

## 2 Maximum likelihood estimation

## 3 First-order hidden Markov model

## 4 Second-order hidden Markov model

## 5 Design challenge

### 5.1 Preprocessing

#### 5.1.1 Tokenizing

Each file can be considered a corpus, the documents being each sentence in the file. Tokenizing was carried out by converting each sentence into a list of words. During tokenization, all whitespace was stripped from the beginning and end of each word. Also, each token is normalized by converting to lowercase.

#### 5.1.2 Replacing irrelevant words

In the initial problem, the training data contains a large vocabulary of words which occurred only a single time. To mitigate this issue, several types of words were converted into various placeholders instead, as they do not carry semantic meaning.

The placeholders used are as follows:

1. #PUNC#: strings that do not contain alphanumeric characters
2. #HASH#: hashtags (strings beginning with a # character)
3. #AT#: @ mentions (strings beginning with a @ character)
4. #NUM#: strings that contain only digits
5. #URL#: strings that begin with `http:` and end with `.com`

#### 5.1.3 Stop words

Stop words were also initially converted to a token labeled #STOP#. However, during evaluation on the EN and SG dataset with Part 3, it was shown to reduce the  $F_1$  score on both the *Correct Entity* and *Correct Entity Type* tasks. It was also shown to produce poorer  $F_1$  scores on the EN dataset with Part 5.

Taking into account the negative correlation between removing stop words and  $F_1$  scores, as well as the fact that we do not have French stop words, we decided not to remove stop words from the training data.

## 5.2 Inputs and labels

### 5.3 RNN

#### 5.3.1 Forward

We use a RNN, with a computational graph as shown in Figure 1. The RNN maps an input sequence  $\mathbf{x}$  to an output sequence  $\mathbf{o}$ . The predicted labels are then given by  $\hat{\mathbf{y}} = \text{softmax}(\mathbf{o})$ .

The loss function  $L$  compares the predicted labels  $\hat{\mathbf{y}}$  with the target labels  $\mathbf{y}$  through cross-entropy loss, where  $C$  is the number of classes in the classification task.

$$L = - \sum_i^C \mathbf{y}_i \cdot \log(\hat{\mathbf{y}}_i)$$

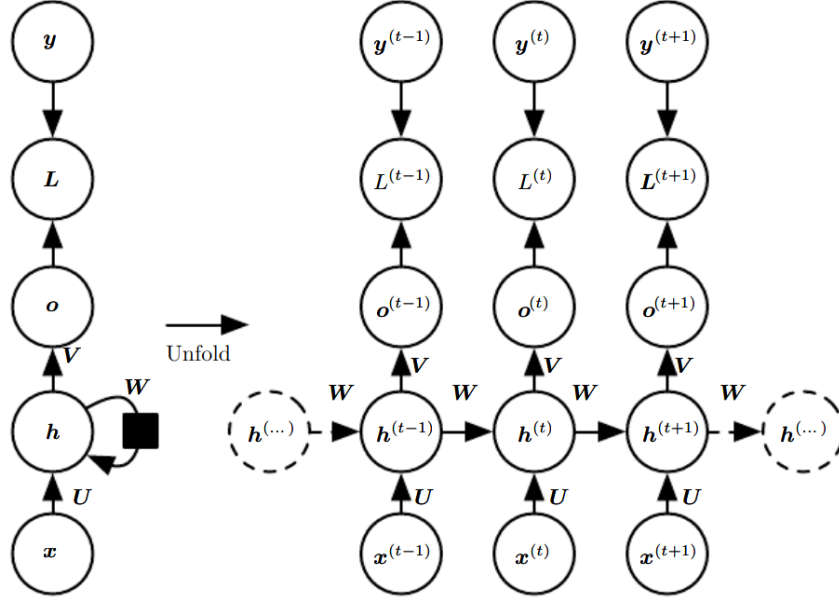


Figure 1: RNN

The remaining nodes in the graph are computed as shown in Equation 1.

$$\begin{aligned}
 \mathbf{a}^{(t)} &= \mathbf{b} + \mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(t-1)} \\
 \mathbf{h}^{(t)} &= \tanh(\mathbf{a}^{(t)}) \\
 \mathbf{o}^{(t)} &= \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)} \\
 \hat{\mathbf{y}}^{(t)} &= \text{softmax}(\mathbf{o}^{(t)})
 \end{aligned} \tag{1}$$

$\mathbf{U}, \mathbf{W}, \mathbf{V}$  are the weight matrices, while  $\mathbf{b}, \mathbf{c}$  are biases. By choice of hyperparameters,  $\mathbf{h}^{(t)} \in \mathbb{R}^{128 \times 1}$ .

For an input sequence,  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ , we first compute the hidden units  $\mathbf{h}^{(t)}$  as shown in Equation 2. At  $t = 1$ , the hidden unit is computed using only  $\mathbf{x}^{(1)}$ . Once we have the value of  $\mathbf{h}^{(t)}$ , we can then compute its predicted state  $\hat{\mathbf{y}}^{(t)}$  with Equation 1.

$$\begin{aligned}
 \mathbf{h}^{(1)} &= \tanh(\mathbf{b} + \mathbf{U}\mathbf{x}^{(1)}) \\
 \mathbf{h}^{(t)} &= \tanh(\mathbf{b} + \mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(t-1)}), \forall t \in \{2, \dots, n\}
 \end{aligned} \tag{2}$$

### 5.3.2 Backward

To perform backpropagation, our objective is to obtain the partial differential of  $L$  with respect to the weights and biases. The gradients are taken to be the **sum** over all time steps  $t \in \{1, \dots, n\}$ .

#### Output layer

$$\begin{aligned}
 \frac{\partial L}{\partial \mathbf{o}^{(t)}} &= \text{softmax}(\mathbf{o}^{(t)}) - \mathbf{y}^{(t)} \\
 &= \hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}
 \end{aligned} \tag{3}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{c}} &= \sum_{t=1}^n \frac{\partial L}{\partial \mathbf{o}^{(t)}} \frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{c}} \\
&= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}) \frac{\partial(\mathbf{c} + \mathbf{V}\mathbf{h}^{(t)})}{\partial \mathbf{c}} \\
&= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)})
\end{aligned} \tag{4}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{V}} &= \sum_{t=1}^n \frac{\partial L}{\partial \mathbf{o}^{(t)}} \frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{V}} \\
&= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}) \frac{\partial(\mathbf{c} + \mathbf{V}\mathbf{h}^{(t)})}{\partial \mathbf{V}} \\
&= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}) \mathbf{h}^{(t)T}
\end{aligned} \tag{5}$$

### Hidden layer

To compute the gradients of the hidden units  $\frac{\partial L}{\partial \mathbf{h}^{(t)}}$ , we initialize the gradient at  $t = n$ , since it has no subsequent time step, and its gradient is dependent only on  $\mathbf{o}^{(n)}$ .

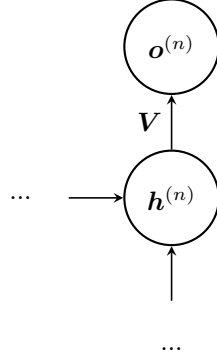


Figure 2: Computational graph showing gradients for  $\mathbf{h}^{(n)}$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{h}^{(n)}} &= \frac{\partial L}{\partial \mathbf{o}^{(n)}} \frac{\partial \mathbf{o}^{(n)}}{\partial \mathbf{h}^{(n)}} \\
&= (\hat{\mathbf{y}}^{(n)} - \mathbf{y}^{(n)}) \frac{\partial(\mathbf{c} + \mathbf{V}\mathbf{h}^{(n)})}{\partial \mathbf{h}^{(n)}} \\
&= \mathbf{V}^T (\hat{\mathbf{y}}^{(n)} - \mathbf{y}^{(n)})
\end{aligned} \tag{6}$$

Now, we can formulate the rest of the gradients recursively.

$$\frac{\partial L}{\partial \mathbf{h}^{(t)}} = \frac{\partial L}{\partial \mathbf{o}^{(t)}} \frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{h}^{(t)}} + \frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{h}^{(t)}}$$

From the result in Equation 6, we have that

$$\frac{\partial L}{\partial \mathbf{o}^{(t)}} \frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{h}^{(t)}} = \mathbf{V}^T (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)})$$

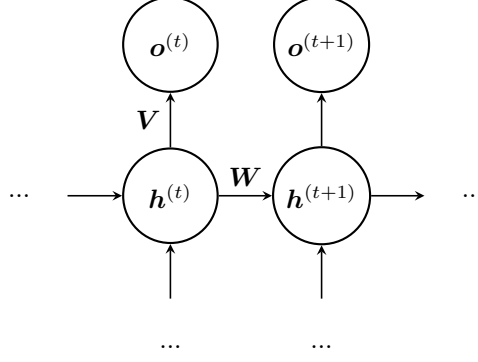


Figure 3: Computational graph showing gradients for  $\mathbf{h}^{(t)}$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{h}^{(t)}} &= \frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{a}^{(t+1)}} \frac{\partial \mathbf{a}^{(t+1)}}{\partial \mathbf{h}^{(t)}} \\
&= \frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial (\tanh(\mathbf{a}^{(t+1)}))}{\partial \mathbf{a}^{(t+1)}} \frac{\partial (\mathbf{b} + \mathbf{U}\mathbf{x}^{(t+1)} + \mathbf{W}\mathbf{h}^{(t)})}{\partial \mathbf{h}^{(t)}} \\
&= \mathbf{W}^T \frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial (\tanh(\mathbf{a}^{(t+1)}))}{\partial \mathbf{a}^{(t+1)}} \\
&= \mathbf{W}^T \text{diag}(1 - (\mathbf{h}^{(t+1)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t+1)}}
\end{aligned}$$

Combining the above results, we obtain

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{h}^{(t)}} &= \frac{\partial L}{\partial \mathbf{o}^{(t)}} \frac{\partial \mathbf{o}^{(t)}}{\partial \mathbf{h}^{(t)}} + \frac{\partial L}{\partial \mathbf{h}^{(t+1)}} \frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{h}^{(t)}} \\
&= \mathbf{V}^T (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}) + \mathbf{W}^T \text{diag}(1 - (\mathbf{h}^{(t+1)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t+1)}}
\end{aligned} \tag{7}$$

It is next imperative to show that the Jacobian  $J^{(t+1)} = \frac{\partial \mathbf{h}^{(t+1)}}{\partial \mathbf{a}^{(t+1)}} = \frac{\partial (\tanh(\mathbf{a}^{(t+1)}))}{\partial \mathbf{a}^{(t+1)}} = \text{diag}(1 - (\mathbf{h}^{(t+1)})^2)$ .

We begin with the knowledge that  $\frac{d}{dx} \tanh(x) = 1 - \tanh^2(x)$ .

$$J_{ij}^{(t+1)} = \frac{\partial \mathbf{h}_i^{(t+1)}}{\partial \mathbf{a}_j^{(t+1)}}$$

For  $i \neq j$ ,  $J_{ij}^{(t)} = 0$ .

For  $i = j$ ,

$$\begin{aligned}
J_{ii}^{(t+1)} &= \frac{\partial \mathbf{h}_i^{(t+1)}}{\partial \mathbf{a}_i^{(t+1)}} = \frac{\partial (\tanh(\mathbf{a}^{(t+1)}))_i}{\partial \mathbf{a}_i^{(t+1)}} \\
&= 1 - \tanh^2(\mathbf{a}^{(t+1)})_i \\
&= 1 - (\mathbf{h}_i^{(t+1)})^2
\end{aligned}$$

$$\therefore J^{(t)} = \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{a}^{(t)}} = \begin{pmatrix} 1 - (\mathbf{h}_1^{(t)})^2 & 0 & \dots & 0 \\ 0 & 1 - (\mathbf{h}_2^{(t)})^2 & \dots & 0 \\ \dots & \dots & 1 - (\mathbf{h}_i^{(t)})^2 & \dots \\ 0 & 0 & \dots & 1 - (\mathbf{h}_n^{(t)})^2 \end{pmatrix} = \text{diag}(1 - (\mathbf{h}^{(t)})^2)$$

We can then use this result to obtain the gradients with respect to  $\mathbf{b}, \mathbf{W}, \mathbf{U}$ .

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{b}} &= \sum_{t=1}^n \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \mathbf{b}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial (\mathbf{b} + \mathbf{U} \mathbf{x}^{(t)} + \mathbf{W} \mathbf{h}^{(t-1)})}{\partial \mathbf{b}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{W}} &= \sum_{t=1}^n \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \mathbf{W}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial (\mathbf{b} + \mathbf{U} \mathbf{x}^{(t)} + \mathbf{W} \mathbf{h}^{(t-1)})}{\partial \mathbf{W}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \mathbf{h}^{(t-1)^T}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{U}} &= \sum_{t=1}^n \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{a}^{(t)}} \frac{\partial \mathbf{a}^{(t)}}{\partial \mathbf{U}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \frac{\partial (\mathbf{b} + \mathbf{U} \mathbf{x}^{(t)} + \mathbf{W} \mathbf{h}^{(t-1)})}{\partial \mathbf{U}} \\
&= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \mathbf{x}^{(t)^T}
\end{aligned} \tag{10}$$

In summary,

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{U}} &= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \mathbf{x}^{(t)^T} \\
\frac{\partial L}{\partial \mathbf{W}} &= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \mathbf{h}^{(t-1)^T} \\
\frac{\partial L}{\partial \mathbf{b}} &= \sum_{t=1}^n \text{diag}(1 - (\mathbf{h}^{(t)})^2) \frac{\partial L}{\partial \mathbf{h}^{(t)}} \\
\frac{\partial L}{\partial \mathbf{V}} &= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)}) \mathbf{h}^{(t)^T} \\
\frac{\partial L}{\partial \mathbf{c}} &= \sum_{t=1}^n (\hat{\mathbf{y}}^{(t)} - \mathbf{y}^{(t)})
\end{aligned} \tag{11}$$

### 5.3.3 Training

### 5.3.4 Evaluation

### 5.3.5 Finetuning

### 5.3.6 Summary