Part A: Parts of Speech Tagging using Hidden Markov Model and Viterbi Algorithm on Hindi Dataset (Total: 40 Points out of 100)

For this assignment, we will implement the Viterbi Decoder using the Forward Algorithm of Hidden Markov Model as explained in class.

Then, we will create an HMM-based PoS Tagger for Hindi language using the annotated Tagset in nltk.indian

You need to first implement the missing code in hmm.py, then run the cells here to get the points

In [1]:
```
from tqdm.autonotebook import tqdm
```

```
/var/folders/31/yzh9j02x7bxd463cl1x0_21h0000gn/T/ipykernel_96500/987820437.py:1: TqdmExperimentalWarning: Using
`tqdm.autonotebook.tqdm` in notebook mode. Use `tqdm.tqdm` instead to force console mode (e.g. in jupyter conso
le)
  from tqdm.autonotebook import tqdm
```

In [2]:
```
# This is so that you don't have to restart the kernel everytime you edit hmm.py

%load_ext autoreload
%autoreload 2
```

In [3]:
```
from hmm import *
```

```
[nltk_data] Downloading package indian to
[nltk_data]     /Users/xingyuchen/nltk_data...
[nltk_data]   Package indian is already up-to-date!
```

1st-Order Hidden Markov Model Class:

The hidden markov model class would have the following attributes:

```
1. initial state log-probs vector (pi)
2. state transition log-prob matrix (A)
3. observation log-prob matrix (B)
```

The following methods:

```
1. fit method to count the probabilitis of the training set
2. path probability
3. viterbi decoding algorithm
```

function VITERBI(*observations* of len T,*state-graph* of len N) **returns** *best-path*, *path-prob*

create a path probability matrix *viterbi[N,T]*
for each state s **from** 1 **to** N **do** ; initialization step
  $viterbi[s,1] \leftarrow \pi_s * b_s(o_1)$
  $backpointer[s,1] \leftarrow 0$
for each time step t **from** 2 **to** T **do** ; recursion step
  for each state s **from** 1 **to** N **do**
    $viterbi[s,t] \leftarrow \max_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
    $backpointer[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
$bestpathprob \leftarrow \max_{s=1}^{N} viterbi[s,T]$; termination step
$bestpathpointer \leftarrow \operatorname{argmax}_{s=1}^{N} viterbi[s,T]$; termination step
bestpath ← the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

Figure A.9 Viterbi algorithm for finding optimal sequence of hidden states. Given an observation sequence and an HMM $\lambda = (A,B)$, the algorithm returns the state path through the HMM that assigns maximum likelihood to the observation sequence.

Task 1: Testing the HMM (20 Points)

In [4]:
```
### DO NOT EDIT ###

# 5 points for the fit test case
# 15 points for the decode test case

# run the funtion that tests the HMM with synthetic parameters!
run_tests()
```

```
Testing the fit function of the HMM
All Test Cases Passed!
Testing the decode function of the HMM
All Test Cases Passed!
Yay! You have a working HMM. Now try creating a pos-tagger using this class.
```

Task 2: PoS Tagging on Hindi Tagset (20 Points)

For this assignment, we will use the Hindi Tagged Dataset available with nltk.indian

Helper methods to load the dataset is provided in hmm.py

Please go through the functions and explore the dataset

Report the Accuracy for the Dev and Test Sets. You should get something between 65-85%

In [5]:
```
words, tags, observation_dict, state_dict, all_observation_ids, all_state_ids = get_hindi_dataset()

# we need to add the id for unknown word (<unk>) in our observations vocab
UNK_TOKEN = '<unk>'

observation_dict[UNK_TOKEN] = len(observation_dict)
print("id of the <unk> token:", observation_dict[UNK_TOKEN])
```

```
id of the <unk> token: 2186
```

In [6]:
```
print("No. of unique words in the corpus:", len(observation_dict))
print("No. of tags in the corpus", len(state_dict))
```

```
No. of unique words in the corpus: 2187
No. of tags in the corpus 26
```

In [7]:
```
# Split the dataset into train, validation and development sets

import random
random.seed(42)
from sklearn.model_selection import train_test_split

data_indices = list(range(len(all_observation_ids)))

train_indices, dev_indices = train_test_split(data_indices, test_size=0.2, random_state=42)

dev_indices, test_indices = train_test_split(dev_indices, test_size=0.5, random_state=42)

print(len(train_indices), len(dev_indices), len(test_indices))


def get_state_obs(state_ids, obs_ids, indices):
    return [state_ids[i] for i in indices], [obs_ids[i] for i in indices]


train_state_ids, train_observation_ids = get_state_obs(all_state_ids, all_observation_ids, train_indices)
dev_state_ids, dev_observation_ids = get_state_obs(all_state_ids, all_observation_ids, dev_indices)
test_state_ids, test_observation_ids = get_state_obs(all_state_ids, all_observation_ids, test_indices)
```

```
432 54 54
```

In [8]:
```
def add_unk_id(observation_ids, unk_id, ratio=0.05):
    """
    make 1% of observations unknown
    """
    for obs in observation_ids:
        for i in range(len(obs)):
            if random.random() < ratio:
                obs[i] = unk_id

add_unk_id(train_observation_ids, observation_dict[UNK_TOKEN])
add_unk_id(dev_observation_ids, observation_dict[UNK_TOKEN])
add_unk_id(test_observation_ids, observation_dict[UNK_TOKEN])
```

In [9]:
```
pos_tagger = HMM(len(state_dict), len(observation_dict))
pos_tagger.fit(train_state_ids, train_observation_ids)
```

In [10]:
```
assert np.round(np.exp(pos_tagger.pi).sum()) == 1
assert np.round(np.exp(pos_tagger.A).sum()) == len(state_dict)
assert np.round(np.exp(pos_tagger.B).sum()) == len(state_dict)

print('All Test Cases Passed!')
```

```
All Test Cases Passed!
```

In [11]:
```
def accuracy(my_pos_tagger, observation_ids, true_labels):
    tag_predictions = my_pos_tagger.decode(observation_ids)
    tag_predictions = np.array([t for ts in tag_predictions for t in ts])
    true_labels_flat = np.array([t for ts in true_labels for t in ts])
    acc = np.sum(tag_predictions == true_labels_flat)/len(tag_predictions)
    return acc
```

In [12]:
```
print('dev accuracy:', accuracy(pos_tagger, dev_observation_ids, dev_state_ids))
```

```
dev accuracy: 0.8127659574468085
```

In [13]:
```
print('test accuracy:', accuracy(pos_tagger, test_observation_ids, test_state_ids))
```

```
test accuracy: 0.7987012987012987
```

In [14]:
```
# Fit a pos_tagger on the entire dataset.
import pickle

full_state_ids = train_state_ids + dev_state_ids + test_state_ids
full_observation_ids = train_observation_ids + dev_observation_ids + test_state_ids

hindi_pos_tagger = HMM(len(state_dict), len(observation_dict))
hindi_pos_tagger.fit(full_state_ids, full_observation_ids)

pickle.dump(hindi_pos_tagger, open('hindi_pos_tagger.pkl', 'wb'))
```

In [15]:
```
### Finally we will use the hindi_pos_tagger as a pre-processing step for our NER tagger
```

In [ ]:

Task B: Named Entity Recognition with CRF on Hindi Dataset. (Total: 60 Points out of 100)

In this part, you will use a CRF to implement a named entity recognition tagger. We have implemented a CRF for you in crf.py along with some functions to build, and pad feature vectors. Your job is to add more features to learn a better tagger. Then you need to complete the traiing loop implementation.

Finally, you can checkout the code in `crf.py` -- reflect on CRFs and span tagging, and answer the discussion questions.

We will use the Hindi NER dataset at: https://github.com/cfiltnlp/HiNER

The first step would be to download the repo into your current folder of the Notebook

In [1]:
```
#!git clone https://github.com/cfiltnlp/HiNER.git
```

In [2]:
```
import torch
```

In [3]:
```
# This is so that you don't have to restart the kernel everytime you edit hmm.py

%load_ext autoreload
%autoreload 2
```

First we load the data and labels. Feel free to explore them below.

Since we have provided a seperate train and dev split, there is not need to split the data yourself.

In [4]:
```
from crf import load_data, make_labels2i

train_filepath = "./HiNER/data/collapsed/train.conll"
dev_filepath = "./HiNER/data/collapsed/validation.conll"
labels_filepath = "./HiNER/data/collapsed/label_list"

train_sents, train_tag_sents = load_data(train_filepath)
dev_sents, dev_tag_sents = load_data(dev_filepath)
labels2i = make_labels2i(labels_filepath)

print("train sample", train_sents[2], train_tag_sents[2])
print()
print("labels2i", labels2i)
```

```
train sample ['रामनगर', 'इगलास', ',', 'अलीगढ़', ',', 'उत्तर', 'प्रदेश', 'स्थित', 'एक', 'गाँव', 'है।'] ['B-LOCATION', 'B-LOCA
TION', 'O', 'B-LOCATION', 'O', 'B-LOCATION', 'I-LOCATION', 'O', 'O', 'O', 'O']

labels2i {'<PAD>': 0, 'B-LOCATION': 1, 'B-ORGANIZATION': 2, 'B-PERSON': 3, 'I-LOCATION': 4, 'I-ORGANIZATION':
5, 'I-PERSON': 6, 'O': 7}
```

Feature engineering. (Total 30 points)

Notice that we are **learning** features to some extent: we start with one unique feature for every possible word. You can refer to figure 8.15 in the textbook for some good baseline features to try.

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Figure 8.15 Typical features for a feature-based NER system.

There is no need to worry about embeddings now.

Hindi POS Tagger (10 Points)

Although this step is not entirely necessary, if you want to use the HMM pos tagger to extract feature corresponding to the pos of the word in the sentence, we need to add this into the pipeline.

You get 10 points if you use your pos_tagger to featurize the sentences

In [5]:
```
from hmm import get_hindi_dataset
import pickle
from typing import List

words, tags, observation_dict, state_dict, all_observation_ids, all_state_ids = get_hindi_dataset()

# we need to add the id for unknown word (<unk>) in our observations vocab
UNK_TOKEN = '<unk>'

observation_dict[UNK_TOKEN] = len(observation_dict)
print("id of the <unk> token:", observation_dict[UNK_TOKEN])

## load the pos tagger
pos_tagger = pickle.load(open('hindi_pos_tagger.pkl', 'rb'))

def encode(sentences: List[List[str]]) -> List[List[int]]:
    """
    Using the observation_dict, convert the tokens to ids
    unknown words take the id for UNK_TOKEN
    """
    return [
        [observation_dict[t] if t in observation_dict else observation_dict[UNK_TOKEN]
            for t in sentence]
        for sentence in sentences]

def get_pos(pos_tagger, sentences) -> List[List[str]]:
    """
    The the pos tag for input sentences
    """
    sentence_ids = encode(sentences)
    decoded_pos_ids = pos_tagger.decode(sentence_ids)
    return [
        [tags[int(i)] for i in d_ids]
        for d_ids in decoded_pos_ids
    ]
```

```
[nltk_data] Downloading package indian to
[nltk_data]     /Users/xingyuchen/nltk_data...
[nltk_data]   Package indian is already up-to-date!
id of the <unk> token: 2186
```

Feature Engineering Functions (20 Points)

In [48]:
```
from typing import List
import numpy as np
# TODO: Update this function to add more features
#       You can check crf.py for how they are encoded, if interested.

with open('gazetteer_hindi.txt', 'r', encoding= 'utf-16') as f:
    gazetteer = f.read()
with open('hindi_suffixes.txt', 'r', encoding= 'utf-8') as f:
    suffixes = f.read()

def make_features(text: List[str]) -> List[List[int]]:
    """Turn a text into a feature vector.

    Args:
        text (List[str]): List of tokens.

    Returns:
        List[List[int]]: List of feature Lists.
    """
    feature_lists = []
    for i, token in enumerate(text):
        feats = []
        # We add a feature for each unigram.
        feats.append(f"word={token}")
        # TODO: Add more features here
        feats.append(f"pos={get_pos(pos_tagger, token)[0]}")
        if len(text) == 1:
            feats.append(f"prev_word={'<S>'}")
            feats.append(f"next_word={'<E>'}")
        else:
            if i == 0:
                feats.append(f"prev_word={'<S>'}")
                feats.append(f"next_word={text[i+1]}")
#                 feats.append(f"prev_pos={'<S>'}")
#                 feats.append(f"next_pos={get_pos(pos_tagger, text[i+1])[0]}")
                prev_word = token
            elif i == len(text) - 1:
                feats.append(f"prev_word={prev_word}")
#                 feats.append(f"prev_pos={get_pos(pos_tagger, text[i-1])[0]}")
                feats.append(f"next_word={'<E>'}")
#                 feats.append(f"next_pos={'<E>'}")
            else:
                feats.append(f"prev_word={prev_word}")
#                 feats.append(f"prev_pos={get_pos(pos_tagger, text[i-1])[0]}")
                feats.append(f"next_word={text[i+1]}")
#                 feats.append(f"next_pos={get_pos(pos_tagger, text[i+1])[0]}")
                prev_word = token
        if token in gazetteer:
            feats.append(f"gazetteer={'1'}")
        else:
            feats.append(f"gazetteer={'0'}")
        if list(filter(token.endswith, suffixes)) != []:
            feats.append(f"suffixes={set(list(filter(token.endswith, suffixes)))}")
        else:
            feats.append(f"suffixes={set(['NaN'])}")
        # We append each feature to a List for the token.
#         special_characters = "!\"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"
#         if any(c in special_characters for c in token):
#             feats.append(f"specialsymbol={token}")
#         else:
#             feats.append(f"specialsymbol={'NaN'}")
#         print(feats)
        feature_lists.append(feats)
    return feature_lists
```

In [20]:
```
def featurize(sents: List[List[str]]) -> List[List[List[str]]]:
    """Turn the sentences into feature Lists.

    Eg.: For an input of 1 sentence:
         [[['I','am','a','student','at','CU','Boulder']]]
        Return list of features for every token for every sentence like:
        [[
         ['word=I',  'prev_word=<S>','pos=PRON',...],
         ['word=an', 'prev_word=I'  , 'pos=VB' ,...],
         [...]
        ]]

    Args:
        sents (List[List[str]]): A List of sentences, which are Lists of tokens.

    Returns:
        List[List[List[str]]]: A List of sentences, which are Lists of feature Lists
    """
    feats = []
    for sent in tqdm(sents):
        # Gets a List of Lists of feature strings
        feats.append(make_features(sent))
    return feats
```

Finish the training loop. (10 Points)

See the previous homework, and fill in the missing parts of the training loop.

In [51]:
```
from crf import f1_score, predict, PAD_SYMBOL, pad_features, pad_labels
import numpy as np
from tqdm.autonotebook import tqdm
import random
# TODO: Implement the training loop
# HINT: Build upon what we gave you for HW2.
# See cell below for how we call this training loop.

def training_loop(
    num_epochs,
    batch_size,
    train_features,
    train_labels,
    dev_features,
    dev_labels,
    optimizer,
    model,
    labels2i,
    pad_feature_idx
):
    samples = list(zip(train_features, train_labels))
    random.shuffle(samples)
    batches = []
    for i in range(0, len(samples), batch_size):
        batches.append(samples[i:i+batch_size])
    print("Training...")
    for i in range(num_epochs):
        losses = []
        for batch in tqdm(batches):
            # Here we get the features and labels, pad them,
            # and build a mask so that our model ignores PADs
            # We have abstracted the padding from you for simplicity,
            # but please reach out if you'd like learn more.
            features, labels = zip(*batch)
            features = pad_features(features, pad_feature_idx)
            features = torch.stack(features)
            # Pad the label sequences to all be the same size, so we
            # can form a proper matrix.
            labels = pad_labels(labels, labels2i[PAD_SYMBOL])
            labels = torch.stack(labels)
            mask = (labels != labels2i[PAD_SYMBOL])
            # TODO: Empty the dynamic computation graph
            optimizer.zero_grad()
            # TODO: Run the model. Since we use the pytorch-crf model,
            # our forward function returns the positive log-likelihood already.
            # We want the negative log-likelihood. See crf.py forward method in NERTagger
            loss = model.forward(features, labels, mask)
            # TODO: Backpropogate the loss through our model
            loss.backward()
            # TODO: Update our coefficients in the direction of the gradient.
            optimizer.step()
            # TODO: Store the losses for logging
            losses.append(loss.item())
        # TODO: Log the average Loss for the epoch
        print(f"epoch {i}, loss: {np.log(sum(losses)/len(losses))}")
        # TODO: make dev predictions with the `predict()` function
        dev_pred = predict(model, dev_features)
#         print('pred',dev_pred)
#         print('label', dev_labels)
        # TODO: Compute F1 score on the dev set and log it.
        print(f"Dev F1 log {np.log(f1_score(dev_pred, dev_labels, labels2i['O']))}")

    # Return the trained model
    return model
```

Run the training loop (10 Points)

We have provided the code here, but you can try different hyperparameters and test multiple runs.

In [49]:
```
from crf import build_features_set
from crf import make_features_dict
from crf import encode_features, encode_labels
from crf import NERTagger

# Build the model and featurized data
train_features = featurize(train_sents)
dev_features = featurize(dev_sents)

# Get the full inventory of possible features
all_features = build_features_set(train_features)
# Hash all features to a unique int.
features_dict = make_features_dict(all_features)
# Initialize the model.
model = NERTagger(len(features_dict), len(labels2i))

encoded_train_features = encode_features(train_features, features_dict)
encoded_dev_features = encode_features(dev_features, features_dict)
encoded_train_labels = encode_labels(train_tag_sents, labels2i)
encoded_dev_labels = encode_labels(dev_tag_sents, labels2i)
```

```
  0%|          | 0/75827 [00:00<?, ?it/s]
  0%|          | 0/10851 [00:00<?, ?it/s]
Building features set!
100%|██████████| 75827/75827 [00:02<00:00, 33097.46it/s]
Found 208208 features
```

In [55]:
```
# TODO: Play with hyperparameters here.
num_epochs = 5
batch_size = 20
LR=0.1
optimizer = torch.optim.SGD(model.parameters(), LR)
```

In [56]:
```
model = training_loop(
    num_epochs,
    batch_size,
    encoded_train_features,
    encoded_train_labels,
    encoded_dev_features,
    encoded_dev_labels,
    optimizer,
    model,
    labels2i,
    features_dict[PAD_SYMBOL]
)
```

```
Training...
  0%|          | 0/3792 [00:00<?, ?it/s]
/var/folders/31/yzh9j02x7bxd463cl1x0_21h0000gn/T/ipykernel_63767/3584098268.py:55: RuntimeWarning: invalid valu
e encountered in log
  print(f"epoch {i}, loss: {np.log(sum(losses)/len(losses))}")
epoch 0, loss: nan
Dev F1 log tensor([-3.8738])
  0%|          | 0/3792 [00:00<?, ?it/s]
epoch 1, loss: nan
Dev F1 log tensor([-3.8738])
  0%|          | 0/3792 [00:00<?, ?it/s]
epoch 2, loss: nan
Dev F1 log tensor([-3.8738])
  0%|          | 0/3792 [00:00<?, ?it/s]
epoch 3, loss: nan
Dev F1 log tensor([-3.8738])
  0%|          | 0/3792 [00:00<?, ?it/s]
epoch 4, loss: nan
Dev F1 log tensor([-3.8738])
```

Quiz (10 Points)

1. Look at the `NERTagger` class in crf.py

```
a) What does the CRF add to our model that makes it different from the sentiment classifier?
    We accept previous tag as a feature along with the other features has the same reult, we
take the argmax of the sum of the feature scores for each element of the sequence
b) Why is this helpful for NER?
    It takes context into account and recognize forms such as 'has lived', 'is moving'. When a
CRF model makes a prediction, it factors in the impact of neighbouring samples by modelling
the prediction as a graphical model. It assumes that the tag for the present word is dependent
only on the tag of just one previous word
```

2. Why computing F1 here is not straightforward?

Hint: Refer to the class in which Jim went over the evaluation metrics for NER

There are lots of O and you can get 80% accuracy for newspaper article because of that. It looks good but actually not. The evaluation metrics for NER can based on Tag or Entities. You can get 0 accuracy for entities such as it is facility but your model recognize as person or other tag. So it is better use span accuracy, find the prediction in the span or category, do I get subset of span or category right? Need to have different measurement to evaluate the model result. That's why it is not strightforward.

In [ ]:

Assignment Title

Programming Assignment (40 points)

The programming assignement will be an implementation of the task described in the assignment

We will make sure you have enough scaffolding to build the code upon where you would only have to implement the interesting parts of the code

Evaluation

The evaluation of the assignment will be done through test scripts that you would need to pass to get the points.

Written Assignment (60 Points)

Written assignment tests the understanding of the student for the assignment's task. We have split the writing into sections. You will need to write 1-2 paragraphs describing the sections. Please be concise.

In your own words, describe what the task is (20 points)

Describe the task, how is it useful and an example.

Section 1: PoS Tagging using HMM and Viterbi on Hindi dataset:

Task one: We will implement the Viterbi Decoder using the Forward Algorithm of Hidden Markov Model. We implement fit method to count the probabilitis of the training set, then path probability, and implement the viterbi decoding algorithm.

Task two: Then, we will create an HMM-based PoS Tagger for Hindi language using the annotated Tagset in nltk.indian

Section 2: NER w/ CRF on Hindi dataset:

I will use a CRF to implement a named entity recognition tagger. My job is to add more features to learn a better tagger. Then I need to complete the traiing loop implementation.

Describe your method for the task (10 points)

Important details about the implementation. Feature engineering, parameter choice etc.

Section 1: PoS Tagging using HMM and Viterbi on Hindi dataset:

For the fit method between state and observation, I simply just count the initial states, state to state transitions, and state to observations emissions. I use zip for creating the bi-grams. Then I fill the viterbi table by calculate product based on initial/state/observation tables. I use max for update viterbi table forwardly and use argmax to fill backpointer for each state and sequence id. I use backpointer to iterate the best path with best probabilities.

Section 2: NER w/ CRF on Hindi dataset:
In order to make more fatures, I need a gazetteer hindi dataset and a suffix hindidataset. I also need to use pos tagger pickle file that I dumped in the section 1. I need to keep track of previous word along with pos tag and next word along with pos tag. I also need to check special characters inside the text. Base on the homework two, it is easy to finish the training loop, simply random shuffle the samples using zip and empty the dynamic computation graph. The forward function is already implement and I just call the method and use loss.backward to do the backpropogate. Calculate the average loss and f1 score from the implemented function.

Experiment Results (10 points)

Typically a table summarizing all the different experiment results for various parameter choices

Section 1: PoS Tagging using HMM and Viterbi on Hindi dataset:

id of the token: 2186
No. of unique words in the corpus: 2187
No. of tags in the corpus 26

Length

train_indices	dev_indices	test_indices
432	54	54

Dev Accuracy	Test Accuracy
81.27	79.87

Section 2: NER w/ CRF on Hindi dataset:
num_epochs = 5
batch_size = 20
LR=0.1
epoch 0, loss: nan
Dev F1 log tensor([-3.8738])

Discussion (20 points)

Key takeaway from the assignment. Why is the method good? shortcomings? how would you improve? Additional thoughts?

Section 1: PoS Tagging using HMM and Viterbi on Hindi dataset:

The way to populate the parameters by counting the bi-grams is straight forward, simply just count the occurance. However, I spend tons of time to implement the decode function because dont know how to use back_pointer to find the best path and forget to use numpy.exp to compute the probability because the three tables have already been normalized and log. The viterbi table and backpointer are very useful to find the best possible sequence by given certain input.

Section 2: NER w/ CRF on Hindi dataset:

Implementing the features is the most difficult task I have met. Due to Hindi language. It do not have upper and lower case features so i have to use other features. We do not have a gazetteer and suffixes text file in the repository so I have to search the web to find decent and relatively clean gazetteer and suffixes to featurize the text. The training loop took a lot of time because featurize the text took tons of time. We built from scratch for featurize the text but there are many exist model that can help us to do similar task.