Mining Unstructured Data
Review - Exercises

1 Dependency parsing (Exercise 2 in the list)

In a global linear model for dependency parsing, the feacture vector $f(x,y)$ for any sentence x paired with a dependency tree y is defined as:

$$f(x,y) = \sum_{(h,m)\in y} \mathbf{f}(x,h,m)$$

where $\mathbf{f}(x,h,m)$ is a function that maps a dependency (h,m) and a sentence x to a local feature vector.

We want the vector $f(x,y)$ to have exactly two dimensions, each dimension having the following value:

$$\begin{aligned} f_1(x,y) &= \text{num of times a dependency with head } \textit{car} \text{ and modifier } \textit{the} \text{ is seen in } (x,y) \\ f_2(x,y) &= \text{num of times a dependency with head part-of-speech NN, modifier part-of-speech DT,} \\ &\quad \text{and no adjective (JJ) between the DT and the NN is seen in } (x,y) \end{aligned}$$

Assuming that each element in the sentence x_i is a pair $(word, PoS)$, and that the functions $word(x_i)$ and $pos(x_i)$ return the value for each component of the pair:

1. Give a definition of the function $\mathbf{f}(x,h,m) = \langle \mathbf{f}_1(x,h,m), \mathbf{f}_2(x,h,m) \rangle$ that leads to the above definition of $f(x,y)$.
2. Compute the value of $f(x,y)$ for the following pair (x,y):

   $x =$ *The/DT car/NN with/IN the/DT red/JJ hood/NN won/VBD the/DT car/NN race/NN*
   $y = \{(2,1),(7,2),(2,3),(3,6),(6,4),(6,5),(0,7),(7,10),(10,8),(10,9)\}$

2 Constituent parsing (Exercise 6 in the list)

Given the following PCFG:

S → NP VP	1.0	N → time	0.4
NP → N N	0.25	N → flies	0.2
NP → D N	0.4	N → arrow	0.4
NP → N	0.35	D → an	1.0
VP → V NP	0.6	ADV → like	1.0
VP → V ADV NP	0.4	V → flies	0.5
		V → like	0.5

and the sentence *time flies like an arrow*

1. Write two parse trees that this grammar generates for this sentence
2. Compute the probability of each tree.
3. Convert the grammar to CNF and emulate the behaviour of the CKY algorithm on this sentence. Provide the final chart with all the information involved.

3 Word sequences (Exercise 4 in the list)

We are performing PoS tagging with a trigram-factored CRF, using tagset $\mathcal{T} = \{\texttt{DT}, \texttt{V}, \texttt{NN}, \texttt{ADV}, \texttt{PREP}\}$, and we defined a history as $h = \langle t_{i-2}, t_{i-1}, w_{[1:n]}, i \rangle$.

1. How many possible histories are there for a given input sequence $\mathcal{X}$ and a fixed value of i?
2. Which of the following are valid features?

$$\mathbf{f}_1(h,t) = \begin{cases} 1 & \text{if } t = \texttt{V} \text{ and } t_{i-1} = \texttt{PREP} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h,t) = \begin{cases} 1 & \text{if } t = \texttt{V} \text{ and } w_{i-2} = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h,t) = \begin{cases} 1 & \text{if } t = \texttt{V} \text{ and } t_{i-3} = \texttt{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_4(h,t) = \begin{cases} 1 & \text{if } t = \texttt{V} \text{ and } t_{i+1} = \texttt{PREP} \text{ and } w_2 = \texttt{cow} \\ 0 & \text{otherwise} \end{cases}$$

3. Compute the global feature vector $\mathbf{f}(\mathcal{X}, \mathcal{Y})$ for the input sequence is $\mathcal{X} =$`the dog walked to a park` and the tag sequence $\mathcal{Y} =$ `DT NN V PREP DT NN`, when using the following features:

$$\mathbf{f}_1(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } w_i = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } t_{i-1} = \texttt{DT} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } t_{i-1} = \texttt{DT} \text{ and } w_{i-1} = \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

4. Given the history $h = (t_{i-2}, t_{i-1}, w_{[1:n]}, 5) = (\texttt{V}, \texttt{DT}, \texttt{the man saw the dog in the park}, 5)$, which of the following features yield $\mathbf{f}(h, \texttt{NN}) = 1$?

$$\mathbf{f}_1(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } w_i = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(h,t) = \begin{cases} 1 & \text{if } t = \texttt{DT} \text{ and } w_i = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_3(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } w_{i+1} = \texttt{dog} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_4(h,t) = \begin{cases} 1 & \text{if } t = \texttt{NN} \text{ and } t_{i-1} = \texttt{DT} \\ 0 & \text{otherwise} \end{cases}$$

4 PoS tagging

Given the following sequences of pairs (y_i, x_i):

(D, the) $N, wine)$ $(V, ages)$ $(A, alone)$ $(FF, .)$

(D, the) $N, wine)$ $(N, waits)$ $(V, last)$ $(N, ages)$ $(FF, .)$

$(D, some)$ $(N flies)$ $(V, dove)$ $(P, into)$ (D, the) $(N, wine)$ $(FF, .)$

(D, the) $(N, dove)$ $(V, flies)$ (P, for) $(D, some)$ $(N, flies)$ $(FF, .)$

(D, the) $(A, last)$ $(N, dove)$ $(V, waits)$ $(A, alone)$ $(FF, .)$

1. Draw the graph of the bigram HMM and list all the non-zero parameters that we can achieve by maximum likelihood estimation from the data.
2. Compute the probability of sequence $y_{1:n} = (D, N, V, P, D, A, N, FF)$ given the input sequence $x_{1:n} = (the, dove, waits, for, some, last, wine, .)$