COMS W4705 Spring 25

Homework 4 - Semantic Role Labelling with BERT

The goal of this assignment is to train and evaluate a PropBank-style semantic role labeling (SRL) system. Following (Collobert et al. 2011) and others, we will treat this problem as a sequence-labeling task. For each input token, the system will predict a B-I-O tag, as illustrated in the following example:

The	judge	scheduled	to	preside	over	his	trial	was	removed	from	the	case	today	.
B-ARG1	I-ARG1	B-V schedule.01	B-ARG2	I-ARG2	I-ARG2	I-ARG2	I-ARG2	O	O	O	O	O	O	O

Note that the same sentence may have multiple annotations for different predicates

The	judge	scheduled	to	preside	over	his	trial	was	removed	from	the	case	today	.
B-ARG1	I-ARG1	I-ARG1	I-ARG1	I-ARG1	I-ARG1	I-ARG1	I-ARG1	O	B-V remove.01	B-ARG2	I-ARG2	I-ARG2	B-ARGM-TMP	O

and not all predicates need to be verbs

The	judge	scheduled	to	preside	over	his	trial	was	removed	from	the	case	today	.
O	O	O	O	O	O	B-ARG1	B-V try.02	O	O	O	O	O	O	O

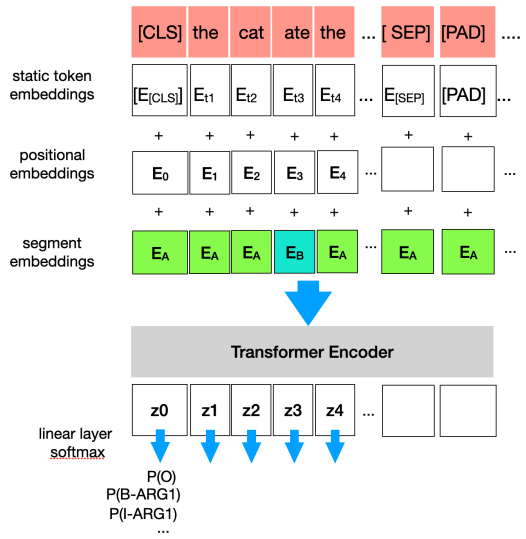
The SRL system will be implemented in PyTorch. We will use BERT (in the implementation provided by the Huggingface transformers library) to compute contextualized token representations and a custom classification head to predict semantic roles. We will fine-tune the pretrained BERT model on the SRL task.

Overview of the Approach

The model we will train is pretty straightforward. Essentially, we will just encode the sentence with BERT, then take the contextualized embedding for each token and feed it into a classifier to predict the corresponding tag.

Because we are only working on argument identification and labeling (not predicate identification), it is essentially that we tell the model where the predicate is. This can be accomplished in various ways. The approach we will choose here repurposes Bert's *segment embeddings*.

Recall that BERT is trained on two input sentences, seperated by [SEP], and on a next-sentence-prediction objective (in addition to the masked LM objective). To help BERT comprehend which sentence a given token belongs to, the original BERT uses a segment embedding, using A for the first sentene, and B for the second sentence 2. Because we are labeling only a single sentence at a time, we can use the segment embeddings to indicate the predicate position instead: The predicate is labeled as segment B (1) and all other tokens will be labeled as segment A (0).



Setup: GCP, Jupyter, PyTorch, GPU

To make sure that PyTorch is available and can use the GPU,run the following cell which should return True. If it doesn't, make sure the GPU drivers and CUDA are installed correctly.

GPU support is required for this assignment -- you will not be able to fine-tune BERT on a CPU.

```
import torch
torch.cuda.is_available()
```

```
True
```

Dataset: Ontonotes 5.0 English SRL annotations

We will work with the English part of the Ontonotes 5.0 data. This is an extension of PropBank, using the same type of annotation. Ontonotes contains annotations other than predicate/argument structures, but we will use the PropBank style SRL annotations only. *Important*: This data set is provided to you for use in COMS 4705 only! Columbia is a subscriber to LDC and is allowed to use the data for educational purposes. However, you may not use the dataset in projects unrelated to Columbia teaching or research.

If you haven't done so already, you can download the data here:

```
! wget https://storage.googleapis.com/4705-bert-srl-data/ontonotes_srl.zip
```

```
--2025-04-30 20:51:41--  https://storage.googleapis.com/4705-bert-srl-data/ontonotes_srl.zip
Resolving storage.googleapis.com (storage.googleapis.com)... 172.217.194.207, 142.250.4.207, 172.253.118.207, ...
Connecting to storage.googleapis.com (storage.googleapis.com)|172.217.194.207|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 12369688 (12M) [application/zip]
Saving to: ‘ontonotes_srl.zip’

ontonotes_srl.zip   100%[===================>]  11.80M  5.64MB/s    in 2.1s

2025-04-30 20:51:43 (5.64 MB/s) - ‘ontonotes_srl.zip’ saved [12369688/12369688]
```

```
! unzip ontonotes_srl.zip
```

```
Archive:  ontonotes_srl.zip
  inflating: propbank_dev.tsv
  inflating: propbank_test.tsv
  inflating: propbank_train.tsv
  inflating: role_list.txt
```

The data has been pre-processed in the following format. There are three files:

```
propbank_dev.tsv propbank_test.tsv propbank_train.tsv
```

Each of these files is in a tab-separated value format. A single predicate/argument structure annotation consists of four rows. For example

```
ontonotes/bc/cnn/00/cnn_0000.152.1
The     judge   scheduled       to      preside over    his     trial   was     removed from    the     case    today
                schedule.01
B-ARG1  I-ARG1  B-V     B-ARG2  I-ARG2  I-ARG2  I-ARG2  I-ARG2  0       0       0       0       0       0       0
```

- The first row is a unique identifier (1st annotation of the 152nd sentence in the file ontonotes/bc/cnn/00/cnn_0000).
- The second row contains the tokens of the sentence (tab-separated).
- The third row contains the probank frame name for the predicate (empty field for all other tokens).
- The fourth row contains the B-I-O tag for each token.

The file `rolelist.txt` contains a list of propbank BIO labels in the dataset (i.e. possible output tokens). This list has been filtered to contain only roles that appeared more than 1000 times in the training data. We will load this list and create mappings from numeric ids to BIO tags and back.

```
role_to_id = {}
with open("role_list.txt",'r') as f:
    role_list = [x.strip() for x in f.readlines()]
```

```
role_to_id = dict((role, index) for (index, role) in enumerate(role_list))
role_to_id['[PAD]'] = -100

id_to_role = dict((index, role) for (role, index) in role_to_id.items())
```

Note that we are also mapping the '[PAD]' token to the value -100. This allows the loss function to ignore these tokens during training.

Double-click (or enter) to edit

Part 1 - Data Preparation

Before you can build the SRL model, you first need to preprocess the data.

1.1 - Tokenization

One challenge is that the pre-trained BERT model uses subword ("WordPiece") tokenization, but the Ontonotes data does not. Fortunately Huggingface transformers provides a tokenizer.

```
from transformers import BertTokenizerFast
tokenizer = BertTokenizerFast.from_pretrained('bert-base-uncased', do_lower_case=True)
tokenizer.tokenize("This is an unbelievably boring test sentence.")
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
tokenizer_config.json: 100%  48.0/48.0 [00:00<00:00, 5.08kB/s]
vocab.txt: 100%  232k/232k [00:00<00:00, 552kB/s]
tokenizer.json: 100%  466k/466k [00:00<00:00, 1.30MB/s]
config.json: 100%  570/570 [00:00<00:00, 79.3kB/s]
['this',
 'is',
 'an',
 'un',
 '##bel',
 '##ie',
 '##va',
 '##bly',
 'boring',
 'test',
 'sentence',
 '.']
```

TODO: We need to be able to maintain the correct labels (B-I-O tags) for each of the subwords. Complete the following function that takes a list of tokens and a list of B-I-O labels of the same length as parameters, and returns a new token / label pair, as illustrated in the following example.

```
>>> tokenize_with_labels("the fancyful penguin devoured yummy fish .".split(), "B-ARG0 I-ARG0 I-ARG0 B-V B-ARG1 I-ARG1
(['the',
  'fancy',
  '##ful',
  'penguin',
  'dev',
  '##oured',
  'yu',
  '##mmy',
  'fish',
  '.'],
 ['B-ARG0',
  'I-ARG0',
  'I-ARG0',
  'I-ARG0',
```

```
    'B-V',
    'I-V',
    'B-ARG1',
    'I-ARG1',
    'I-ARG1',
    'O'])
```

To approach this problem, iterate through each word/label pair in the sentence. Call the tokenizer on the word. This may result in one or more tokens. Create the correct number of labels to match the number of tokens. Take care to not generate multiple B- tokens.

This approach is a bit slower than tokenizing the entire sentence, but is necessary to produce proper input tokenization for the pre-trained BERT model, and the matching target labels.

```
def tokenize_with_labels(sentence, text_labels, tokenizer):
    """
    Word piece tokenization makes it difficult to match word labels
    back up with individual word pieces.
    """

    tokenized_sentence = []
    labels = []

    for word, label in zip(sentence, text_labels):

      word_tokens = tokenizer.tokenize(word)

      if not word_tokens:
        continue
      tokenized_sentence.append(word_tokens[0])

      if label.startswith("B-") and len(word_tokens)>1: ## built-in function
        labels.append(label)
        new_label = "I-" + label[1:]
        labels.extend([new_label] * (len(word_tokens)-1))
      elif label.startswith("I-") and len(word_tokens)>1:
        labels.extend([label] * len(word_tokens))
      else:
        labels.extend([label] * len(word_tokens))

      if len(word_tokens) > 1:
        tokenized_sentence.extend(word_tokens[1:])


    return tokenized_sentence, labels

tokenize_with_labels("the fancyful penguin devoured yummy fish .".split(), "B-ARG0 I-ARG0 I-ARG0 B-V B-ARG1 I-ARG1 O".split(), t
```

```
([the',
  'fancy',
  '##ful',
  'penguin',
  'dev',
  '##oured',
  'yu',
  '##mmy',
  'fish',
  '.'],
 ['B-ARG0',
  'I-ARG0',
  'I-ARG0',
  'I-ARG0',
  'B-V',
  'I--V',
  'B-ARG1',
  'I--ARG1',
  'I-ARG1',
  'O'])
```

1.2 Loading the Dataset

Next, we are creating a PyTorch Dataset class. This class acts as a contained for the training, development, and testing data in memory. You should already be familiar with Datasets and Dataloaders from homework 3.

1.2.1 **TODO**: Write the __init__(self, filename) method that reads in the data from a data file (specified by the filename).

For each annotation you start with the tokens in the sentence, and the BIO tags. Then you need to create the following

1. call the `tokenize_with_labels` function to tokenize the sentence.
2. Add the (token, label) pair to the self.items list.

1.2.2 **TODO**: Write the __len__(self) method that returns the total number of items.

1.2.3 **TODO**: Write the __getitem__(self, k) method that returns a single item in a format BERT will understand.

- We need to process the sentence by adding "[CLS]" as the first token and "[SEP]" as the last token. The need to pad the token sequence to 128 tokens using the "[PAD]" symbol. This needs to happen both for the inputs (sentence token sequence) and outputs (BIO tag sequence).
- We need to create an *attention mask*, which is a sequence of 128 tokens indicating the actual input symbols (as a 1) and [PAD] symbols (as a 0).
- We need to create a *predicate indicator* mask, which is a sequence of 128 tokens with at most one 1, in the position of the "B-V" tag. All other entries should be 0. The model will use this information to understand where the predicate is located.
- Finally, we need to convert the token and tag sequence into numeric indices. For the tokens, this can be done using the `tokenizer.convert_tokens_to_ids` method. For the tags, use the `role_to_id` dictionary. Each sequence must be a pytorch tensor of shape (1,128). You can convert a list of integer values like this `torch.tensor(token_ids, dtype=torch.long)`.

To keep everything organized, we will return a dictionary in the following format

```
 {'ids': token_tensor,
  'targets': tag_tensor,
  'mask': attention_mask_tensor,
  'pred': predicate_indicator_tensor}
```

(Hint: To debug these, read in the first annotation only / the first few annotations)

```
from torch.utils.data import Dataset, DataLoader

class SrlData(Dataset):

    def __init__(self, filename):

        super(SrlData, self).__init__()

        self.max_len = 128 # the max number of tokens inputted to the transformer.

        self.tokenizer = BertTokenizerFast.from_pretrained('bert-base-uncased', do_lower_case=True)

        self.items = []
        self.role_to_id = role_to_id
        self.id_to_role = id_to_role
        # complete this method
        tokens = []
        labels = []
        with open(filename, 'r') as file:
          for i, line in enumerate(file):
            line = line.strip()
            if (i+1) % 2 == 0 and (i+1) % 4 != 0:
              tokens = line.split('\t')
            elif (i+1) % 4 == 0:
              labels = line.split('\t')
              tokenized_sentence, tokenized_labels = tokenize_with_labels(tokens, labels, self.tokenizer)
              self.items.append((tokenized_sentence, tokenized_labels))


    def __len__(self):
        return len(self.items) # replace

    def __getitem__(self, k):


        #complete this method
        tokenized_sentence, tokenized_labels = self.items[k]
        # Pad or truncate the sequences to max_len
        padded_tokens = tokenized_sentence[:self.max_len-2]
        padded_labels = tokenized_labels[:self.max_len-2]
        padded_tokens = ['[CLS]'] + padded_tokens + ['[SEP]'] + ['[PAD]'] * (self.max_len - 2 - len(padded_tokens))
```

```
        padded_labels = ['[CLS]'] + padded_labels + ['[SEP]'] + ['[PAD]'] * (self.max_len - 2 - len(padded_labels))

        # Convert tokens and labels to IDs
        token_ids = self.tokenizer.convert_tokens_to_ids(padded_tokens)
        label_ids = [self.role_to_id.get(label, self.role_to_id['O']) for label in padded_labels]

        # Create attention mask
        attention_mask = [1 if token != '[PAD]' else 0 for token in padded_tokens]

        # Create predicate indicator
        predicate_indicator = [1 if label == 'B-V' else 0 for label in padded_labels]

        # Convert all to PyTorch tensors
        token_tensor = torch.tensor(token_ids, dtype=torch.long)
        label_tensor = torch.tensor(label_ids, dtype=torch.long)
        attn_mask = torch.tensor(attention_mask, dtype=torch.long)
        pred_tensor = torch.tensor(predicate_indicator, dtype=torch.long)

        return {'ids': token_tensor,#token_tensor,
                'mask': attn_mask, #attn_mask,
                'targets': label_tensor, #label_tensor,
                'pred': pred_tensor#pred_tensor
               }


# Reading the training data takes a while for the entire data because we preprocess all data offline
data = SrlData("propbank_train.tsv")
```

2. Model Definition

```
from torch.nn import Module, Linear, CrossEntropyLoss
from transformers import BertModel
```

We will define the pyTorch model as a subclass of the torch.nn.Module class. The code for the model is provided for you. It may help to take a look at the documentation to remind you of how Module works. Take a look at how the huggingface BERT model simply becomes another sub-module.

```
class SrlModel(Module):

    def __init__(self):

        super(SrlModel, self).__init__()

        self.encoder = BertModel.from_pretrained("bert-base-uncased")

        # The following two lines would freeze the BERT parameters and allow us to train the classifier by itself.
        # We are fine-tuning the model, so you can leave this commented out!
        # for param in self.encoder.parameters():
        #    param.requires_grad = False

        # The linear classifier head, see model figure in the introduction.
        self.classifier = Linear(768, len(role_to_id))


    def forward(self, input_ids, attn_mask, pred_indicator):

        # This defines the flow of data through the model

        # Note the use of the "token type ids" which represents the segment encoding explained in the introduction.
        # In our segment encoding, 1 indicates the predicate, and 0 indicates everything else.
        bert_output =  self.encoder(input_ids=input_ids, attention_mask=attn_mask, token_type_ids=pred_indicator)

        enc_tokens = bert_output[0] # the result of encoding the input with BERT
        logits = self.classifier(enc_tokens) #feed into the classification layer to produce scores for each tag.

        # Note that we are only interested in the argmax for each token, so we do not have to normalize
        # to a probability distribution using softmax. The CrossEntropyLoss loss function takes this into account.
        # It essentially computes the softmax first and then computes the negative log-likelihood for the target classes.
        return logits
```

```
model = SrlModel().to('cuda') # create new model and store weights in GPU memory
```

```
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falling back to regular HTTP download. For
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed. Falli
model.safetensors: 100%  440M/440M [00:00<00:00, 602MB/s]
```

Now we are ready to try running the model with just a single input example to check if it is working correctly. Clearly it has not been trained, so the output is not what we expect. But we can see what the loss looks like for an initial sanity check.

TODO:

- Take a single data item from the dev set, as provided by your Dataset class defined above. Obtain the input token ids, attention mask, predicate indicator mask, and target labels.
- Run the model on the ids, attention mask, and predicate mask like this:

```
# pick an item from the dataset. Then run
device = 'cuda'
dev_data = SrlData("propbank_dev.tsv")
ids = dev_data[0]['ids'].to(device,dtype=torch.long).unsqueeze(0)
mask = dev_data[0]['mask'].to(device,dtype=torch.long).unsqueeze(0)
pred = dev_data[0]['pred'].to(device,dtype=torch.long).unsqueeze(0)
targets = dev_data[0]['targets'].to(device,dtype=torch.long).unsqueeze(0)

outputs = model(ids, mask, pred)
outputs
outputs.size()
```

```
torch.Size([1, 128, 53])
```

TODO: Compute the loss on this one item only. The initial loss should be close to -ln(1/num_labels)

Without training we would assume that all labels for each token (including the target label) are equally likely, so the negative log probability for the targets should be approximately

$$-\ln(\frac{1}{\text{num\_labels}}).$$

This is what the loss function should return on a single example. This is a good sanity check to run for any multi-class prediction problem.

```
import math
-math.log(1 / len(role_to_id), math.e)
```

```
3.970291913552122
```

```
loss_function = CrossEntropyLoss(ignore_index = -100, reduction='mean')

# complete this. Note that you still have to provide a (batch_size, input_pos)
# tensor for each parameter, where batch_size =1

# outputs = model(ids, mask, pred)
# loss = loss_function(...)
# loss.item()   #this should be approximately the score from the previous cell
model.eval()
outputs = model(ids, mask, pred)
loss = loss_function(outputs.transpose(2, 1),targets)
loss.item()   #this should be approximately the score from the previous cell
predictions = torch.argmax(outputs, dim=2)
predictions
```

```
tensor([[ 2, 13, 52, 35, 35, 27, 15, 32, 36, 52, 32, 32, 17, 32, 36, 45, 49, 32,
         23, 35, 18, 18, 35, 35, 35, 35, 23, 15, 44, 40,  8, 40,  8, 32,  8, 40,
          8,  8, 22, 32, 11, 11, 11, 11,  8,  8,  8,  8, 32, 36, 32, 11, 11, 11,
          8, 11, 19,  8, 35, 11, 11, 11, 36, 40,  8, 36, 32, 11, 11,  9, 47, 36,
          8,  8,  8, 32, 11, 11, 11, 11, 11,  8,  8, 47,  8,  8,  8,  8,  7, 11,
         11, 36, 40, 22, 32, 32,  9, 11, 11,  9, 11, 11,  8,  8, 36, 32, 11, 11,
         11, 11,  8, 42, 15, 44, 40, 44,  8, 40, 40, 19, 40,  8, 40, 40, 32, 32,
          8, 32]], device='cuda:0')
```

TODO: At this point you should also obtain the actual predictions by taking the argmax over each position. The result should look something like this (values will differ).

```
tensor([[ 1,  4,  4,  4,  4,  4,  5, 29, 29, 29,  4, 28,  6, 32, 32, 32, 32, 32,
         32, 32, 30, 30, 32, 30, 32,  4, 32, 32, 30,  4, 49,  4, 49, 32, 30,  4,
         32,  4, 32, 32,  4,  2,  4,  4, 32,  4, 32, 32, 32, 32, 30, 32, 32, 30,
         32,  4,  4, 49,  4,  4,  4,  4,  4,  4,  4,  4,  4,  4,  6,  6, 32, 32,
         30, 32, 32, 32, 32, 32, 30, 30, 30, 32, 30, 49, 49, 32, 32, 30,  4,  4,
          4,  4, 29,  4,  4,  4,  4,  4,  4, 32,  4,  4,  4, 32,  4, 30,  4, 32,
         30,  4, 32,  4,  4,  4,  4,  4, 32,  4,  4,  4,  4,  4,  4,  4,  4,  4,
          4,  4]], device='cuda:0')
```

Then use the id_to_role dictionary to decode to actual tokens.

```
['[CLS]', 'O', 'O', 'O', 'O', 'O', 'B-ARG0', 'I-ARG0', 'I-ARG0', 'I-ARG0', 'O', 'B-V', 'B-ARG1', 'I-ARG2', 'I-ARG2', '
```

For now, just make sure you understand how to do this for a single example. Later, you will write a more formal function to do this once we have trained the model.

```
# act_tokens = [role_to_id.get(label, self.role_to_id['O']) for label in predictions]
pred_list = predictions.tolist()[0]
act_tokens = [tokens for pred_ids in pred_list for tokens, ids in role_to_id.items() if ids == pred_ids]
act_tokens
```

```
        'I-ARG2',
       'B-ARG2',
       'I-ARG2']
```

```
predictions.shape
#act_tokens = [role_to_id.get(label, self.role_to_id['O']) for label in predictions]
pred_list = predictions.tolist()[0]
act_tokens = [tokens for pred_ids in pred_list for tokens, ids in role_to_id.items() if ids == pred_ids]
act_tokens
```

```
    'I-ARGM-PRP',
    'I-ARGM-ADV',
    'B-ARG2',
    'B-ARG2',
    'B-ARG2',
    'I-ARG2',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARG2',
    'B-ARG2',
    'I-ARGM-PRP',
    'B-ARG2',
    'B-ARG2',
    'B-ARG2',
    'B-ARG2',
    'B-ARG1-DSP',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'I-ARGM-ADV',
    'I-ARGM-EXT',
    'B-ARGM-PRD',
    'I-ARG2',
    'I-ARG2',
    'B-ARG3',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARG3',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARG2',
    'B-ARG2',
    'I-ARGM-ADV',
    'I-ARG2',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARGM-ADJ',
    'B-ARG2',
    'I-ARGM-LOC',
    'B-ARGM-DIS',
    'I-ARGM-MOD',
    'I-ARGM-EXT',
    'I-ARGM-MOD',
    'B-ARG2',
    'I-ARGM-EXT',
    'I-ARGM-EXT',
    'B-ARGM-MNR',
    'I-ARGM-EXT',
    'B-ARG2',
    'I-ARGM-EXT',
    'I-ARGM-EXT',
    'I-ARG2',
    'I-ARG2',
    'B-ARG2',
    'I-ARG2']
```

3. Training loop

pytorch provides a DataLoader class that can be wrapped around a Dataset to easily use the dataset for training. The DataLoader allows us to easily adjust the batch size and shuffle the data.

```
from torch.utils.data import DataLoader
loader = DataLoader(data, batch_size = 32, shuffle = True)
```

The following cell contains the main training loop. The code should work as written and report the loss after each batch, cumulative average loss after each 100 batches, and print out the final average loss after the epoch.

TODO: Modify the training loop belowso that it also computes the accuracy for each batch and reports the average accuracy after the epoch. The accuracy is the number of correctly predicted token labels out of the number of total predictions. Make sure you exclude [PAD] tokens, i.e. tokens for which the target label is -100. It's okay to include [CLS] and [SEP] in the accuracy calculation.

```
loss_function = CrossEntropyLoss(ignore_index = -100, reduction='mean')

LEARNING_RATE = 1e-05
optimizer = torch.optim.AdamW(params=model.parameters(), lr=LEARNING_RATE)

device = 'cuda'

def train():
    """
    Train the model for one epoch.
    """
    tr_loss = 0
    nb_tr_examples, nb_tr_steps = 0, 0
    tr_preds, tr_labels = 0, 0
    # put model in training mode
    model.train()

    for idx, batch in enumerate(loader):

        # Get the encoded data for this batch and push it to the GPU
        ids = batch['ids'].to(device, dtype = torch.long)
        mask = batch['mask'].to(device, dtype = torch.long)
        targets = batch['targets'].to(device, dtype = torch.long)
        pred_mask = batch['pred'].to(device, dtype = torch.long)

        # Run the forward pass of the model
        logits = model(input_ids=ids, attn_mask=mask, pred_indicator=pred_mask)
        loss = loss_function(logits.transpose(2,1), targets)
        tr_loss += loss.item()
        print("Batch loss: ", loss.item()) # can comment out if too verbose.

        nb_tr_steps += 1
        nb_tr_examples += targets.size(0)

        if idx % 100==0:
            #torch.cuda.empty_cache() # can help if you run into memory issues
            curr_avg_loss = tr_loss/nb_tr_steps
            print(f"Current average loss: {curr_avg_loss}")

        # Compute accuracy for this batch
        mask = targets != -100
        matching = torch.sum(torch.argmax(logits,dim=2) == targets)
        predictions = torch.sum(mask)
        tr_preds += matching.item()
        tr_labels += predictions.item()

        # Run the backward pass to update parameters
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

    epoch_loss = tr_loss / nb_tr_steps
    print(f"Training loss epoch: {epoch_loss}")
    epoch_accuracy = tr_preds/tr_labels
    print(f"Training accuracy epoch: {epoch_accuracy}")
```

Now let's train the model for one epoch. This will take a while (up to a few hours).

```
train()
```

```
Batch loss:  0.1872660517692566
```

```
Batch loss:  0.18726605176925
Batch loss:  0.27660953998565674
Batch loss:  0.5148825645446777
Batch loss:  0.33062228560447693
Batch loss:  0.08504681289196014
Batch loss:  0.09550785273313522
Batch loss:  0.26235073804855347
Batch loss:  0.2453237920999527
Batch loss:  0.21078424155712128
Batch loss:  0.2539408206939697
Batch loss:  0.20544610917568207
Batch loss:  0.18833865225315094
Batch loss:  0.34785914421081543
Batch loss:  0.25071388483047485
Batch loss:  0.30084383487701416
Batch loss:  0.17232771217823029
Batch loss:  0.2206701785326004
Batch loss:  0.10365457832813263
Batch loss:  0.19513128697872162
Batch loss:  0.30219364166259766
Batch loss:  0.16839951276779175
Batch loss:  0.18506203591823578
Batch loss:  0.30679041147232056
Batch loss:  0.2732817828655243
Batch loss:  0.12508951127529144
Batch loss:  0.17021547257900238
Batch loss:  0.21608057618141174
Batch loss:  0.2757944166660309
Batch loss:  0.2826422452926636
Batch loss:  0.2137814462184906
Batch loss:  0.2765103578567505
Current average loss: 0.3781939327424916
Batch loss:  0.12736555933952332
Batch loss:  0.3131615221500397
Batch loss:  0.11787880957126617
Batch loss:  0.21946276724338531
Batch loss:  0.12062471359968185
Batch loss:  0.17384004592895508
Batch loss:  0.14561496675014496
Batch loss:  0.26347601413726807
Batch loss:  0.135414257645607
Batch loss:  0.42493775486946106
Batch loss:  0.4512430429458618
Batch loss:  0.10924618691205978
Batch loss:  0.2946852147579193
Batch loss:  0.16103704273700714
Batch loss:  0.4145125448703766
Batch loss:  0.1456981748342514
Batch loss:  0.19904030859470367
Training loss epoch: 0.3779278848783915
Training accuracy epoch: 0.8944697931130284
```

In my experiments, I found that two epochs are needed for good performance.

```
train()
```

```
Batch loss:  0.10320004090010986
```

```
Batch loss:  0.10350004990100860
Batch loss:  0.14916974306106567
Batch loss:  0.46469664573669434
Batch loss:  0.16225016117095947
Batch loss:  0.2793416976928711
Batch loss:  0.22681599855422974
Batch loss:  0.1714252233505249
Batch loss:  0.08115319162607193
Batch loss:  0.18460622429847717
Batch loss:  0.26494184136390686
Batch loss:  0.30225756764411926
Current average loss: 0.1905501493936126
Batch loss:  0.12720121443271637
Batch loss:  0.13102883100509644
Batch loss:  0.22392123937606812
Batch loss:  0.13447971642017365
Batch loss:  0.24064308404922485
Batch loss:  0.11189209669828415
Batch loss:  0.11062154173851013
Batch loss:  0.2952254116535187
Batch loss:  0.2738935351371765
Batch loss:  0.22720274329185486
Batch loss:  0.20320335030555725
Batch loss:  0.32136985659599304
Batch loss:  0.2478887438774109
Batch loss:  0.1335618793964386
Batch loss:  0.10079118609428406
Batch loss:  0.2068949192762375
Batch loss:  0.13203148543834686
Training loss epoch: 0.19054836677954612
Training accuracy epoch: 0.9409072068136571
```

I ended up with a training loss of about 0.19 and a training accuracy of 0.94. Specific values may differ.

At this point, it's a good idea to save the model (or rather the parameter dictionary) so you can continue evaluating the model without having to retrain.

```
torch.save(model.state_dict(), "srl_model_fulltrain_2epoch_finetune_1e-05.pt")
```

4. Decoding

```
# Optional step: If you stopped working after part 3, first load the trained model

model = SrlModel().to('cuda')
model.load_state_dict(torch.load("srl_model_fulltrain_2epoch_finetune_1e-05.pt"))
model = model.to('cuda')
```

TODO (this is the fun part): Now that we have a trained model, let's try labeling an unseen example sentence. Complete the functions decode_output and label_sentence below. decode_output takes the logits returned by the model, extracts the argmax to obtain the label predictions for each token, and then translate the result into a list of string labels.

label_sentence takes a list of input tokens and a predicate index, prepares the model input, call the model and then call decode_output to produce a final result.

Note that you have already implemented all components necessary (preparing the input data from the token list and predicate index, decoding the model output). But now you are putting it together in one convenient function.

```
tokens = "A U. N. team spent an hour inside the hospital , where it found evident signs of shelling and gunfire .".split()


def decode_output(logits): # it will be useful to have this in a separate function later on
    """
    Given the model output, return a list of string labels for each token.
    """
    predicted_indices = torch.argmax(logits, dim=2)
    pred_list = predicted_indices.tolist()[0]
    act_labels = [tokens for pred_ids in pred_list for tokens, ids in role_to_id.items() if ids == pred_ids]
    return act_labels



def label_sentence(tokens, pred_idx):

    # complete this function to prepare token_ids, attention mask, predicate mask, then call the model.
    # Decode the output to produce a list of labels.
```

```
    n = len(tokens)
    pad_tokens = ['[CLS]'] + tokens + ['[SEP]'] + ['[PAD]'] * (128 - 2 - n)
    ids = tokenizer.convert_tokens_to_ids(pad_tokens)
    ids = torch.tensor(ids, dtype=torch.long).unsqueeze(0)
    ids = ids.to(device,dtype=torch.long)
    mask = [1 if pad_tokens != '[PAD]' else 0 for token in pad_tokens]
    mask = torch.tensor(mask, dtype=torch.long).unsqueeze(0)
    mask = mask.to(device,dtype=torch.long)
    pred_mask = torch.zeros(len(pad_tokens))
    pred_mask[pred_idx+1] = 1
    pred_mask = torch.tensor(pred_mask, dtype=torch.long).unsqueeze(0)
    pred_mask = pred_mask.to(device,dtype=torch.long)
    logits = model(ids,mask,pred_mask)
    return decode_output(logits)[1:n+1]


# Now you should be able to run

label_test = label_sentence(tokens, 13) # Predicate is "found"
zip(tokens, label_test)
(tokens, label_test)
```

```
<ipython-input-26-62a4960459b6>:15: UserWarning: To copy construct from a tensor, it is recommended to use sourceTensor.clon
  pred_mask = torch.tensor(pred_mask, dtype=torch.long).unsqueeze(0)
(['A',
  'U.',
  'N.',
  'team',
  'spent',
  'an',
  'hour',
  'inside',
  'the',
  'hospital',
  ',',
  'where',
  'it',
  'found',
  'evident',
  'signs',
  'of',
  'shelling',
  'and',
  'gunfire',
  '.'],
 ['O',
  'O',
  'O',
  'O',
  'O',
  'O',
  'O',
  'O',
  'B-ARGM-LOC',
  'I-ARGM-LOC',
  'O',
  'B-ARGM-LOC',
  'B-ARG0',
  'B-V',
  'B-ARG1',
  'I-ARG1',
  'I-ARG1',
  'I-ARG1',
  'I-ARG1',
  'I-ARG1',
  'O'])
```

The expected output is somethign like this:

```
('A', 'O'),
('U.', 'O'),
('N.', 'O'),
('team', 'O'),
('spent', 'O'),
('an', 'O'),
('hour', 'O'),
```

```
  ('inside', 'O'),
  ('the', 'B-ARGM-LOC'),
  ('hospital', 'I-ARGM-LOC'),
  (',', 'O'),
  ('where', 'B-ARGM-LOC'),
  ('it', 'B-ARG0'),
  ('found', 'B-V'),
  ('evident', 'B-ARG1'),
  ('signs', 'I-ARG1'),
  ('of', 'I-ARG1'),
  ('shelling', 'I-ARG1'),
  ('and', 'I-ARG1'),
  ('gunfire', 'I-ARG1'),
  ('.', 'O'),
```

5. Evaluation 1: Token-Based Accuracy

We want to evaluate the model on the dev or test set.

```
dev_data = SrlData("propbank_dev.tsv") # Takes a while because we preprocess all data offline
```

```
from torch.utils.data import DataLoader
loader = DataLoader(dev_data, batch_size = 1, shuffle = False)
```

```
# Optional: Load the model again if you stopped working prior to this step.
# model = SrlModel()
# model.load_state_dict(torch.load("srl_model_fulltrain_2epoch_finetune_1e-05.pt"))
# model = mode.to('cuda')
```

TODO: Complete the evaluate_token_accuracy function below. The function should iterate through the items in the data loader (see training loop in part 3). Run the model on each sentence/predicate pair and extract the predictions.

For each sentence, count the correct predictions and the total predictions. Finally, compute the accuracy as #correct_predictions / #total_predictions

Careful: You need to filter out the padded positions ([PAD] target tokens), as well as [CLS] and [SEP]. It's okay to include [B-V] in the count though.

```
def evaluate_token_accuracy(model, loader):

    model.eval() # put model in evaluation mode

    # for the accuracy
    total_correct = 0 # number of correct token label predictions.
    total_predictions = 0 # number of total predictions = number of tokens in the data.

    # iterate over the data here.
    for idx, batch in enumerate(loader):
      ids = batch['ids'].to(device, dtype = torch.long)
      mask = batch['mask'].to(device, dtype = torch.long)
      targets = batch['targets'].to(device, dtype = torch.long)
      pred_mask = batch['pred'].to(device, dtype = torch.long)


      with torch.no_grad():  # no need to track gradients during evaluation
            outputs = model(input_ids=ids, attn_mask=mask, pred_indicator=pred_mask)  # get model predictions, adjust as necessa

        # Assume outputs are logits; we take the max index to get the predicted class
      predictions = torch.argmax(outputs, dim=-1)

        # Now we need to compare predictions to the ground truth labels and calculate accuracy
      mask = (targets != -100) & \
               (targets  != 1) & \
               (targets  != 2)  # mask out [PAD], [CLS], [SEP]

      correct_predictions = (predictions == targets ) & mask  # correct where pred equals label and not masked
      total_correct += correct_predictions.sum().item()  # count correct predictions
      total_predictions += mask.sum().item()  # count all non-masked tokens
```

```
    acc = total_correct / total_predictions
    print(f"Accuracy: {acc}")
    evaluate_token_accuracy(model,loader)
```

6. Span-Based evaluation

While the accuracy score in part 5 is encouraging, an accuracy-based evaluation is problematic for two reasons. First, most of the target labels are actually O. Second, it only tells us that per-token prediction works, but does not directly evaluate the SRL performance.

Instead, SRL systems are typically evaluated on micro-averaged precision, recall, and F1-score for predicting labeled spans.

More specifically, for each sentence/predicate input, we run the model, decode the output, and extract a set of labeled spans (from the output and the target labels). These spans are (i,j,label) tuples.

We then compute the true_positives, false_positives, and false_negatives based on these spans.

In the end, we can compute

- Precision: true_positive / (true_positives + false_positives) , that is the number of correct spans out of all predicted spans.
- Recall: true_positives / (true_positives + false_negatives) , that is the number of correct spans out of all target spans.
- F1-score: (2 * precision * recall) / (precision + recall)

For example, consider

	[CLS]	The	judge	scheduled	to	preside	over	his	trial	was	removed	from	the	case	today	.
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
target	[CLS]	B-ARG1	I-ARG1	B-V	B-ARG2	I-ARG2	I-ARG2	I-ARG2	I-ARG2	O	O	O	O	O	O	O
prediction	[CLS]	B-ARG1	I-ARG1	B-V	I-ARG2	I-ARG2	O	O	O	O	O	O	O	O	B-ARGM-TMP	O

The target spans are (1,2,"ARG1"), and (4,8,"ARG2").

The predicted spans would be (1,2,"ARG1"), (14,14,"ARGM-TMP"). Note that in the prediction, there is no proper ARG2 span because we are missing the B-ARG2 token, so this span should not be created.

So for this sentence we woudl get: true_positives: 1 false_positives: 1 false_negatives: 1

TODO: Complete the function evaluate_spans that performs the span-based evaluation on the given model and data loader. You can use the provided extract_spans function, which returns the spans as a dictionary. For example {(1,2): "ARG1", (4,8):"ARG2"}

```
def extract_spans(labels):
    spans = {} # map (start,end) ids to label
    current_span_start = 0
    current_span_type = ""
    inside = False
    for i, label in enumerate(labels):
        if label.startswith("B"):
            if inside:
                if current_span_type != "V":
                    spans[(current_span_start,i)] = current_span_type
            current_span_start = i
            current_span_type = label[2:]
            inside = True
        elif inside and label.startswith("O"):
            if current_span_type != "V":
                spans[(current_span_start,i)] = current_span_type
            inside = False
        elif inside and label.startswith("I") and label[2:] != current_span_type:
            if current_span_type != "V":
                spans[(current_span_start,i)] = current_span_type
            inside = False
    return spans


def evaluate_spans(model, loader):


    total_tp = 0
    total_fp = 0
    total_fn = 0

    for idx, batch in enumerate(loader):
```

```
            # compelte this

            ids = batch['ids'].to(device, dtype = torch.long)
            mask = batch['mask'].to(device, dtype = torch.long)
            targets = batch['targets'].to(device, dtype = torch.long)
            pred_mask = batch['pred'].to(device, dtype = torch.long)

            model.eval()
            with torch.no_grad():
                outputs = model(input_ids=ids, attn_mask=mask, pred_indicator=pred_mask)
            predictions = decode_output(outputs)
            targets_list = targets.tolist()[0]
            act_targets = [tokens for target in targets_list for tokens, ids in role_to_id.items() if ids == target]

            true_spans = extract_spans(act_targets)
            pred_spans = extract_spans(predictions[0:mask.sum()])

            # Compute true positives, false positives, and false negatives
            tp = len(set(true_spans.items()) & set(pred_spans.items()))
            fp = len(pred_spans) - tp
            fn = len(true_spans) - tp

            total_tp += tp
            total_fp += fp
            total_fn += fn



        total_p = total_tp / (total_tp + total_fp)
        total_r = total_tp / (total_tp + total_fn)
        total_f = (2 * total_p *total_r) / (total_p + total_r)

        print(f"Overall P: {total_p}  Overall R: {total_r}  Overall F1: {total_f}")

    evaluate_spans(model, loader)
```

```
Overall P: 0.8212997602992899  Overall R: 0.8246434837374601  Overall F1: 0.8229682256417165
```

In my evaluation, I got an F score of 0.82 (which slightly below the state-of-the art in 2018)

OPTIONAL:

Repeat the span-based evaluation, but print out precision/recall/f1-score for each role separately.