Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

5. Exercise for "Sprachverarbeitung und Text Mining"

03.12.2021

1 Knowledge Questions

1. How do Maximum-Entropy Markov Model (MEMM) and Hidden Markov Model (HMM) differ?

   The Maximum Entropy Markov Model (MEMM) is a combination of the Hidden Markov Model (HMM) and the Maximum Entropy Model. HMM is based on probabilities of the form $p(tag|tag)$ and $p(word|tag)$, requiring all information to be encoded in these probabilities. MEMM uses features to encode information, making it easier to encode additional information directly and allowing the use of different patterns, e.g. taking the following word into consideration in POS-tagging.

2. What is the meaning of the parameters λ in the Maximum Entropy Model?

   Feature scores λ indicate how well a feature conforms to the data.

3. Describe the **label bias** problem in MEMM!

   MEMMs are normalized locally over each observation and therefore suffer from the label bias problem, where transitions originating from one state only compete against each other, as opposed to all other transitions in the model.

   Name a solution presented in the lecture to avoid this!

The problem can be solved with Conditional Random Fields (CRF).

4. How do Maximum Entropy Markov Model (MEMM) and Conditional Random Field (CRF) differ?

CRFs are not based on the assumption of independence (of labels). In MEMMs, local normalization of probabilities is realized, which leads to "label bias", while CRFs realize global normalization. CRF is a global model.

5. What does the **order** of a MEMM state?

The order of a MEMM indicates how many time steps it can look back.

2 MEMM and CRF Models

1. **Number of terms:** You want to calculate the probability of a certain POS-tag sequence (e.g., ART NN ADV) for a sentence by using both a MEMM and a CRF. For both MEMM and CRF, calculate the number of terms that are summed over in the denominator and numerator during probability calculation when the following constraints apply:

   a) The length of the sequence is 4 and there are 4 possible tags per word. For every timestep there are two active features: one feature from the template $f(x, y_i)$ and one feature from the template $f(x, y_i, y_{i-1})$.

      i. MEMM:
      at each local decision(timestep):
      numerator: 2 terms as 2 features are active.
      denominator: 4 tags x 2 active features $= 8$ terms.

      ii. CRF:
      Sum over 8 terms in numerator and sum over 2048 terms in denominator
      numerator: 4 words x 2 active features = 8 terms
      denominator: 4^4, 4 words ^{4} tags (all possible sequences) x 8 (nr. terms of numerator) = 2048 terms

b) The length of the sequence is 5 and there are 5 possible tags per word. For every timestep there are two active features: one feature from the template $f(x, y_i)$ and one feature from the template $f(x, y_i, y_{i-1})$. Furthermore there are two more active features on a single timestep of the templates $f(x, y_i)$ and $f(x, y_i, y_{i-1})$.

i. MEMM:
at each local decision (timestep) except one:
numerator: 2 terms as 2 features are active.
denominator: 5 tags x 2 active features $= 10$ terms.
at one local decision (timestep):
numerator: 4 terms as 4 features are active.
denominator: 5 tags x 4 active features $= 20$ terms.

ii. CRF:
Sum over 12 terms in the numerator and sum over 37500 terms in the denominator
numerator: 5 words x 2 active features (all timesteps) + 2 active features (at an arbitrary time step) = 12 terms
denominator: 5^5, 5 words and ^{5} tags x 12 (nr. terms of numerator) = 37500 terms

2. **Calculation:**

   In the following λ tables, mark the values that are relevant in a MEMM and CRF to calculate the numerators when the probability of the tag sequence

   PP ADV ADV ADV NN

   for the sentence "Wir rennen oft zum Bus" (En:"We often run to the bus") is to be calculated.

   Notes: The transition tables are obviously incomplete. For space reasons, tables that are not relevant have been omitted. Further, specific transition scores have simply been copied since they are not relevant for the task of marking the relevant entries.

Feature	PP	VVFIN	ADV	ART	NN
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

Feature	<S> $\rightarrow$ PP	PP $\rightarrow$ PP	VVFIN $\rightarrow$ PP	ADV $\rightarrow$ PP	ART $\rightarrow$ PP
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

Feature	<S> → ADV	PP → ADV	VVFIN → ADV	ADV → ADV	ART → ADV
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

Feature	<S> → NN	PP → NN	VVFIN → NN	ADV → NN	ART → NN
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

Feature	PP	VVFIN	ADV	ART	NN
$x_i = Wir$	**0.9**	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	**0.05**	-0.1	0.4
$x_i = oft$	0.1	0.05	**0.9**	0.15	0.01
$x_i = zum$	0.01	0.05	**0.2**	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	**0.9**
$x_{i-1} = Wir$	0.01	0.99	**0.1**	0.2	0.01

Feature	<S> → PP	PP → PP	VVFIN → PP	ADV → PP	ART → PP
$x_i = Wir$	**0.9**	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

Feature	<S> → ADV	PP → ADV	VVFIN → ADV	ADV → ADV	ART → ADV
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	**0.7**	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	**0.15**	0.01
$x_i = zum$	0.01	0.05	0.2	**0.8**	0.15
$x_i = Bus$	0.1	0.02	0.1	0.2	0.9
$x_{i-1} = Wir$	0.01	**0.99**	0.1	0.2	0.01

Feature	<S> → NN	PP → NN	VVFIN → NN	ADV → NN	ART → NN
$x_i = Wir$	0.9	-0.2	0.1	0.1	0.2
$x_i = rennen$	0.1	0.7	0.05	-0.1	0.4
$x_i = oft$	0.1	0.05	0.9	0.15	0.01
$x_i = zum$	0.01	0.05	0.2	0.8	0.15
$x_i = Bus$	0.1	0.02	0.1	**0.2**	0.9
$x_{i-1} = Wir$	0.01	0.99	0.1	0.2	0.01

In the given example, MEMM and CRF use the same feature functions, so the relevant values for the calculation of the numerators are the same.

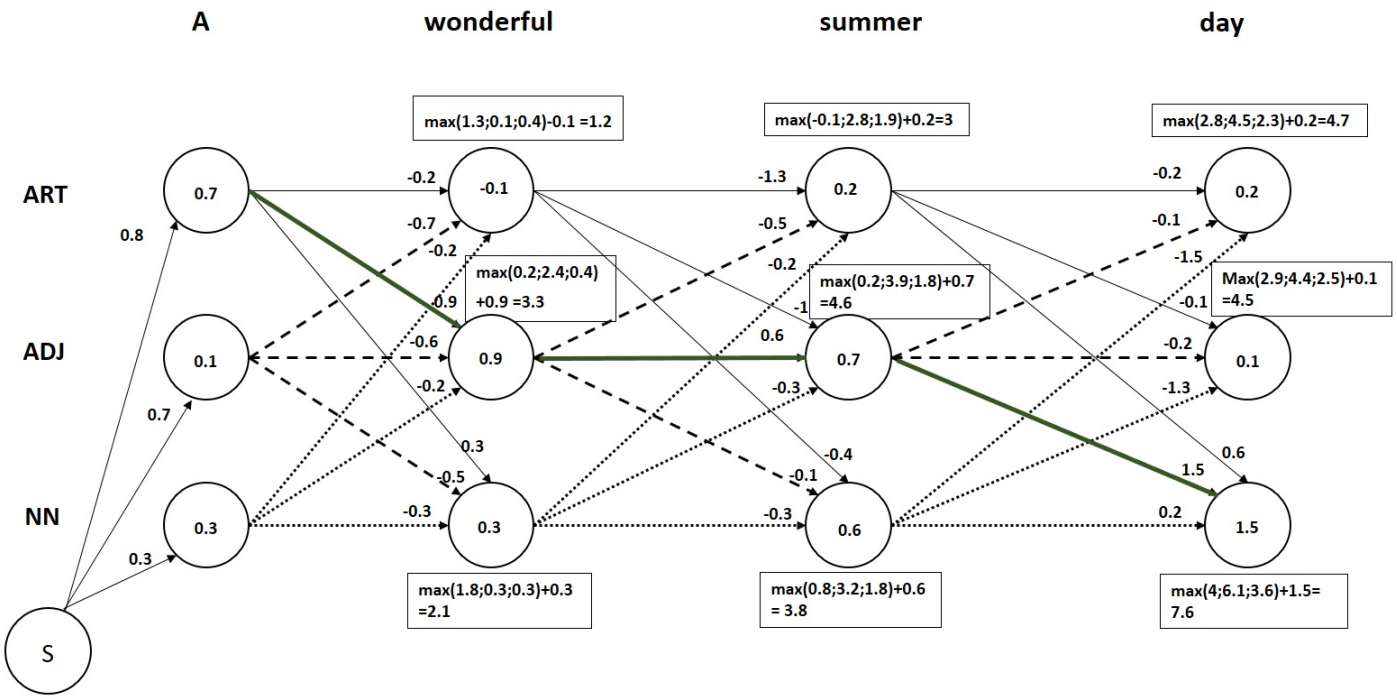
3. **Viterbi:** Given is a CRF with the following λ tables with feature scores. Find the most likely sequence of tags for the sentence:
„`A wonderful summer day`“
Apply the Viterbi algorithm!

Feature	ART	ADJ	NN
$CurrentWord = A$	0.7	0.1	0.3
$CurrentWord = wonderful$	-0.1	0.9	0.3
$CurrentWord = summer$	0.2	0.7	0.6
$CurrentWord = day$	0.2	0.1	1.5

Feature	ART → ART	ADJ → ART	NN → ART	<S> → ART
$CurrentWord = A$	-0.1	0.1	0.1	0.8
$CurrentWord = wonderful$	-0.2	-0.7	-0.2	0.2
$CurrentWord = summer$	-1.3	-0.5	-0.2	-0.5
$CurrentWord = day$	-0.2	-0.1	-1.5	-1

Feature	ART → ADJ	ADJ → ADJ	NN → ADJ	<S> → ADJ
$CurrentWord = A$	-0.3	0.1	-0.2	0.7
$CurrentWord = wonderful$	0.9	-0.6	-0.2	0.1
$CurrentWord = summer$	-1	0.6	-0.3	-0.5
$CurrentWord = day$	-0.1	-0.2	-1.3	-1

Feature	ART → NN	ADJ → NN	NN → NN	<S> → NN
$CurrentWord = A$	-0.1	0.25	-0.45	0.3
$CurrentWord = wonderful$	0.3	-0.5	-0.3	-0.1
$CurrentWord = summer$	-0.4	-0.1	-0.3	-0.5
$CurrentWord = day$	0.6	1.5	0.2	0.3


A
wonderful
summer
day
ART
ADJ
NN
S
0.8
0.7
0.3
0.7
0.1
0.3
-0.2
-0.7
-0.2
0.9
-0.6
-0.2
0.3
-0.5
-0.3
-1.3
-0.5
-0.2
-1
0.6
-0.3
-0.4
-0.1
-0.3
-0.2
-0.1
-1.5
-0.1
-0.2
-1.3
0.6
1.5
0.2
-0.1
0.2
0.9
0.3
0.7
0.2
0.1
1.5
max(1.3;0.1;0.4)-0.1 =1.2
max(0.2;2.4;0.4) +0.9 =3.3
max(1.8;0.3;0.3)+0.3 =2.1
max(-0.1;2.8;1.9)+0.2=3
max(0.2;3.9;1.8)+0.7 =4.6
max(0.8;3.2;1.8)+0.6 = 3.8
max(2.8;4.5;2.3)+0.2=4.7
Max(2.9;4.4;2.5)+0.1 =4.5
max(4;6.1;3.6)+1.5= 7.6