64,␣
↪layers=3):
    """
    Initializer. Creates network modules and loss function.
    Arguments:
        hf_src_tokenizer: hf src tokenizer
        hf_tgt_tokenizer: hf tgt tokenizer
        hidden_size: hidden layer size of both encoder and decoder
        layers: number of layers of both encoder and decoder
    """
    super().__init__()
    self.hf_src_tokenizer = hf_src_tokenizer
    self.hf_tgt_tokenizer = hf_tgt_tokenizer
    self.src_vocab = src_tokenizer.get_vocab()
    self.tgt_vocab = tgt_tokenizer.get_vocab()


    # Keep the vocabulary sizes available
    self.V_src = len(self.hf_src_tokenizer)
    self.V_tgt = len(self.hf_tgt_tokenizer)

    # Get special word ids
    self.padding_id_src = self.src_vocab[pad_token]
    self.padding_id_tgt = self.tgt_vocab[pad_token]
    self.bos_id = self.tgt_vocab[bos_token]
    self.eos_id = self.tgt_vocab[eos_token]

    # Keep hyper-parameters available
    self.embedding_size = hidden_size
    self.hidden_size = hidden_size
    self.layers = layers

    # Create essential modules
    self.word_embeddings_src = nn.Embedding(self.V_src, self.embedding_size)
    self.word_embeddings_tgt = nn.Embedding(self.V_tgt, self.embedding_size)

    # RNN cells
    self.encoder_rnn = nn.LSTM(
      input_size    = self.embedding_size,
      hidden_size   = hidden_size // 2, # to match decoder hidden size
      num_layers    = layers,
```

```
      batch_first=True,
      bidirectional = True              # bidirectional encoder
    )
    self.decoder_rnn = nn.LSTM(
      input_size    = self.embedding_size,
      hidden_size   = hidden_size,
      num_layers    = layers,
      batch_first=True,
      bidirectional = False             # unidirectional decoder
    )

    # Final projection layer
    self.hidden2output = nn.Linear(2*hidden_size, self.V_tgt) # project the␣
↪concatenation to logits

    # Create loss function
    self.loss_function = nn.CrossEntropyLoss(reduction='sum',
                                             ignore_index=self.padding_id_tgt)

  def forward_encoder(self, src, src_lengths):
    """
    Encodes source words `src`.
    Arguments:
        src: src batch of size (bsz, max_src_len)
        src_lengths: src lengths of size (bsz)
    Returns:
        memory_bank: a tensor of size (bsz, src_len, hidden_size)
        (final_state, context): `final_state` is a tuple (h, c) where h/c is of␣
↪size
                                (layers, bsz, hidden_size), and `context` is␣
↪`None`.
    """
    word_embeddings = self.word_embeddings_src(src)

    packed_embeddings = pack(word_embeddings, src_lengths.cpu().numpy(),␣
↪batch_first=True, enforce_sorted=False)

    out, (h,c) = self.encoder_rnn(packed_embeddings)

    h = h.view(2, self.layers, src.shape[0], self.hidden_size//2 )
    c = c.view(2, self.layers, src.shape[0], self.hidden_size//2 )
    h = torch.cat((h[0], h[1]), dim=2 )
    c = torch.cat((c[0], c[1]), dim=2 )

    memory_bank,_ = unpack(out, batch_first=True, padding_value=self.
↪padding_id_src,total_length=src.shape[1])
    final_state = (h, c)
```

```
    context = None

    return memory_bank, (final_state, context)

  def forward_decoder(self, encoder_final_state, tgt_in, memory_bank, src_mask):
    """
    Decodes based on encoder final state, memory bank, src_mask, and ground␣
↪truth
    target words.
    Arguments:
        encoder_final_state: (final_state, None) where final_state is the␣
↪encoder
                             final state used to initialize decoder. None is the
                             initial context (there's no previous context at the
                             first step).
        tgt_in: a tensor of size (bsz, tgt_len)
        memory_bank: a tensor of size (bsz, src_len, hidden_size), encoder␣
↪outputs
                     at every position
        src_mask: a tensor of size (bsz, src_len): a boolean tensor, `False`␣
↪where
                  src is padding (we disallow decoder to attend to those␣
↪places).
    Returns:
        Logits of size (bsz, tgt_len, V_tgt) (before the softmax operation)
    """
    max_tgt_length = tgt_in.size(1)

    # Initialize decoder state, note that it's a tuple (state, context) here
    decoder_states = encoder_final_state

    all_logits = []
    for i in range(max_tgt_length):
      logits, decoder_states, attn = \
        self.forward_decoder_incrementally(decoder_states,
                                           tgt_in[:, i],
                                           memory_bank,
                                           src_mask,
                                           normalize=False)
      all_logits.append(logits)             # list of bsz, vocab_tgt
    all_logits = torch.stack(all_logits, 1) # bsz, tgt_len, vocab_tgt
    return all_logits

  def forward(self, src, src_lengths, tgt_in):
    """
    Performs forward computation, returns logits.
    Arguments:
```

```
        src: src batch of size (bsz, max_src_len)
        src_lengths: src lengths of size (bsz)
        tgt_in:  a tensor of size (bsz, tgt_len)
    """
    src_mask = src.ne(self.padding_id_src) # bsz, max_src_len
    # Forward encoder
    memory_bank, encoder_final_state = self.forward_encoder(src, src_lengths)
    # Forward decoder
    logits = self.forward_decoder(encoder_final_state, tgt_in, memory_bank,␣
↪src_mask)
    return logits

def forward_decoder_incrementally(self, prev_decoder_states, tgt_in_onestep,
                                  memory_bank, src_mask,
                                  normalize=True):
    """
    Forward the decoder for a single step with token `tgt_in_onestep`.
    This function will be used both in `forward_decoder` and in beam search.
    Note that bsz can be greater than 1.
    Arguments:
        prev_decoder_states: a tuple (prev_decoder_state, prev_context).␣
↪`prev_context`
                             is `None` for the first step
        tgt_in_onestep: a tensor of size (bsz), tokens at one step
        memory_bank: a tensor of size (bsz, src_len, hidden_size), encoder␣
↪outputs
                     at every position
        src_mask: a tensor of size (bsz, src_len): a boolean tensor, `False`␣
↪where
                  src is padding (we disallow decoder to attend to those␣
↪places).
        normalize: use log_softmax to normalize or not. Beam search needs to␣
↪normalize,
                   while `forward_decoder` does not
    Returns:
        logits: log probabilities for `tgt_in_token` of size (bsz, V_tgt)
        decoder_states: (`decoder_state`, `context`) which will be used for the
                        next incremental update
        attn: normalized attention scores at this step (bsz, src_len)
    """
    prev_decoder_state, prev_context = prev_decoder_states
    word_embeddings = self.word_embeddings_tgt(torch.unsqueeze(tgt_in_onestep,␣
↪1))

    if prev_context != None:
        word_embeddings += prev_context
```

```
    rnn_out, rnn_states = self.decoder_rnn(word_embeddings, prev_decoder_state)
    attn, context = attention(rnn_out, memory_bank, memory_bank, torch.
↪unsqueeze(src_mask, 1))

    dec_out_e_cntxt = torch.cat((rnn_out, context), dim=2)
    logits = self.hidden2output(dec_out_e_cntxt)
    decoder_states = (rnn_states, context)
    #decoder_states = (h, context)
    if normalize:
      logits = torch.log_softmax(logits, dim=-1)
    return logits, decoder_states, attn.squeeze(1)

  def evaluate_ppl(self, iterator):
    """Returns the model's perplexity on a given dataset `iterator`."""
    # Switch to eval mode
    self.eval()
    total_loss = 0
    total_words = 0
    for batch in iterator:
      # Input and target
      src = batch['src_ids']              # bsz, max_src_len
      src_lengths = batch['src_lengths']  # bsz
      tgt_in = batch['tgt_ids'][:, :-1] # Remove <eos> for decode input␣
↪(y_0=<bos>, y_1, y_2)
      tgt_out = batch['tgt_ids'][:, 1:] # Remove <bos> as target        (y_1,␣
↪y_2, y_3=<eos>)
      # Forward to get logits
      logits = self.forward(src, src_lengths, tgt_in) # bsz, tgt_len, V_tgt
      # Compute cross entropy loss
      loss = self.loss_function(logits.reshape(-1, self.V_tgt), tgt_out.
↪reshape(-1))
      total_loss += loss.item()
      total_words += tgt_out.ne(self.padding_id_tgt).float().sum().item()
    return math.exp(total_loss/total_words)

  def train_all(self, train_iter, val_iter, epochs=10, learning_rate=0.001):
    """Train the model."""
    # Switch the module to training mode
    self.train()
    # Use Adam to optimize the parameters
    optim = torch.optim.Adam(self.parameters(), lr=learning_rate)
    best_validation_ppl = float('inf')
    best_model = None
    # Run the optimization for multiple epochs
    for epoch in range(epochs):
      total_words = 0
```

```
      total_loss = 0.0
      for batch in tqdm(train_iter):
        # Zero the parameter gradients
        self.zero_grad()
        # Input and target
        tgt = batch['tgt_ids']              # bsz, max_tgt_len
        src = batch['src_ids']              # bsz, max_src_len
        src_lengths = batch['src_lengths']  # bsz
        tgt_in = tgt[:, :-1].contiguous() # Remove <eos> for decode input (y_0=<bos>, y_1, y_2)
        tgt_out = tgt[:, 1:].contiguous() # Remove <bos> as target        (y_1, y_2, y_3=<eos>)
        bsz = tgt.size(0)
        # Run forward pass and compute loss along the way.
        logits = self.forward(src, src_lengths, tgt_in)
        loss = self.loss_function(logits.view(-1, self.V_tgt), tgt_out.view(-1))
        # Training stats
        num_tgt_words = tgt_out.ne(self.padding_id_tgt).float().sum().item()
        total_words += num_tgt_words
        total_loss += loss.item()
        # Perform backpropagation
        loss.div(bsz).backward()
        optim.step()

      # Evaluate and track improvements on the validation dataset
      validation_ppl = self.evaluate_ppl(val_iter)
      self.train()
      if validation_ppl < best_validation_ppl:
        best_validation_ppl = validation_ppl
        self.best_model = copy.deepcopy(self.state_dict())
      epoch_loss = total_loss / total_words
      print (f'Epoch: {epoch} Training Perplexity: {math.exp(epoch_loss):.4f} '
             f'Validation Perplexity: {validation_ppl:.4f}')

  def predict(self, tokens, K=1, max_T=400):
    beam_searcher = BeamSearcher(model)
    src_ids = self.hf_src_tokenizer(tokens).input_ids
    src_lengths= torch.IntTensor([len(src_ids)]).to(device)
    src_max_length = max(src_lengths)
    src = torch.zeros(1,src_lengths).long().fill_(self.padding_id_src).to(device)
    src[0][:len(src_ids)] = torch.LongTensor(src_ids).to(device)
    prediction, _ = beam_searcher.beam_search(src, src_lengths, K, max_T)
    prediction = hf_tgt_tokenizer.decode(prediction, skip_special_tokens=True)
    # Strip <bos> and <eos>
    prediction = prediction.lstrip('<bos>').rstrip('<eos>').strip()
    return prediction
```

We provide the recommended hyperparameters for the final model in the script below, but you are free to tune the hyperparameters or change any part of the provided code.

> For quick debugging, we recommend starting with smaller models (by using a very small `hidden_size`), and only a single epoch. If the model runs smoothly, then you can train the full model on GPUs.

[65]:
```python
EPOCHS = 15 # epochs; we recommend starting with a smaller number like 1
LEARNING_RATE = 1e-4 # learning rate

# Instantiate and train classifier
model = AttnEncoderDecoder(hf_src_tokenizer, hf_tgt_tokenizer,
  hidden_size    = 1024,
  layers         = 1,
).to(device)

model.train_all(train_iter, val_iter, epochs=EPOCHS,␣
 ↪learning_rate=LEARNING_RATE)
model.load_state_dict(model.best_model)

# Evaluate model performance, the expected value should be < 1.2
print (f'Validation perplexity: {model.evaluate_ppl(val_iter):.3f}')
```

```
100%|      | 229/229 [01:40<00:00,  2.27it/s]

Epoch: 0 Training Perplexity: 4.1195 Validation Perplexity: 1.6977

100%|      | 229/229 [01:39<00:00,  2.31it/s]

Epoch: 1 Training Perplexity: 1.4546 Validation Perplexity: 1.3618

100%|      | 229/229 [01:37<00:00,  2.34it/s]

Epoch: 2 Training Perplexity: 1.2739 Validation Perplexity: 1.2620

100%|      | 229/229 [01:37<00:00,  2.34it/s]

Epoch: 3 Training Perplexity: 1.1957 Validation Perplexity: 1.2016

100%|      | 229/229 [01:39<00:00,  2.30it/s]

Epoch: 4 Training Perplexity: 1.1509 Validation Perplexity: 1.1726

100%|      | 229/229 [01:35<00:00,  2.39it/s]

Epoch: 5 Training Perplexity: 1.1203 Validation Perplexity: 1.1478

100%|      | 229/229 [01:37<00:00,  2.36it/s]

Epoch: 6 Training Perplexity: 1.0984 Validation Perplexity: 1.1386

100%|      | 229/229 [01:36<00:00,  2.36it/s]

Epoch: 7 Training Perplexity: 1.0817 Validation Perplexity: 1.1221

100%|      | 229/229 [01:36<00:00,  2.38it/s]
```

```
Epoch: 8 Training Perplexity: 1.0675 Validation Perplexity: 1.1166

100%|      | 229/229 [01:38<00:00,  2.33it/s]

Epoch: 9 Training Perplexity: 1.0572 Validation Perplexity: 1.1098

100%|      | 229/229 [01:34<00:00,  2.42it/s]

Epoch: 10 Training Perplexity: 1.0513 Validation Perplexity: 1.1062

100%|      | 229/229 [01:36<00:00,  2.38it/s]

Epoch: 11 Training Perplexity: 1.0428 Validation Perplexity: 1.1003

100%|      | 229/229 [01:37<00:00,  2.35it/s]

Epoch: 12 Training Perplexity: 1.0373 Validation Perplexity: 1.0967

100%|      | 229/229 [01:37<00:00,  2.35it/s]

Epoch: 13 Training Perplexity: 1.0320 Validation Perplexity: 1.1016

100%|      | 229/229 [01:37<00:00,  2.36it/s]

Epoch: 14 Training Perplexity: 1.0294 Validation Perplexity: 1.0937
Validation perplexity: 1.094
```

With a trained model, we can convert questions to SQL statements. We recommend making sure that the model can generate at least reasonable results on the examples from before, before evaluating on the full test set.

[66]:
```
def seq2seq_trial(sentence, gold_sql):
  print("Sentence: ", sentence, "\n")

  predicted_sql = model.predict(sentence, K=1, max_T=400)
  print("Predicted SQL:\n\n", predicted_sql, "\n")

  if verify(predicted_sql, gold_sql, silent=False):
    print ('Correct!')
  else:
    print ('Incorrect!')
```

[67]:
```
seq2seq_trial(example_1, gold_sql_1)
```

```
Sentence:  flights from phoenix to milwaukee

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1, airport_service airport_service_2, city city_2
WHERE flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'PHOENIX'
AND flight_1.to_airport = airport_service_2.airport_code AND
airport_service_2.city_code = city_2.city_code AND city_2.city_name =
```

```
'MILWAUKEE'

Predicted DB result:

 [(108086,), (108087,), (301763,), (301764,), (301765,), (301766,), (302323,),
(304881,), (310619,), (310620,)]

Gold DB result:

 [(108086,), (108087,), (301763,), (301764,), (301765,), (301766,), (302323,),
(304881,), (310619,), (310620,)]

Correct!
```

[68]:

```
seq2seq_trial(example_2, gold_sql_2)
```

```
Sentence:  i would like a united flight

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1 WHERE flight_1.airline_code = 'UA' AND
flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'DENVER'

Predicted DB result:

 [(100094,), (100099,), (100699,), (100703,), (100704,), (100705,), (100706,),
(101082,), (101083,), (101084,)]

Gold DB result:

 [(100094,), (100099,), (100145,), (100158,), (100164,), (100167,), (100169,),
(100203,), (100204,), (100296,)]

Incorrect!
```

[69]:

```
seq2seq_trial(example_3, gold_sql_3)
```

```
Sentence:  i would like a flight between boston and dallas

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1, airport_service airport_service_2, city city_2
WHERE flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'BOSTON'
AND flight_1.to_airport = airport_service_2.airport_code AND
```

```
airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'DALLAS'

Predicted DB result:

 [(103171,), (103172,), (103173,), (103174,), (103175,), (103176,), (103177,),
(103178,), (103179,), (103180,)]

Gold DB result:

 [(103171,), (103172,), (103173,), (103174,), (103175,), (103176,), (103177,),
(103178,), (103179,), (103180,)]

Correct!
```

[70]:
```
seq2seq_trial(example_4, gold_sql_4)
```

```
Sentence:  show me the united flights from denver to baltimore

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1, airport_service airport_service_2, city city_2
WHERE flight_1.airline_code = 'UA' AND ( flight_1.from_airport =
airport_service_1.airport_code AND airport_service_1.city_code =
city_1.city_code AND city_1.city_name = 'DENVER' AND flight_1.to_airport =
airport_service_2.airport_code AND airport_service_2.city_code =
city_2.city_code AND city_2.city_name = 'BALTIMORE' )

Predicted DB result:

 [(101231,), (101233,), (305983,)]

Gold DB result:

 [(101231,), (101233,), (305983,)]

Correct!
```

[71]:
```
seq2seq_trial(example_5, gold_sql_5)
```

```
Sentence:  show flights from cleveland to miami that arrive before 4pm

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1, airport_service airport_service_2, city city_2
WHERE flight_1.from_airport = airport_service_1.airport_code AND
airport_service_1.city_code = city_1.city_code AND city_1.city_name =
```

'CLEVELAND' AND (flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'MIAMI' AND flight_1.arrival_time < 1600)

Predicted DB result:

 [(107698,), (301117,)]

Gold DB result:

 [(107698,), (301117,)]

Correct!

[72]:
```
seq2seq_trial(example_6, gold_sql_6b)
```

Sentence: okay how about a flight on sunday from tampa to charlotte

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2, days days_1, date_day date_day_1 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'TAMPA' AND (flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'CHARLOTTE' AND flight_1.flight_days = days_1.days_code AND days_1.day_name = date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 8 AND date_day_1.day_number = 27)

Predicted DB result:

 [(101860,), (101861,), (101862,), (101863,), (101864,), (101865,), (305231,)]

Gold DB result:

 [(101860,), (101861,), (101862,), (101863,), (101864,), (101865,), (305231,)]

Correct!

[73]:
```
seq2seq_trial(example_7, gold_sql_7b)
```

Sentence: list all flights going from boston to atlanta that leaves before 7 am on thursday

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service

```
airport_service_1, city city_1, airport_service airport_service_2, city city_2,
days days_1, date_day date_day_1 WHERE flight_1.from_airport =
airport_service_1.airport_code AND airport_service_1.city_code =
city_1.city_code AND city_1.city_name = 'BOSTON' AND ( flight_1.to_airport =
airport_service_2.airport_code AND airport_service_2.city_code =
city_2.city_code AND city_2.city_name = 'ATLANTA' AND ( flight_1.flight_days =
days_1.days_code AND days_1.day_name = date_day_1.day_name AND date_day_1.year =
1991 AND date_day_1.month_number = 5 AND date_day_1.day_number = 24 AND
flight_1.departure_time < 700 ) )

Predicted DB result:

 [(100014,)]

Gold DB result:

 [(100014,)]

Correct!
```

[74]:

```
seq2seq_trial(example_8, gold_sql_8)
```

```
Sentence:  list the flights from dallas to san francisco on american airlines

Predicted SQL:

 SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service
airport_service_1, city city_1, airport_service airport_service_2, city city_2
WHERE flight_1.airline_code = 'AA' AND ( flight_1.from_airport =
airport_service_1.airport_code AND airport_service_1.city_code =
city_1.city_code AND city_1.city_name = 'DALLAS' AND flight_1.to_airport =
airport_service_2.airport_code AND airport_service_2.city_code =
city_2.city_code AND city_2.city_name = 'SAN FRANCISCO' )

Predicted DB result:

 [(108452,), (108454,), (108456,), (111083,), (111085,), (111086,), (111090,),
(111091,), (111092,), (111094,)]

Gold DB result:

 [(108452,), (108454,), (108456,), (111083,), (111085,), (111086,), (111090,),
(111091,), (111092,), (111094,)]

Correct!
```

6.1.1 Evaluation

Now we are ready to run the full evaluation. A proper implementation should reach more than 35% precision/recall/F1.

[75]:
```
def seq2seq_predictor(tokens):
  prediction = model.predict(tokens, K=1, max_T=400)
  return prediction
```

[76]:
```
precision, recall, f1 = evaluate(seq2seq_predictor, test_data, num_examples=0)
print(f"precision: {precision:3.2f}")
print(f"recall:    {recall:3.2f}")
print(f"F1:        {f1:3.2f}")
```

```
332it [02:15,  2.45it/s]

precision: 0.39
recall:    0.39
F1:        0.39
```

6.2 Goal 3: Implement a seq2seq model (with cross attention and self attention)

In the previous section, you have implemented a seq2seq model with attention. The attention mechanism used in that section is usually referred to as "cross-attention", as at each decoding step, the decoder attends to encoder outputs, enabling a dynamic view on the encoder side as decoding proceeds.

Similarly, we can have a dynamic view on the decoder side as well as decoding proceeds, i.e., the decoder attends to decoder outputs at previous steps. This is called "self attention", and has been found very useful in modern neural architectures such as transformers.

Augment the seq2seq model you implemented before with a decoder self-attention mechanism as class `AttnEncoderDecoder2`. A model diagram can be found below:

At each decoding step, the decoder LSTM first produces an output state o_t, then it attends to all previous output states $o_1, \ldots, o_{t-1}$ (decoder self-attention). You need to special case the first decoding step to not perform self-attention, as there are no previous decoder states. The attention result is added to o_t itself and the sum is used as q_t to attend to the encoder side (encoder-decoder cross-attention). The rest of the model is the same as encoder-decoder with attention.

[89]:
```
#TODO - implement the `AttnEncoderDecoder2` class.
class AttnEncoderDecoder2(nn.Module):
  def __init__(self, hf_src_tokenizer, hf_tgt_tokenizer, hidden_size=64,␣
 ↪layers=3):
    """
    Initializer. Creates network modules and loss function.
    Arguments:
        hf_src_tokenizer: hf src tokenizer
        hf_tgt_tokenizer: hf tgt tokenizer
```

```
        hidden_size: hidden layer size of both encoder and decoder
        layers: number of layers of both encoder and decoder
    """
    super().__init__()
    self.hf_src_tokenizer = hf_src_tokenizer
    self.hf_tgt_tokenizer = hf_tgt_tokenizer
    self.src_vocab = src_tokenizer.get_vocab()
    self.tgt_vocab = tgt_tokenizer.get_vocab()


    # Keep the vocabulary sizes available
    self.V_src = len(self.hf_src_tokenizer)
    self.V_tgt = len(self.hf_tgt_tokenizer)

    # Get special word ids
    self.padding_id_src = self.src_vocab[pad_token]
    self.padding_id_tgt = self.tgt_vocab[pad_token]
    self.bos_id = self.tgt_vocab[bos_token]
    self.eos_id = self.tgt_vocab[eos_token]

    # Keep hyper-parameters available
    self.embedding_size = hidden_size
    self.hidden_size = hidden_size
    self.layers = layers

    # Create essential modules
    self.word_embeddings_src = nn.Embedding(self.V_src, self.embedding_size)
    self.word_embeddings_tgt = nn.Embedding(self.V_tgt, self.embedding_size)

    # RNN cells
    self.encoder_rnn = nn.LSTM(
      input_size    = self.embedding_size,
      hidden_size   = hidden_size // 2, # to match decoder hidden size
      num_layers    = layers,
      batch_first=True,
      bidirectional = True              # bidirectional encoder
    )
    self.decoder_rnn = nn.LSTM(
      input_size    = self.embedding_size,
      hidden_size   = hidden_size,
      num_layers    = layers,
      batch_first=True,
      bidirectional = False             # unidirectional decoder
    )

    # Final projection layer
```

```
    self.hidden2output = nn.Linear(2*hidden_size, self.V_tgt) # project the␣
↪concatenation to logits

    # Create loss function
    self.loss_function = nn.CrossEntropyLoss(reduction='sum',
                                             ignore_index=self.padding_id_tgt)

  def forward_encoder(self, src, src_lengths):
    """
    Encodes source words `src`.
    Arguments:
        src: src batch of size (bsz, max_src_len)
        src_lengths: src lengths of size (bsz)
    Returns:
        memory_bank: a tensor of size (bsz, src_len, hidden_size)
        (final_state, context): `final_state` is a tuple (h, c) where h/c is of␣
↪size
                                (layers, bsz, hidden_size), and `context` is␣
↪`None`.
    """
    word_embeddings = self.word_embeddings_src(src)

    packed_embeddings = pack(word_embeddings, src_lengths.cpu().numpy(),␣
↪batch_first=True, enforce_sorted=False)

    out, (h,c) = self.encoder_rnn(packed_embeddings)

    h = h.view(2, self.layers, src.shape[0], self.hidden_size//2 )
    c = c.view(2, self.layers, src.shape[0], self.hidden_size//2 )
    h = torch.cat((h[0], h[1]), dim=2 )
    c = torch.cat((c[0], c[1]), dim=2 )

    memory_bank,_ = unpack(out, batch_first=True, padding_value=self.
↪padding_id_src,total_length=src.shape[1])
    final_state = (h, c)
    context = None

    return memory_bank, (final_state, context, None)

  def forward_decoder(self, encoder_final_state, tgt_in, memory_bank, src_mask):
    """
    Decodes based on encoder final state, memory bank, src_mask, and ground␣
↪truth
    target words.
    Arguments:
```

```
        encoder_final_state: (final_state, None) where final_state is the␣
↪encoder
                              final state used to initialize decoder. None is the
                              initial context (there's no previous context at the
                              first step).
        tgt_in: a tensor of size (bsz, tgt_len)
        memory_bank: a tensor of size (bsz, src_len, hidden_size), encoder␣
↪outputs
                      at every position
        src_mask: a tensor of size (bsz, src_len): a boolean tensor, `False`␣
↪where
                   src is padding (we disallow decoder to attend to those␣
↪places).
    Returns:
        Logits of size (bsz, tgt_len, V_tgt) (before the softmax operation)
    """
    max_tgt_length = tgt_in.size(1)

    # Initialize decoder state, note that it's a tuple (state, context) here
    decoder_states = encoder_final_state

    all_logits = []
    for i in range(max_tgt_length):
      logits, decoder_states, attn = \
        self.forward_decoder_incrementally(decoder_states,
                                            tgt_in[:, i],
                                            memory_bank,
                                            src_mask,
                                            normalize=False)
      all_logits.append(logits)             # list of bsz, vocab_tgt
    all_logits = torch.stack(all_logits, 1) # bsz, tgt_len, vocab_tgt
    return all_logits

  def forward(self, src, src_lengths, tgt_in):
    """
    Performs forward computation, returns logits.
    Arguments:
        src: src batch of size (bsz, max_src_len)
        src_lengths: src lengths of size (bsz)
        tgt_in:  a tensor of size (bsz, tgt_len)
    """
    src_mask = src.ne(self.padding_id_src) # bsz, max_src_len
    # Forward encoder
    memory_bank, encoder_final_state = self.forward_encoder(src, src_lengths)
    # Forward decoder
    logits = self.forward_decoder(encoder_final_state, tgt_in, memory_bank,␣
↪src_mask)
```

```
    return logits

def forward_decoder_incrementally(self, prev_decoder_states, tgt_in_onestep,
                                  memory_bank, src_mask,
                                  normalize=True):
    """
    Forward the decoder for a single step with token `tgt_in_onestep`.
    This function will be used both in `forward_decoder` and in beam search.
    Note that bsz can be greater than 1.
    Arguments:
        prev_decoder_states: a tuple (prev_decoder_state, prev_context,␣
↪past_outputs). `prev_context`
                             is `None` for the first step
        tgt_in_onestep: a tensor of size (bsz), tokens at one step
        memory_bank: a tensor of size (bsz, src_len, hidden_size), encoder␣
↪outputs
                     at every position
        src_mask: a tensor of size (bsz, src_len): a boolean tensor, `False`␣
↪where
                  src is padding (we disallow decoder to attend to those␣
↪places).
        normalize: use log_softmax to normalize or not. Beam search needs to␣
↪normalize,
                   while `forward_decoder` does not
    Returns:
        logits: log probabilities for `tgt_in_token` of size (bsz, V_tgt)
        decoder_states: (`decoder_state`, `context`) which will be used for the
                        next incremental update
        attn: normalized attention scores at this step (bsz, src_len)
    """
    prev_decoder_state, prev_context, past_outputs = prev_decoder_states
    word_embeddings = self.word_embeddings_tgt(torch.unsqueeze(tgt_in_onestep,␣
↪1))

    if prev_context != None:
        word_embeddings += prev_context

    rnn_out, rnn_states = self.decoder_rnn(word_embeddings, prev_decoder_state)

    # Compute self-attention if not none
    if past_outputs is not None:
        attn, context = attention(rnn_out, past_outputs,past_outputs, None)
        past_outputs = torch.cat((past_outputs, rnn_out), dim=1)
        rnn_out = context + rnn_out
    else:
        past_outputs = rnn_out
```

```
    attn, context = attention(rnn_out, memory_bank, memory_bank, torch.
↪unsqueeze(src_mask, 1))

    dec_out_e_cntxt = torch.cat((rnn_out, context), dim=2)
    logits = self.hidden2output(dec_out_e_cntxt)
    decoder_states = (rnn_states, context, past_outputs)
    #decoder_states = (h, context)
    if normalize:
      logits = torch.log_softmax(logits, dim=-1)
    return logits, decoder_states, attn.squeeze(1)

  def evaluate_ppl(self, iterator):
    """Returns the model's perplexity on a given dataset `iterator`."""
    # Switch to eval mode
    self.eval()
    total_loss = 0
    total_words = 0
    for batch in iterator:
      # Input and target
      src = batch['src_ids']              # bsz, max_src_len
      src_lengths = batch['src_lengths']  # bsz
      tgt_in = batch['tgt_ids'][:, :-1] # Remove <eos> for decode input␣
↪(y_0=<bos>, y_1, y_2)
      tgt_out = batch['tgt_ids'][:, 1:] # Remove <bos> as target        (y_1,␣
↪y_2, y_3=<eos>)
      # Forward to get logits
      logits = self.forward(src, src_lengths, tgt_in) # bsz, tgt_len, V_tgt
      # Compute cross entropy loss
      loss = self.loss_function(logits.reshape(-1, self.V_tgt), tgt_out.
↪reshape(-1))
      total_loss += loss.item()
      total_words += tgt_out.ne(self.padding_id_tgt).float().sum().item()
    return math.exp(total_loss/total_words)

  def train_all(self, train_iter, val_iter, epochs=10, learning_rate=0.001):
    """Train the model."""
    # Switch the module to training mode
    self.train()
    # Use Adam to optimize the parameters
    optim = torch.optim.Adam(self.parameters(), lr=learning_rate)
    best_validation_ppl = float('inf')
    best_model = None
    # Run the optimization for multiple epochs
    for epoch in range(epochs):
      total_words = 0
      total_loss = 0.0
```

```
    for batch in tqdm(train_iter):
      # Zero the parameter gradients
      self.zero_grad()
      # Input and target
      tgt = batch['tgt_ids']               # bsz, max_tgt_len
      src = batch['src_ids']               # bsz, max_src_len
      src_lengths = batch['src_lengths']   # bsz
      tgt_in = tgt[:, :-1].contiguous() # Remove <eos> for decode input␣
↪(y_0=<bos>, y_1, y_2)
      tgt_out = tgt[:, 1:].contiguous() # Remove <bos> as target        (y_1,␣
↪y_2, y_3=<eos>)
      bsz = tgt.size(0)
      # Run forward pass and compute loss along the way.
      logits = self.forward(src, src_lengths, tgt_in)
      loss = self.loss_function(logits.view(-1, self.V_tgt), tgt_out.view(-1))
      # Training stats
      num_tgt_words = tgt_out.ne(self.padding_id_tgt).float().sum().item()
      total_words += num_tgt_words
      total_loss += loss.item()
      # Perform backpropagation
      loss.div(bsz).backward()
      optim.step()

    # Evaluate and track improvements on the validation dataset
    validation_ppl = self.evaluate_ppl(val_iter)
    self.train()
    if validation_ppl < best_validation_ppl:
      best_validation_ppl = validation_ppl
      self.best_model = copy.deepcopy(self.state_dict())
    epoch_loss = total_loss / total_words
    print (f'Epoch: {epoch} Training Perplexity: {math.exp(epoch_loss):.4f} '
           f'Validation Perplexity: {validation_ppl:.4f}')

def predict(self, tokens, K=1, max_T=400):
  beam_searcher = BeamSearcher(model)
  src_ids = self.hf_src_tokenizer(tokens).input_ids
  src_lengths= torch.IntTensor([len(src_ids)]).to(device)
  src_max_length = max(src_lengths)
  src = torch.zeros(1,src_lengths).long().fill_(self.padding_id_src).
↪to(device)
  src[0][:len(src_ids)] = torch.LongTensor(src_ids).to(device)
  prediction, _ = beam_searcher.beam_search(src, src_lengths, K, max_T)
  prediction = hf_tgt_tokenizer.decode(prediction, skip_special_tokens=True)
  # Strip <bos> and <eos>
  prediction = prediction.lstrip('<bos>').rstrip('<eos>').strip()
  return prediction
```

[94]:
```
EPOCHS = 15 # epochs, we recommend starting with a smaller number like 1
LEARNING_RATE = 1e-4 # learning rate

# Instantiate and train classifier
model2 = AttnEncoderDecoder2(hf_src_tokenizer, hf_tgt_tokenizer,
  hidden_size    = 1024,
  layers         = 1,
).to(device)

model2.train_all(train_iter, val_iter, epochs=EPOCHS,␣
 ↪learning_rate=LEARNING_RATE)
model2.load_state_dict(model2.best_model)

# Evaluate model performance, the expected value should be < 1.2
print (f'Validation perplexity: {model2.evaluate_ppl(val_iter):.3f}')
```

```
100%|      | 229/229 [06:42<00:00,  1.76s/it]

Epoch: 0 Training Perplexity: 3.8551 Validation Perplexity: 1.7790

100%|      | 229/229 [06:23<00:00,  1.68s/it]

Epoch: 1 Training Perplexity: 1.5354 Validation Perplexity: 1.4477

100%|      | 229/229 [05:26<00:00,  1.43s/it]

Epoch: 2 Training Perplexity: 1.3439 Validation Perplexity: 1.3207

100%|      | 229/229 [04:57<00:00,  1.30s/it]

Epoch: 3 Training Perplexity: 1.2762 Validation Perplexity: 1.2932

100%|      | 229/229 [06:20<00:00,  1.66s/it]

Epoch: 4 Training Perplexity: 1.2105 Validation Perplexity: 1.2239

100%|      | 229/229 [06:04<00:00,  1.59s/it]

Epoch: 5 Training Perplexity: 1.1734 Validation Perplexity: 1.2059

100%|      | 229/229 [04:57<00:00,  1.30s/it]

Epoch: 6 Training Perplexity: 1.1429 Validation Perplexity: 1.1744

100%|      | 229/229 [06:07<00:00,  1.60s/it]

Epoch: 7 Training Perplexity: 1.1230 Validation Perplexity: 1.1592

100%|      | 229/229 [05:27<00:00,  1.43s/it]

Epoch: 8 Training Perplexity: 1.1092 Validation Perplexity: 1.1530

100%|      | 229/229 [04:33<00:00,  1.20s/it]

Epoch: 9 Training Perplexity: 1.0958 Validation Perplexity: 1.1393

100%|      | 229/229 [04:21<00:00,  1.14s/it]
```

```
Epoch: 10 Training Perplexity: 1.0863 Validation Perplexity: 1.1328

100%|      | 229/229 [04:52<00:00,  1.28s/it]

Epoch: 11 Training Perplexity: 1.0740 Validation Perplexity: 1.1375

100%|      | 229/229 [04:04<00:00,  1.07s/it]

Epoch: 12 Training Perplexity: 1.0680 Validation Perplexity: 1.1245

100%|      | 229/229 [03:59<00:00,  1.05s/it]

Epoch: 13 Training Perplexity: 1.0579 Validation Perplexity: 1.1236

100%|      | 229/229 [04:05<00:00,  1.07s/it]

Epoch: 14 Training Perplexity: 1.0517 Validation Perplexity: 1.1140
Validation perplexity: 1.114
```

6.2.1 Evaluation

Now we are ready to run the full evaluation. A proper implementation should reach more than 35% precision/recall/F1.

[95]:
```
def seq2seq_predictor2(tokens):
  prediction = model2.predict(tokens, K=1, max_T=400)
  return prediction
```

[96]:
```
precision, recall, f1 = evaluate(seq2seq_predictor2, test_data, num_examples=0)
print(f"precision: {precision:3.2f}")
print(f"recall:    {recall:3.2f}")
print(f"F1:        {f1:3.2f}")
```

```
332it [03:15,  1.69it/s]

precision: 0.39
recall:    0.39
F1:        0.39
```

6.3 Goal 4: Use state-of-the-art pretrained transformers

The most recent breakthrough in natural-language processing stems from the use of pretrained transformer models. For example, you might have heard of pretrained transformers such as GPT-3 and BERT. (BERT is already used in Google search.) These models are usually trained on vast amounts of text data using variants of language modeling objectives, and researchers have found that finetuning them on downstream tasks usually results in better performance as compared to training a model from scratch.

In the previous part, you implemented an LSTM-based sequence-to-sequence approach. To "upgrade" the model to be a state-of-the-art pretrained transformer only requires minor modifications.

The pretrained model that we will use is BART, which uses a bidirectional transformer encoder and a unidirectional transformer decoder, as illustrated in the below diagram (image courtesy https://arxiv.org/pdf/1910.13461):

We can see that this model is strikingly similar to the LSTM-based encoder-decoder model we've been using. The only difference is that they use transformers instead of LSTMs. Therefore, we only need to change the modeling parts of the code, as we will see later.

First, we download and load the pretrained BART model from the transformers package by Huggingface. Note that we also need to use the "tokenizer" of BART, which is actually a combination of a tokenizer and a mapping from strings to word ids.

[ ]:
```
pretrained_bart = BartForConditionalGeneration.from_pretrained('facebook/
 ↪bart-base')
bart_tokenizer = BartTokenizer.from_pretrained('facebook/bart-base')
```

We need to reprocess the data using our new tokenizer. Note that we use the same tokenizer for both the source and target.

[ ]:
```
def bart_encode(example):
    example['src_ids'] = bart_tokenizer(example['src']).input_ids[:1024] # BART␣
 ↪model can process at most 1024 tokens
    example['tgt_ids'] = bart_tokenizer(example['tgt']).input_ids[:1024]
    return example

train_bart_data = dataset['train'].map(encode)
val_bart_data = dataset['val'].map(encode)
test_bart_data = dataset['test'].map(encode)

BATCH_SIZE = 1 # batch size for training/validation
TEST_BATCH_SIZE = 1 # batch size for test, we use 1 to make beam search␣
 ↪implementation easier

# we use the same collate function as before
train_iter_bart = torch.utils.data.DataLoader(train_data,
                                         batch_size=BATCH_SIZE,
                                         shuffle=True,
                                         collate_fn=collate_fn)
val_iter_bart = torch.utils.data.DataLoader(val_data,
                                       batch_size=BATCH_SIZE,
                                       shuffle=False,
                                       collate_fn=collate_fn)
test_iter_bart = torch.utils.data.DataLoader(test_data,
                                        batch_size=TEST_BATCH_SIZE,
                                        shuffle=False,
                                        collate_fn=collate_fn)
```

Let's take a look at the batch.

```
[ ]: batch = next(iter(train_iter_bart))
     src_ids = batch['src_ids']
     src_example = src_ids[0]
     print (f"Size of text batch: {src_ids.size()}")
     print (f"First sentence in batch: {src_example}")
     print (f"Length of the third sentence in batch: {len(src_example)}")
     print (f"Converted back to string: {hf_src_tokenizer.decode(src_example)}")

     tgt_ids = batch['tgt_ids']
     tgt_example = tgt_ids[0]
     print (f"Size of sql batch: {tgt_ids.size()}")
     print (f"First sql in batch: {tgt_example}")
     print (f"Converted back to string: {hf_tgt_tokenizer.decode(tgt_example)}")
```

Now we are ready to implement the BART-based approach for the text-to-SQL conversion problem. In the below `BART` class, we have provided the constructer `__init__`, the `forward` function, and the `predict` function. Your job is to implement the main optimization `train_all`, and `evaluate_ppl` for evaluating validation perplexity for model selection.

> Hint: you can use almost the same `train_all` and `evaluate_ppl` function you implemented before, but here a major difference is that due to setting `batch_first=True`, the batched source/target tensors are of size `batch_size x max_len`, as opposed to `max_len x batch_size` in the LSTM-based approach, and you need to make changes in `train_all` and `evaluate_ppl` accordingly.

```
[ ]: #TODO - finish implementing the `BART` class.
     class BART(nn.Module):
       def __init__(self, tokenizer, pretrained_bart):
         """
         Initializer. Creates network modules and loss function.
         Arguments:
             tokenizer: BART tokenizer
             pretrained_bart: pretrained BART
         """
         super(BART, self).__init__()

         self.tokenizer = tokenizer

         self.V_tgt = len(tokenizer)

         # Get special word ids
         self.padding_id_tgt = tokenizer.pad_token_id

         # Create essential modules
         self.bart = pretrained_bart

         # Create loss function
         self.loss_function = nn.CrossEntropyLoss(reduction="sum",
```

```
                                                  ignore_index=self.padding_id_tgt)

  def forward(self, src, src_lengths, tgt_in):
    """
    Performs forward computation, returns logits.
    Arguments:
        src: src batch of size (batch_size, max_src_len)
        src_lengths: src lengths of size (batch_size)
        tgt_in:  a tensor of size (tgt_len, bsz)
    """
    # BART assumes inputs to be batch-first
    # This single function is forwarding both encoder and decoder (w/ cross␣
↪attn),
    # using `input_ids` as encoder inputs, and `decoder_input_ids`
    # as decoder inputs.
    logits = self.bart(input_ids=src,
                       decoder_input_ids=tgt_in,
                       use_cache=False
                      ).logits
    return logits

  def evaluate_ppl(self, iterator):
    """Returns the model's perplexity on a given dataset `iterator`."""
    #TODO - implement this function
    ...
    ppl = ...
    return ppl

  def train_all(self, train_iter, val_iter, epochs=10, learning_rate=0.001):
    """Train the model."""
    #TODO - implement this function
    ...

  def predict(self, tokens, K=1, max_T=400):
    """
    Generates the target sequence given the source sequence using beam search␣
↪decoding.
    Note that for simplicity, we only use batch size 1.
    Arguments:
        tokens: the source sentence.
        max_T: at most proceed this many steps of decoding
    Returns:
        a string of the generated target sentence.
    """
    # Tokenize and map to a list of word ids
    inputs = torch.LongTensor(self.tokenizer([tokens])['input_ids'][:1024]).
↪to(device)
```

```python
    # The `transformers` package provides built-in beam search support
    prediction = self.bart.generate(inputs,
                                    num_beams=K,
                                    max_length=max_T,
                                    early_stopping=True,
                                    no_repeat_ngram_size=0,
                                    decoder_start_token_id=0,
                                    use_cache=True)[0]
    return  self.tokenizer.decode(prediction, skip_special_tokens=True)
```

The code below will kick off training, and evaluate the validation perplexity. You should expect to see a value very close to 1.

[ ]:
```python
EPOCHS = 5 # epochs, we recommend starting with a smaller number like 1
LEARNING_RATE = 1e-5 # learning rate

# Instantiate and train classifier
bart_model = BART(bart_tokenizer,
                  pretrained_bart
).to(device)

bart_model.train_all(train_iter_bart, val_iter_bart, epochs=EPOCHS,
 ↪learning_rate=LEARNING_RATE)
bart_model.load_state_dict(bart_model.best_model)

# Evaluate model performance, the expected value should be < 1.2
print (f'Validation perplexity: {bart_model.evaluate_ppl(val_iter_bart):.3f}')
```

As before, make sure that your model is making reasonable predictions on a few examples before evaluating on the entire test set.

[ ]:
```python
def bart_trial(sentence, gold_sql):
  predicted_sql = bart_model.predict(sentence, K=1, max_T=300)
  print("Predicted SQL:\n\n", predicted_sql, "\n")

  if verify(predicted_sql, gold_sql, silent=False):
    print ('Correct!')
  else:
    print ('Incorrect!')
```

[ ]:
```python
bart_trial(example_1, gold_sql_1)
```

[ ]:
```python
bart_trial(example_2, gold_sql_2)
```

[ ]:
```python
bart_trial(example_3, gold_sql_3)
```

[ ]:
```python
bart_trial(example_4, gold_sql_4)
```

[ ]:
```
bart_trial(example_5, gold_sql_5)
```

[ ]:
```
bart_trial(example_6, gold_sql_6b)
```

[ ]:
```
bart_trial(example_7, gold_sql_7b)
```

[ ]:
```
bart_trial(example_8, gold_sql_8)
```

6.3.1 Evaluation

The code below will evaluate on the entire test set. You should expect to see precision/recall/F1 greater than 40%.

[ ]:
```
def seq2seq_predictor_bart(tokens):
  prediction = bart_model.predict(tokens, K=4, max_T=400)
  return prediction
```

[ ]:
```
precision, recall, f1 = evaluate(seq2seq_predictor_bart, test_data,␣
 ↪num_examples=0)
print(f"precision: {precision:3.2f}")
print(f"recall:    {recall:3.2f}")
print(f"F1:        {f1:3.2f}")
```

7 Discussion

7.1 Goal 5: Compare the pros and cons of rule-based and neural approaches.

Compare the pros and cons of the rule-based approach and the neural approaches with relevant examples from your experiments above. Concerning the accuracy, which approach would you choose to be used in a product? Explain.

we learned that the rule-based approach is quite different than the neural approach in how they handle the data processing.

- the rule-based approach yeilds higher precision whereas the neural approach yeilds better recall.
- the rule-based approach does not need a training phase, however it needs manual structuring for the derivation rules. this manual step requires knowledge and understanding of the language and the spesific problem the model is solving. Whereas the neural approach doesnt require such in-depth understanding and expertise however it needs an enourmus amount of tagged data to be trained on in order to yield the needed results.
- the rule-based approach is much faster than the neural approach. this is because the complexity of of the rule-based apprach over a CNF grammar is O(logn) whereas the neural approach using RNN/LSTM requires at least O(n) and involves complex matrix multiplications.
- the rule-based approach performs better at tasks where converting flight queries from English to SQL is needed, however the neural approach allows adaptbility and can be retrained to work in different domains too, this is because neural models can build rules on their own

based on data, in contrarly to the rule-based which requires a human(or an outside tool) to establish the rules.

- rule-based models can give exact control on dataset focus areas, whereas neural models may be more general and less specific, this is because the neural model is adaptable and universal (we know that we can add spesific rules for the rule-based approach but it will be very hard to reach the same universality as the neural model).
- the neural model can easily add extra layers and increase the hidden size whereas the rule-based approach would demand considerable effort for similar improvements.
- the neural model is scalable, adaptable, has a broafer applicability even thoughr ule-based approaches can guarantee high accuracy for specific tasks.

8 Debrief

Question: We're interested in any thoughts you have about this project segment so that we can improve it for later years, and to inform later segments for this year. Please list any issues that arose or comments you have to improve the project segment. Useful things to comment on might include the following:

- Was the project segment clear or unclear? Which portions?
- Were the readings appropriate background for the project segment?
- Are there additions or changes you think would make the project segment better?

but you should comment on whatever aspects you found especially positive or negative.

Type your answer here, replacing this text.

9 Instructions for submission of the project segment

This project segment should be submitted to Gradescope at `https://rebrand.ly/project4-submit-code` and `https://rebrand.ly/project4-submit-pdf`, which will be made available some time before the due date.

Project segment notebooks are manually graded, not autograded using otter as labs are. (Otter is used within project segment notebooks to synchronize distribution and solution code however.) **We will not run your notebook before grading it.** Instead, we ask that you submit the already freshly run notebook. The best method is to "restart kernel and run all cells", allowing time for all cells to be run to completion. You should submit your code to Gradescope at the code submission assignment at `https://rebrand.ly/project4-submit-code`. Make sure that you are also submitting your `data/grammar` file as part of your solution code as well.

We also request that you **submit a PDF of the freshly run notebook**. The simplest method is to use "Export notebook to PDF", which will render the notebook to PDF via LaTeX. If that doesn't work, the method that seems to be most reliable is to export the notebook as HTML (if you are using Jupyter Notebook, you can do so using `File -> Print Preview`), open the HTML in a browser, and print it to a file. Then make sure to add the file to your git commit. Please name the file the same name as this notebook, but with a `.pdf` extension. (Conveniently, the methods just described will use that name by default.) You can then perform a git commit and push and submit the commit to Gradescope at `https://rebrand.ly/project4-submit-pdf`.

10 End of project segment 4