HW4: Word Segmentation

Virgile Audi
vaudi@g.harvard.edu
Nicolas Drizard
nicolasdrizard@g.harvard.edu
Github: virgodi/cs287/HW5

April 17, 2016

1 Introduction

The goal of this assignement is to tackle the NLP task of identifying and labeling contiguous segments of text. We will use sequence models and a dynamic programming method to find the best scoring sequence.

2 Problem Description

The idea is here to label continuous sequence of words with BIO tagging of different entities. The entities are the following:

1. PER: a person
2. LOC: a location
3. ORG: an organization
4. MISC:

Furthermore, this tagging method identifies the continuous group of words belonging to the same entity: the prefix B stop the current tag and begins a new one whereas the prefix I continues adding to the previous tag. However, in our solution we just cared about predicting the entity tag and then we were grouping the contiguous predictions into the same entity because the training text does not contain any B-tag.

3 Model and Algorithms

We used three different methods to solve this problem. The first two are the equivalent of first the Naive Bayes and second the logistic regression from text classification tasks. The last one introduces a customized way to train a neural architecture for this task.

3.1 Hidden Markov Model

We implement here a standard first order hidden Markov Model. The hidden states are the tags and the observed states are the features we built (word counts, capitalization...). The model can be represented with the following graphical model and requires two distribution: emission and transition.

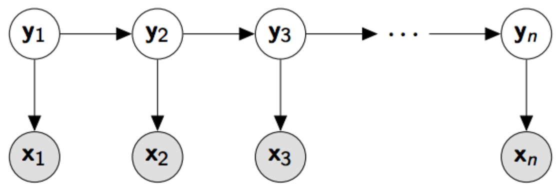


Figure 1: Graphical model of 1st order HMM with one feature

We represent the two distrubitions with multinomial as they model feature counts. As a result, we can infer them simply with the maximum likelihood estimator:

$$p(x_i = \delta(f) | y_i = \delta(c)) = \frac{F_{f,c}}{F_{.,c}}$$

$$p(y_i = \delta(c_i) | y_{i-1} = \delta(c_{i-1})) = \frac{T_{c_{i-1},c_i}}{T_{c_{i-1},.}}$$

with T_{c_{i-1},c_i} the counts of class c_{i-1} preceding class c_i and $F_{f,c}$ the counts of emission f with class c.

If we consider multiple features, then we still assume that the feature are indepent with each other (it's the main assumption in the Naive Bayes approach also). Only the emission distribution is changed and we can combine the probability together:

$$p(x_i = (\delta(f_1), \delta(f_2)) | y_i = \delta(c)) = p(x_i = \delta(f_1) | y_i = \delta(c)) p(x_i = \delta(f_2) | y_i = \delta(c)) = \frac{F_{f_1,c}}{F_{.,c}} \frac{F_{f_2,c}}{F_{.,c}}$$

3.2 Maximum-Entropy Markov Model

Next, we implemented a Maximum-Entropy Markov Model. The objective of the MEMM is to evaluate at each time step a distribution over the possible tags using features of the current word, denoted as $feat(x_i)$ and the tag of the previous word, c_{i-1}, using multi-class logistic regression, i.e.

$$p(\mathbf{y}_i | \mathbf{y}_{i-1}, feat(x_i)) = \text{softmax}([feat(x_i), c_{i-1}]\mathbf{W} + \mathbf{b})$$

3.3 Viterbi algorithm

The search algorithm that we implemented is the dynamic programming algorithm named after Andrew Viterbi. Its main difference with a greedy approach is that it evaluates at every step and for every previous state, the best possible next step. This guarantees a solution closer to the true optimal solution. The pseudo-code of the algorithm is given by:

procedure VITERBIWITHBP
  $\pi \in \mathbb{R}^{n+1 \times \mathcal{C}}$ initialized to $-\infty$
  $bp \in \mathcal{C}^{n \times \mathcal{C}}$ initialized to ϵ
  $\pi[0, \langle s \rangle] = 0$
  for $i = 1$ to n **do**
    for $c_{i-1} \in \mathcal{C}$ **do**
      compute $\hat{\boldsymbol{y}}(c_{i-1})$
      for $c_i \in \mathcal{C}$ **do**
        $score = \pi[i-1, c_{i-1}] + \log \hat{\boldsymbol{y}}(c_{i-1})_{c_i}$
        if $score > \pi[i, c_i]$ **then**
          $\pi[i, c_i] = score$
          $bp[i, c_i] = c_{i-1}$
  return sequence from bp

3.4 Structured Perceptron

The final model, we implemented is the structure perceptron train algorithm. The way the model is trained uses the Viterbi search algorithm, presented above. At each epoch, we uses Viterbi to predict the highest scored sequence given the state of the model. We can then find the timesteps where the actual sequence for the given sentence and the predicted one differ and compute at each of these time steps, the gradient of a hinge type loss. These gradients have a -1 entry on the true class for this given word, and a 1 on the predicted class by the model. We can then propagate these gradients in the network, and update the weights with a learning rate that can be tuned.

The model itself is similar to the model of the MEMM without the final logsoftmax layer.

4 Experiments

4.1 Feature Engineering

The original paper suggests several features to use. We focus on the word counts and a capitalization feature. We defined our capitalization feature as follow:

1. 1 : word in low caps;
2. 2 : whole word in caps;
3. 3 : first letter in cap;
4. 4 : one cap in the word;

5. 5 : other

We then produced an embedding of the word counts using a pre-trained version.

We also used the Python "pattern.en" package to extract Part-of-Speach (PoS) features. The packages generates 41 features to which we added special feature for the opening and closing tabs <s> and < \s>.

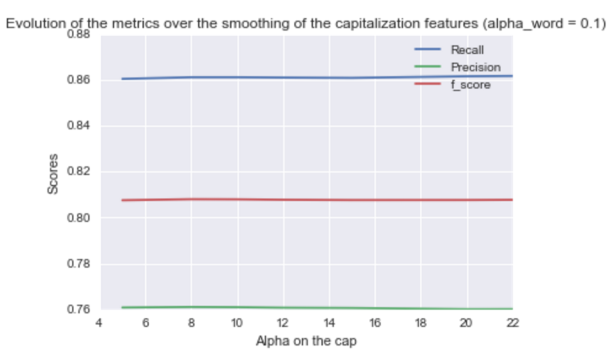
4.2 Model Evaluation

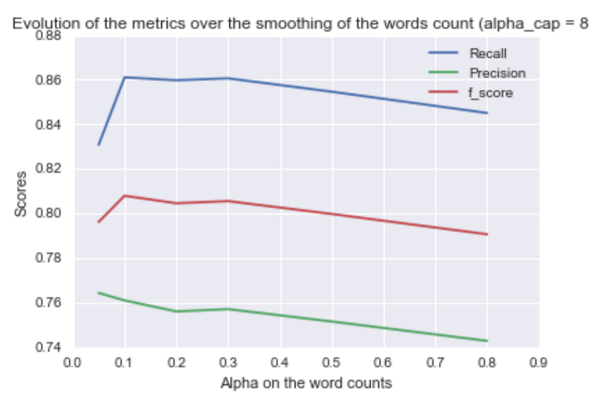
As used in the Kaggle competition, we used the f-score with the precision and recall measure to evaluate our model while tuning the hyperparameters. A positive prediction stands for a label (in the notation of the task, everything which is not the **O** tag):

1. recall: ratio of the true positive predictions among the positives tags in the correct sequence
2. precision: ratio of the true positive predictions among the positive predictions,
3. f-score (with $\beta = 1$): harmonic mean of the precision and the recall, i.e. $f_1 = \frac{2pr}{p+r}$

4.3 Hidden Markov Model

There is only the smoothing parameter α and eventually feature selection here to tune here. We evaluate the impact of adding more features and run experiments with different alpha values to tune them . One important details is to make sure to use a specific smoothing parameter for each distribution, i.e a smoothing parameter may be applied to the transition matrix but also to the emission matricx of each different feature. Each of this distribution has a different tail and need a different smoothing. For instance, the transition matrix need a very small α (around 0.1) because we are pretty confident in it but the capitalizations feature need one much bigger (around 20) because the counts are already high.





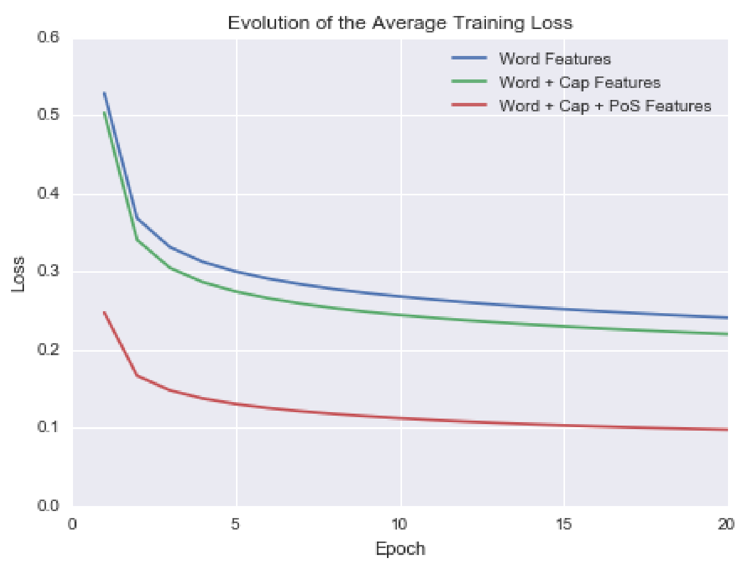
We notice that the model is less sensitive to the changes of the smoothing parameter on the capitalization feature as on the word counts. This is pretty reasonable as the feature coutns are much higher in the capitalization feature than in the word counts. Tuning this parameter provides a model with a f-score of **0.808**. Using only the word counts features provide a best f-score of **0.764**.

Adding the part-of-speech tagging feature increased the performance of the model. We got on the dev set a f-score of **0.843** We obtained a Kaggle score on the test set with the three different features of :

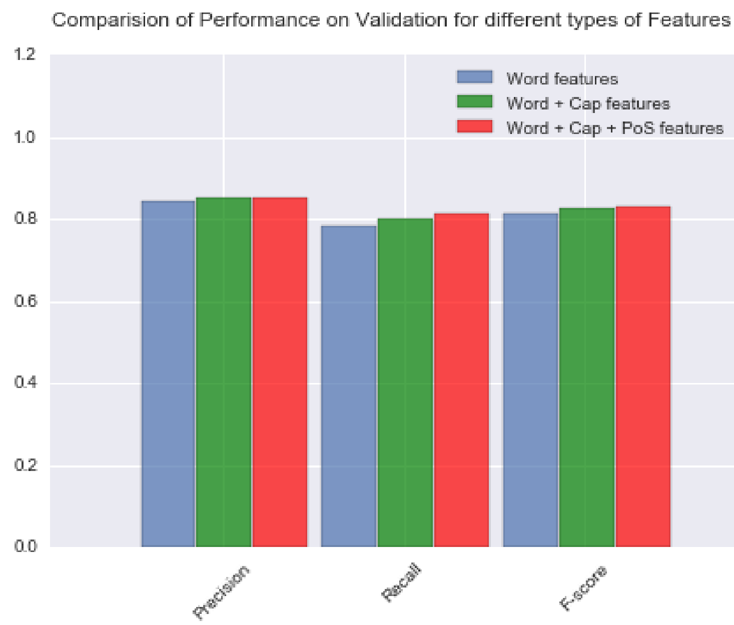
$$K_{HMM} = 0.54392$$

4.4 Maximum-Entropy Markov Model

We coded the MEMM using the nn module and trained using stochastic gradient descent. We also used the Glove embedggins using a lookup table. As for the HMM, we used two different sets of features, i.e. the words and the words and capitalisation of the words. We also added the Part of Speech features that were evaluated using the python package "pattern.en" in order to gain some time. We observed that the training algorithm converges quite rapidely, and that if adding caps to the features helped decrease the loss, the impact was not as strong as expected. On the other hand, adding PoS features impacted greatly the loss. Nevertheless, we trained the model on 20 epochs in order to learn the embeddings for the <s> and < \s> "words" added during pre-processing.



We evaluated the performance of these two models using the f-score presented above:



Adding extra features yielded better results on both Precision and Recall and therefore on the f-score. But as we expected from the small differences in loss, we did not observe an important increase on the f-score using cap features. We were nevertheless surprised to see that the impact on loss using PoS features did not translate on the f-score. These results were later confirmed on the test set, as the kaggle score obtained for these two models were:

$$K_{nocaps} = 0.52057 \quad \text{and} \quad K_{caps} = 0.55482 \quad K_{PoS} = 0.57121$$

which are both slightly better than the results of the HMM.

4.5 Structured Perceptron

We implemented the structured perceptron with the idea described in the model: weighting up the true edges in the lattice and down the incorrectly predicted. However we did not observe convincing results on our model, especially the f-score on the dev set was not increasing over the epochs but simply oscillating randomly around 0.72, which is not so bad but still less than what we obtained from the two other model.

We first tried a simple training version (our first train function in sp.lua) where for each timestep with a wrong prediction we do one forward/backward. The input is the right tag and we use a gradient with -1 on the right tag and 1 on the wrong one. We also coded an advanced version which was treating the two wrong edges of the lattice for each error (in our second train function) but did not observe the expected result.

5 Conclusion

We were disappointed to not be able to get the performance of the structured perceptron to the levels of the hidden and maximum-entropy markov models on the task of finding and labeling named-entities in text. Due to the our difficulties at implementing the perceptron, we did not get the chance of implementing the NNMEM or add more features. Nevertheless, looking at the impact of the the few extra features implemented and the performance of the multi-class logistic regression, we believe that future work in that direction would yield better results.

Appendices

Preprocessing:

```
import numpy as np
import h5py
import re
import pattern.en
import sys
import argparse

from itertools import product


def get_tag2index():
    # Tags mapping
    tag2index = {}

    with open('data/tags.txt', 'r') as f:
        for line in f:
            line_split = line[:-1].split(' ')
            tag2index[line_split[0]] = int(line_split[1])

    # Adding tags for end/start of sentence
    tag2index['<t>'] = 8
    tag2index['<\t>'] = 9
    return tag2index


def get_pos2index():
    '''
    Part of speech tagging tags to feature index mapping
    '''
    # mapping for the POS tags
    tags = ['CC', 'CD', 'DT', 'EX', 'FW', 'IN', 'JJ', 'JJR', 'JJS', 'LS
        ', 'MD',
            'NN', 'NNS', 'NNP', 'NNPS', 'PDT', 'POS',
            'PRP', 'PRP$', 'RB', 'RBR', 'RBS', 'RP', 'SYM', 'TO', 'UH',
                'VB',
```

```
            'VBZ', 'VBP', 'VBD', 'VBN', 'VBG', 'WDT', 'WP', 'WP$', 'WRB',
            '.', ',', ':', '(', ')']

    pos2index = {k: v+1 for v, k in enumerate(tags)}
    return pos2index


def count_elements(filename, tags=True):
    # Counting the number of elements to stored (ie num_words +
    # 2*num_sentences)
    num_words = 0
    num_sentences = 0
    with open(filename, 'r') as f:
        for line in f:
            if tags:
                line_split = line[:-1].split('\t')
            else:
                line_split = line[:-1].split(' ')
            # Case blank
            if len(line_split) == 1:
                num_sentences += 1
            else:
                num_words += 1

    return num_words, num_sentences


def get_cap_feature(word):
    # Return the caps feature for the given word
    # 1 - low caps; 2 - all caps; 3 - first cap; 4 - one cap; 5 - other
    if len(word) == 0 or word.islower() or re.search('[.?\-",]+', word):
        feature = 1
    elif word.isupper():
        feature = 2
    elif len(word) and word[0].isupper():
        feature = 3
    elif sum([w.isupper() for w in word]):
        feature = 4
    else:
        feature = 5
    return feature
```

```
def get_tokenized_sentences(filename, tags=True):
    # Build the part of speech tags
    with open(filename, 'r') as f:
        text = []
        for line in f:
            if tags:
                line_split = line[:-1].split('\t')
            else:
                line_split = line[:-1].split(' ')
            if len(line_split) != 1:
                text.append(line_split[2])

    return pattern.en.tag(' '.join(text))


def build_input_matrix(filename, num_rows, tag2index, pos2index, tags=True, word2index=None, memm = False):
    # Building input matrix with columns: (id, id_in_sentence, id_word, id_caps, id_token, id_tag)
    # caps feature:
    # 1 - low caps; 2 - all caps; 3 - first cap; 4 - one cap; 5 - other
    # Tags: if correct solution given (ie 4th column)
    # word2index: if use of previously built word2index mapping

    # Features for starting/ending of sentence (3 last columns)
    # For the POS tag, we use the same as a point (index 36)
    # initialization
    input_matrix = np.zeros((num_rows, 6), dtype=int)
        if memm ==  False:
        input_matrix[0] = [1, 1, 1, 1, 36, 8]
        start = [1, 1, 36, 8]
        end = [2, 1, 36, 9]
    else:
        input_matrix[0] = [1,1,word2index['<s>'],1,36,8]
        start = [word2index['<s>'],1,36, 8]
                end = [word2index['<\s>'],1,36, 9]
    row = 1

    # Get the POS tokken
    tokenized_sentences = get_tokenized_sentences(filename, tags=tags)
    pos_i = 0

    # Boolean to indicate if a sentence is starting
    starting = False
```

```
    # Boolean if a mapping is defined (last element of the mapping is for
    # unknown words)
    if word2index == None:
        test = False
        word2index = {'<s>': 1, '<\s>': 2}
        id_word = 3
    else:
        test = True
    with open(filename, 'r') as f:
        for line in f:
            if tags:
                line_split = line[:-1].split('\t')
            else:
                line_split = line[:-1].split(' ')
            if starting == True:
                # Start of sentence
                input_matrix[row, 0] = input_matrix[row-1, 0] + 1
                input_matrix[row, 1] = 1
                input_matrix[row, 2:] = start
                row += 1
                starting = False
            if len(line_split) == 1:
                # End of sentence
                input_matrix[row, :2] = input_matrix[row-1, :2] + 1
                input_matrix[row, 2:] = end
                row += 1
                starting = True
            else:
                # Indexing
                input_matrix[row, 0] = input_matrix[row-1, 0] + 1
                input_matrix[row, 1] = int(line_split[1]) + 1
                # Build cap feature
                word = line_split[2]
                input_matrix[row, 3] = get_cap_feature(word)
                # Build pos feature
                pos_tag = tokenized_sentences[pos_i][1].split('-')[0]
                if pos_tag in pos2index.keys():
                    input_matrix[row, 4] = pos2index[pos_tag]
                else:
                    input_matrix[row, 4] = len(pos2index) + 1
                pos_i += 1

                # Build word count feature
                word_clean = word.lower()
```

```
                if not test:
                    if word_clean not in word2index:
                        word2index[word_clean] = id_word
                        id_word += 1
                    input_matrix[row, 2] = word2index[word_clean]
                else:
                    # Unseen word during train
                    if word_clean not in word2index:
                        input_matrix[row, 2] = len(word2index)
                    else:
                        input_matrix[row, 2] = word2index[word_clean]
                if tags:
                    input_matrix[row, 5] = tag2index[line_split[3]]
                row += 1
    # Add special word if training
    if not test:
        word2index['<unk>'] = len(word2index)+1
    if tags:
        return input_matrix, word2index
    else:
        return input_matrix[:, :5], word2index

#Function that formats the output of the previous function in order to run MEMM:
def input_mm_pos(matrix):

    nwords = matrix.shape[0]

    res = np.zeros((nwords,1 + 9 + 5 + 43 + 1),dtype = int)

    res[:,0] = matrix[:,2]

    for i in range(nwords):
        tag_1_hot = np.zeros(9)
        tag_1_hot[matrix[i,5]-1] = 1
        tag_1_hot_cap = np.zeros(5)
        tag_1_hot_cap[matrix[i,3]-1] = 1
        tag_1_hot_pos = np.zeros(43)
        tag_1_hot_pos[matrix[i,4]] = 1
        res[i,1:10] = tag_1_hot
        res[i,10:15] = tag_1_hot_cap
        res[i,15:58] = tag_1_hot_pos
    res[:,58] = matrix[:,5]
    return res
```

```
def train_hmm(input_matrix, num_features, num_pos, num_tags):
    # Emission word_count matrix:
    # size (num_words, num_tags)
    # row: observation / colum: tag
    # (un-normalized if smoothing required)
    emission_w = np.zeros((num_features, num_tags), dtype=int)

    # Emission caos_count matrix:
    # size (5, num_tags)
    # row: observation / colum: caps
    # (un-normalized if smoothing required)
    emission_c = np.zeros((5, num_tags), dtype=int)

    # Emission pos_count matrix:
    # size (5, num_tags)
    # row: observation / colum: pos tag
    # (un-normalized if smoothing required)
    emission_p = np.zeros((num_pos, num_tags), dtype=int)

    # Building
    for r in input_matrix:
        emission_w[r[2]-1, r[5]-1] += 1
        emission_c[r[3]-1, r[5]-1] += 1
        emission_p[r[4]-1, r[5]-1] += 1

    # Transition matrix
    # size (num_tags, num_tags)
    # row: to / colum: from
    # (un-normalized if smoothing required)
    transition = np.zeros((num_tags, num_tags), dtype=int)
    for i in xrange(input_matrix.shape[0] - 1):
        transition[input_matrix[i+1, 5]-1, input_matrix[i, 5]-1] += 1

    return emission_w, emission_c, emission_p, transition


def main(arguments):
    # Args
    global args
    parser = argparse.ArgumentParser(
        description=__doc__,
        formatter_class=argparse.RawDescriptionHelpFormatter)

    parser.add_argument('--f', default='data/features.hdf5',
```

```
                         type=str, help='Filename to save data')
    args = parser.parse_args(arguments)
    filename = args.f

    # Train
    pos2index = get_pos2index()
    tag2index = get_tag2index()
    num_words, num_sentences = count_elements('data/train.num.txt')
    num_rows = num_words + 2*num_sentences
    input_matrix_train, word2index = build_input_matrix('data/train.num.txt',
                                                        num_rows, tag2index,
                                                        pos2index)

    # Building the count matrix
    num_tags = len(tag2index)
    num_features = len(word2index)
    num_pos = len(pos2index) + 1
    emission_w, emission_c, emission_p, transition = train_hmm(input_matrix_train,
                                                               num_features, num_pos,
                                                               num_tags)

    # Dev & test
    num_words, num_sentences = count_elements('data/dev.num.txt')
    # Miss 1 blank line at the end of the file for the dev set
    num_rows = num_words + 2*num_sentences + 1
    input_matrix_dev, word2index = build_input_matrix('data/dev.num.txt',
                                                      num_rows, tag2index,
                                                      pos2index,
                                                      word2index=word2index)

    num_words, num_sentences = count_elements('data/test.num.txt',
                                              tags=False)
    num_rows = num_words + 2*num_sentences
    input_matrix_test, word2index = build_input_matrix('data/test.num.txt',
```

```
                                                            num_rows,
                                                               tag2index,
                                                            pos2index,
                                                            tags=False,
                                                            word2index=
                                                               word2index)

    # Saving pre-processing
    with h5py.File(filename, "w") as f:
        # Model
        f['emission_w'] = emission_w
        f['emission_c'] = emission_c
        f['emission_p'] = emission_p
        f['transition'] = transition

        f['input_matrix_train'] = input_matrix_train
        f['input_matrix_dev'] = input_matrix_dev
        f['input_matrix_test'] = input_matrix_test


if __name__ == '__main__':
    sys.exit(main(sys.argv[1:]))
```

<u>Hidden Markov Model:</u>

```
-- Documentation:
-- ---- How to call it from the command line?
-- For example:
-- $ th count_based.lua -N 5
-- Other argument possible (see below)
--
-- ---- Is there an Output?
-- By default, the predictions on the test set are saved in hdf5 format
   as classifier .. opt.f

-- Only requirements allowed
require("hdf5")
require 'helper.lua';

cmd = torch.CmdLine()

-- Cmd Args
cmd:option('-datafile', 'data/words_feature.hdf5',
           'Datafile with features in hdf5 format')
cmd:option('-alpha_t', 0.1, 'Smoothing parameter alpha in the
   transition counts')
```

```
cmd:option('-alpha_w', 0.1, 'Smoothing parameter alpha in the word counts')
cmd:option('-alpha_c', 8, 'Smoothing parameter alpha in the caps counts')
cmd:option('-alpha_p', 2, 'Smoothing parameter alpha in the pos counts')
cmd:option('-test', 0, 'Boolean (as int) to ask for a prediction on test, will be saved in submission in hdf5 format')
cmd:option('-datafile_test', 'submission/v_seq_hmm', 'Smoothing parameter alpha in the word counts')
cmd:option('-nfeatures', 2, 'Number of type of features to use')
cmd:option('-cv', 0, 'Boolean (as int) to run a cross validation pipeline')



-- Formating as log-probability and smoothing the input
function format_matrix(matrix, alpha)
    local formatted_matrix = matrix:clone():type('torch.DoubleTensor')
    formatted_matrix:add(alpha)
    -- Normalize
    local norm_mat = torch.expandAs(formatted_matrix:sum(1), formatted_matrix)
    formatted_matrix:cdiv(norm_mat)
    return formatted_matrix:log()
end

-- log-scores of transition and emission
-- corresponds to the vector y in the lecture notes
-- i: timestep for the computed score
function score_hmm(observations, i, emissions, transition, C, nfeatures)
    local observation_emission = torch.zeros(C)
    for k=1,nfeatures do
        -- print(i,k)
        -- print(emissions[k][observations[{i,k}]])
        observation_emission:add(emissions[k][observations[{i,k}]])
    end
    observation_emission = observation_emission:view(C, 1):expand(C, C)
    -- NOTE: allocates a new Tensor
    return observation_emission + transition
end

-- Viterbi algorithm.
-- observations: a sequence of observations, represented as integers
```

```
-- logscore: the edge scoring function over classes and observations in a history-based model
function viterbi(observations, logscore, emissions, transition, nfeatures)
    local y
    -- Formating tensors
    local initial = torch.zeros(transition:size(2), 1)
    -- initial started with a start of sentence: <t>
    initial[{8,1}] = 1
    initial:log()

    -- number of classes
    C = initial:size(1)
    local n = observations:size(1)
    local max_table = torch.Tensor(n, C)
    local backpointer_table = torch.Tensor(n, C)

    -- first timestep
    -- the initial most likely paths are the initial state distribution
    -- NOTE: another unnecessary Tensor allocation here
    local init_pred = initial:clone()
    for i=1,nfeatures do
        init_pred:add(emissions[i][observations[{1,i}]])
    end
    local maxes, backpointers = init_pred:max(2)
    max_table[1] = maxes

    -- remaining timesteps ("forwarding" the maxes)
    for i=2,n do
        -- precompute edge scores
        y = logscore(observations, i, emissions, transition, C, nfeatures)
        scores = y + maxes:view(1, C):expand(C, C)

        -- compute new maxes (NOTE: another unnecessary Tensor allocation here)
        maxes, backpointers = scores:max(2)

        -- record
        max_table[i] = maxes
        backpointer_table[i] = backpointers
    end
    -- follow backpointers to recover max path
    local classes = torch.Tensor(n)
    maxes, classes[n] = maxes:max(1)
```

```
    for i=n,2,-1 do
        classes[i-1] = backpointer_table[{i, classes[i]}]
    end

    return classes
end

-- Prediction pipeline
function predict(observations, emissions, transition, alphas, nfeatures)
    -- Formating model parameters (log and alpha smoothing)
    -- Alphas is a tensor : {alpha_t, alpha_w, alpha_c}
    emissions_cleaned = {}
    for i=1,nfeatures do
        emissions_cleaned[i] = format_matrix(emissions[i], alphas[i+1])
    end
    local transition_cleaned = format_matrix(transition, alphas[1])

    return viterbi(observations, score_hmm, emissions_cleaned, transition_cleaned, nfeatures)
end

-- Cross validation pipeline
function cross_validation(observations, emissions, transitions, true_classes,
                            alphas_table, alpha_t)
    -- alphas_table is a table of tensor with the range of parameters to use
    -- Current implementation for 3 features only
    -- alphas_table = {alpha_w_tensor, alpha_c_tensor}
    -- Return a tensor with first columns the alpha value and last the score for each
    local nfeatures = #alphas_table
    local v_seq_dev, precision, recall, f
    local alphas = torch.DoubleTensor(1+nfeatures)
    local size1 = alphas_table[1]:size(1)
    local size2 = alphas_table[2]:size(1)
    local size3 = alphas_table[3]:size(1)
    local num_evaluations = size1*size2*size3
    local score_ind = 1

    -- Columns for 2 features are (alphas_w_value, alphas_c_value, f_score, precision, recall)
    local scores = torch.DoubleTensor(num_evaluations, nfeatures+3)

```

```
    for i=1,size1 do
        alpha_w = alphas_table[1][i]
        for k=1,size2 do
            alpha_c = alphas_table[2][k]
            for j=1,size3 do
                alpha_p = alphas_table[3][j]
                alphas:copy(torch.Tensor({alpha_t, alpha_w, alpha_c,
                    alpha_p}))
                v_seq_dev = predict(observations, emissions, transition
                    , alphas, nfeatures)
                precision, recall = compute_score(v_seq_dev,
                    true_classes)
                f = f_score(precision, recall)
                -- Filling the scores tensor
                scores[{score_ind, 1}] = alpha_w
                scores[{score_ind, 2}] = alpha_c
                scores[{score_ind, 3}] = alpha_p
                scores[{score_ind, 4}] = f
                scores[{score_ind, 5}] = precision
                scores[{score_ind, 6}] = recall
                score_ind = score_ind + 1
            end
        end
    end

    return scores
end


function main()
    -- Parse input params
    opt = cmd:parse(arg)

    -- Reading file from pre-processing
    myFile = hdf5.open(opt.datafile,'r')
    data = myFile:all()
    emission_w = data['emission_w']
    emission_c = data['emission_c']
    emission_p = data['emission_p']
    print(emission_p:size())
    -- Table of emission tensor (one tensor per feature)
    emissions = {emission_w, emission_c, emission_p}
    -- Assertion on number of features
    nfeatures = opt.nfeatures
    if nfeatures > #emissions then
```

```
        error('Too many features specified')
    end
    print('Number of features used:  '..nfeatures)
    transition = data['transition']
    input_matrix_train = data['input_matrix_train']
    input_matrix_dev = data['input_matrix_dev']
    input_matrix_test = data['input_matrix_test']
    myFile:close()

    -- Parameters:
    true_classes = input_matrix_dev:narrow(2,6,1):clone():view(input_matrix_dev:size(1))
    -- contain in each column feature observation
    -- (same order as the feature emission tensor in the emissoins table)
    observations = input_matrix_dev:narrow(2,3,nfeatures):clone()
    -- Alpha parameter
    alphas = torch.Tensor({opt.alpha_t, opt.alpha_w, opt.alpha_c, opt.alpha_p})

    -- Prediction on dev
    v_seq_dev = predict(observations, emissions, transition, alphas, nfeatures)
    print(v_seq_dev:size(1))
    precision, recall = compute_score(v_seq_dev, true_classes)
    f = f_score(precision, recall)

    print('Prediction on dev')
    print('Precision is :  '..precision)
    print('Recall is :  '..recall)
    print('F score (beta = 1) is :  '..f)

    -- Cross validation
    if (opt.cv == 1) then
        alphas_table = {}
        -- alpha_w
        alphas_table[1] = torch.Tensor({0.1, 0.2, 0.3, 0.4, 0.5})
        -- alpha_c
        alphas_table[2] = torch.Tensor({5, 8, 10, 12})
        -- alpha_p
        alphas_table[3] = torch.Tensor({1, 2, 4, 6})

        scores = cross_validation(observations, emissions, transitions, true_classes,
                                    alphas_table, opt.alpha_t)
```

```
        print(scores)

        -- Saving the score
        myFile = hdf5.open('plot_scores.hdf5', 'w')
        myFile:write('scores', scores)
        myFile:close()
        print('CV on dev saved in '..'plot_scores.hdf5')
    end

    -- Prediction on test
    if (opt.test == 1) then
        print('Prediction on test')
        observations_test = input_matrix_test:narrow(2,3,nfeatures):clone()
        v_seq_test = predict(observations_test, emissions, transition, alphas, nfeatures)
        -- Saving predicted sequence on test
        myFile = hdf5.open(opt.datafile_test, 'w')
        myFile:write('v_seq_test', v_seq_test)
        myFile:write('v_seq_dev', v_seq_dev)
        myFile:close()
        print('Sequence predicted on test saved in '..opt.datafile_test)
    end

end

main()
```

Max-Entropy Markov Model:

```
require 'hdf5';
require 'nn';
require 'helper.lua';

-- Loading data
myFile = hdf5.open('../data/MM_data_pos.hdf5','r')
data = myFile:all()
input_matrix_train_pos = data['input_matrix_train_pos']
input_matrix_dev_pos = data['input_matrix_dev_pos']
input_matrix_test_pos = data['input_matrix_test_pos']
myFile:close()

nwords = input_matrix_train_pos:size(1)
train_output = input_matrix_train_pos:narrow(2,59)
train_input_pos = torch.Tensor(nwords-1,1+9+5+43)
```

```
train_input_pos:narrow(2,1,1):copy(input_matrix_train_pos:narrow(2,1,1)
    :narrow(1,2,nwords-1))
train_input_pos:narrow(2,2,9):copy(input_matrix_train_pos:narrow(2,2,9)
    :narrow(1,1,nwords-1))
train_input_pos:narrow(2,11,5):copy(input_matrix_train_pos:narrow
    (2,11,5):narrow(1,1,nwords-1))
train_input_pos:narrow(2,16,43):copy(input_matrix_train_pos:narrow
    (2,16,43):narrow(1,1,nwords-1))

observations_dev = input_matrix_dev_pos:narrow(2,1,1):clone()
dev_feat = input_matrix_dev_pos:narrow(2,11, 5 + 43)
dev_true_classes = input_matrix_dev_pos:narrow(2, 59,1):squeeze()

observations_test_pos = input_matrix_test_pos:narrow(2,1,1)
observations_test_feat = input_matrix_test_pos:narrow(2,2,5+43)

-- Defining the model

model = nn.Sequential()
t1_pos = nn.ParallelTable()

t1_pos_1 = nn.Sequential()
t1_pos_1:add(LT)
t1_pos_1:add(nn.View(-1,50))

t1_pos_2 = nn.Identity()

t1_pos:add(t1_pos_1)
t1_pos:add(t1_pos_2)

model:add(t1_pos)
model:add(nn.JoinTable(2))

model:add(nn.Linear(50 + 9 + 5 + 43,9))
model:add(nn.LogSoftMax())

-- Training function:


function train_model_cap(train_input, train_output, model, criterion,
    din, nclass, eta, nEpochs, batchSize)
    -- Train the model with a mini batch SGD
    -- standard parameters are
    -- nEpochs = 1
    -- batchSize = 32
```

```
    -- eta = 0.01
    local loss = torch.Tensor(nEpochs)

    -- To store the loss
    local av_L = 0

    -- Memory allocation
    local inputs_batch = torch.DoubleTensor(batchSize, din)
    local targets_batch = torch.DoubleTensor(batchSize)
    local outputs = torch.DoubleTensor(batchSize, nclass)
    local  df_do = torch.DoubleTensor(batchSize, nclass)

    for i = 1, nEpochs do
        -- timing the epoch
        timer = torch.Timer()
        av_L = 0

        -- mini batch loop
        for t = 1, train_input:size(1), batchSize do
            -- Mini batch data
            current_batch_size = math.min(batchSize,train_input:size(1)-t)

            inputs_batch:narrow(1,1,current_batch_size):copy(train_input:narrow(1,t,current_batch_size))

            targets_batch:narrow(1,1,current_batch_size):copy(train_output:narrow(1,t,current_batch_size))

            -- reset gradients
            model:zeroGradParameters()

            -- Forward pass (selection of inputs_batch in case the batch is not full, ie last batch)
            outputs:narrow(1,1,current_batch_size):copy(model:forward({inputs_batch:narrow(1,1,current_batch_size):narrow(2,1,1),
            inputs_batch:narrow(1,1,current_batch_size):narrow(2,2,din-1)}))
            -- Average loss computation
            f = criterion:forward(outputs:narrow(1,1,current_batch_size), targets_batch:narrow(1,1,current_batch_size))

            av_L = av_L +f

```

```
            -- Backward pass
            df_do:narrow(1,1,current_batch_size):copy(criterion:
               backward(outputs:narrow(1,1,current_batch_size),
               targets_batch:narrow(1,1,current_batch_size)))
            model:backward({inputs_batch:narrow(1,1,current_batch_size)
               :narrow(2,1,1), inputs_batch:narrow(1,1,
               current_batch_size):narrow(2,2,din-1)},
            df_do:narrow(1,1,current_batch_size))

            model:updateParameters(eta)

        end

        print('Epoch '..i..': '..timer:time().real)
        loss[i] = av_L/math.floor(train_input:size(1)/batchSize)
        print('Average Loss: '.. loss[i])

    end

    return loss
end

-- Viterbi for MEMM:

-- Evaluates the matrix of scores for all possible tags for the
   previous word, using the word features at timestep i

function compute_logscore_extrafeat(observations, feat, i, model, C)
    local y = torch.zeros(C,C)
    local hot_1 = torch.zeros(C+feat:size(2))
    for j = 1, C do
        hot_1:zero()
        hot_1[j] = 1
        hot_1:narrow(1,10,feat:size(2)):copy(feat:narrow(1,i,1))
        y:narrow(1,j,1):copy(model:forward({observations[i]:view(1,1),
           hot_1:view(1,C+feat:size(2))}))
    end
    return y
end

-- Evaluates the highest scoring sequence:
function viterbi_extrafeat(observations, feat, compute_logscore, model,
    C)

    local y = torch.zeros(C,C)
```

```
    -- Formating tensors
    local initial = torch.zeros(C, 1)
    -- initial started with a start of sentence: <t>

    initial[{8,1}] = 1
    initial:log()

    -- number of classes
    local n = observations:size(1)
    local max_table = torch.Tensor(n, C)
    local backpointer_table = torch.Tensor(n, C)
    -- first timestep
    -- the initial most likely paths are the initial state distribution
    local maxes, backpointers = (initial + compute_logscore_extrafeat(observations, feat, 1, model, C)[8]):max(2)
    max_table[1] = maxes
    -- remaining timesteps ("forwarding" the maxes)
    for i=2,n do
        -- precompute edge scores

        y:copy(compute_logscore_extrafeat(observations, feat, i, model, C))
        scores = y:transpose(1,2) + maxes:view(1, C):expand(C, C)

        -- compute new maxes
        maxes, backpointers = scores:max(2)

        -- record
        max_table[i] = maxes
        backpointer_table[i] = backpointers
    end
    -- follow backpointers to recover max path
    local classes = torch.Tensor(n)
    maxes, classes[n] = maxes:max(1)
    for i=n,2,-1 do
        classes[i-1] = backpointer_table[{i, classes[i]}]
    end

    return classes
end

-- Train Model

loss_pos = train_model_cap(train_input_pos, train_output, ultimate_t_pos, criterion, 1 + 9 + 5 + 43, 9, 0.1, 20, 32)
```

```
-- Evaluate performance on dev set:

cl_pos_dev = viterbi_extrafeat(observations_dev, dev_feat,
    compute_logscore_extrafeat, ultimate_t_pos, 9)
f = f_score(cl_pos_dev, dev_true_classes)

-- Predict on test:

v_seq_test_pos = viterbi_extrafeat(observations_test_pos,
    observations_test_feat, compute_logscore_extrafeat, ultimate_t_pos,
    9)

-- Saving predicted sequence on test
myFile = hdf5.open('../submission/v_seq_test_mem_pos', 'w')
myFile:write('v_seq_test', v_seq_test_pos)
myFile:close()
```

Structured Perceptron:

```
function train_model(train_input, sent, train_output, observations_dev,
     model, din, nclass, eta, nEpochs)
    -- Train the model with the structured perceptron approach
    -- V1: only treating the eged leaving the error

    -- For the verbose print
    observations = observations_dev:narrow(2,1,1):narrow(1,1,1000):
        clone()
    true_classes = observations_dev:narrow(2,16,1):narrow(1,1,1000):
        squeeze()

    -- Memory allocation
    inputs_batch = torch.DoubleTensor(100, din)
    gold_sequence = torch.DoubleTensor(100)
    high_score_seq = torch.DoubleTensor(100)
    grad_pos = torch.zeros(9)
    grad_neg = torch.zeros(9)
    pr1 = torch.zeros(9)
    pr2 = torch.zeros(9)

    for i = 1, nEpochs do
        -- timing the epoch
        timer = torch.Timer()

        -- mini batch loop
        for t = 2, sent:size(1)-1 do
```

```
            -- Mini batch data
            sent_size = sent[{t,2}]
--          print('here1')

            inputs_batch:narrow(1,1,sent_size+1):copy(train_input:narrow(1,sent[{t,1}]-1,sent_size+1))
--          print('here2')

            gold_sequence:narrow(1,1,sent_size+1):copy(train_output:narrow(1,sent[{t,1}]-1,sent_size+1))
--          print('here3')

            -- reset gradients
            model:zeroGradParameters()
            --gradParameters:zero()

            -- Forward pass on a batch subsequence:
            high_score_seq:narrow(1,1,sent_size+1):copy(viterbi(inputs_batch:narrow(1,1,sent_size+1):narrow(2,1,1),
                                                                compute_logscore,model,nclass))
--          print('here4')


            for ii = 1, sent_size+1 do
                grad_pos:zero()
                if high_score_seq[ii] ~= gold_sequence[ii] then
                    -- WARNING: Need to call backward right after the forward with the same input to compute correct gradients

                    -- Use of a single gradient (grad_pos) with a penalization on the wrong class predicted (1)
                    -- and a valorisation (-1) on the correct class to predict
                    -- We treat here only the transition after the error
                    model:forward({inputs_batch:narrow(1,ii,1):narrow(2,1,1),inputs_batch:narrow(1,ii,1):narrow(2,2,9)})
                    grad_pos[gold_sequence[ii]] = -1
```

```
                    grad_pos[high_score_seq[ii]] = 1
                    model:backward({inputs_batch:narrow(1,ii,1):narrow
                        (2,1,1),inputs_batch:narrow(1,ii,1):narrow
                        (2,2,9)}, grad_pos:view(1,9))


                end
            end
--              print('here7')
            model:updateParameters(eta)

        end

        print('Epoch '..i..': '..timer:time().real)
        -- Print the f-score on a the first 1000 words to follow the
            improvement of the model
        cl = viterbi(observations, compute_logscore, model, 9)
        print (f_score(cl, true_classes))

    end
end

function train_model2(train_input, sent, train_output, model, din,
    nclass, eta, nEpochs, obs_val, true_val, f_score)
    -- Train the model with the structured perceptron approach
    -- V2: treating the two edges, the one leading to the error and the
    -- one leaving the error.

    val_res = torch.zeros(nEpochs,3)
    -- Memory allocation
    inputs_batch = torch.DoubleTensor(100, din)
    gold_sequence = torch.DoubleTensor(100)
    high_score_seq = torch.DoubleTensor(100)
    grad_pos = torch.zeros(9)
    grad_neg = torch.zeros(9)
    one_hot_true = torch.zeros(1,9)
    one_hot_false = torch.zeros(1,9)

    for i = 1, nEpochs do
        -- timing the epoch
        timer = torch.Timer()

        -- mini batch loop
        for t = 2, sent:size(1)-1 do
            -- Mini batch data
```

```
        sent_size = sent[{t,2}]
--          print('here1')

        inputs_batch:narrow(1,1,sent_size+1):copy(train_input:narrow(1,sent[{t,1}]-1,sent_size+1))
--          print('here2')

        gold_sequence:narrow(1,1,sent_size+1):copy(train_output:narrow(1,sent[{t,1}]-1,sent_size+1))
--          print('here3')

        -- reset gradients
        model:zeroGradParameters()
        --gradParameters:zero()

        -- Forward pass on a batch subsequence:
        high_score_seq:narrow(1,1,sent_size+1):copy(viterbi(inputs_batch:narrow(1,1,sent_size+1):narrow(2,1,1),
                                                              compute_logscore, model, nclass))
--          print('here4')

        previous_error = false

        for ii = 1, sent_size+1 do

            grad_neg:zero()
            grad_pos:zero()

            if high_score_seq[ii] ~= gold_sequence[ii] and not previous_error then
                -- WARNING: Need to call backward right after the forward with the same input to compute correct gradients

                -- Use of a single gradient (grad_pos) with a penalization on the wrong class predicted (1)
                -- and a valorisation (-1) on the correct class to predict

                model:forward({inputs_batch:narrow(1,ii,1):narrow
```

```
(2,1,1),inputs_batch:narrow(1,ii,1):narrow(2,2,9)})
            grad_pos[gold_sequence[ii]] = -1
            grad_pos[high_score_seq[ii]] = 1
            model:backward({inputs_batch:narrow(1,ii,1):narrow(2,1,1),inputs_batch:narrow(1,ii,1):narrow(2,2,9)}, grad_pos:view(1,9))

            grad_neg:zero()
            grad_pos:zero()
            if ii ~= (sent_size + 1) then
                one_hot_true:zero()
                one_hot_true[1][gold_sequence[ii]] = 1
                model:forward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_true})
                grad_neg[gold_sequence[ii+1]] = -1
                model:backward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_true}, grad_neg:view(1,9) )

                one_hot_false:zero()
                one_hot_false[1][high_score_seq[ii]] = 1
                model:forward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_false})
                grad_pos[gold_sequence[ii+1]] = 1
                model:backward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_false}, grad_pos:view(1,9) )
            end

            previous_error = true

        elseif high_score_seq[ii] ~= gold_sequence[ii] and previous_error then

            if ii ~= sent_size + 1 then
                one_hot_true:zero()
                one_hot_true[1][gold_sequence[ii]] = 1
                model:forward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_true})
                grad_neg[gold_sequence[ii+1]] = -1
                model:backward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_true}, grad_neg:view(1,9) )

```

```
                    one_hot_false:zero()
                    one_hot_false[1][high_score_seq[ii]] = 1
                    model:forward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_false})
                    grad_pos[gold_sequence[ii+1]] = 1
                    model:backward({inputs_batch:narrow(1,ii+1,1):narrow(2,1,1),one_hot_false}, grad_pos:view(1,9) )
                end

                previous_error = true

            else
                previous_error = false
            end
        end
--          print('here7')
        model:updateParameters(eta)

    end

    print('Epoch '..i..': '..timer:time().real)
    cl = viterbi(obs_val, compute_logscore, model, 9)
    val_res[i][1], val_res[i][2], val_res[i][3] = f_score(cl, true_val)
    print('f-score: '.. val_res[i][1])

  end
  return val_res
end
```

Helper:

```
-- function to evaluate the predicted sequence
-- need to compute precision and recall (class 1 stands for negative class)
function compute_score(predicted_classes, true_classes)
    print('here')
    local n = predicted_classes:size(1)
    local right_pred = 0
    local positive_true = 0
    local positive_pred = 0
    for i=1,n do
        if predicted_classes[i] > 1 then
            positive_pred = positive_pred + 1
        end
```

```
        if true_classes[i] > 1 then
            positive_true = positive_true + 1
        end
        if (true_classes[i] == predicted_classes[i]) and true_classes[i] > 1 then
            right_pred = right_pred + 1
        end
    end
    local precision = right_pred/positive_pred
    local recall = right_pred/positive_true
    return precision, recall
end

function f_score(precision, recall)
    return 2*precision*recall/(precision+recall)
end
```