

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

## 8. Exercise for “Sprachverarbeitung und Text Mining”

14.01.2022

### 1 Graph-based Dependency Parsing

1. How do graph-based dependency parsing approaches differ from shift-reduce-parsing approaches in terms of machine learning?
2. Given is the following graph in tabular form that contains the nodes A through D. The table entries describe the edges from one node to the other and (if present) their weights.

Table 1: Graph: f = from, t = to

<b>f \ t</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>A</b>	-	1	5	7
<b>B</b>	-	-	9	2
<b>C</b>	-	9	-	6
<b>D</b>	-	3	7	-

Calculate the maximum spanning tree of the graph using the CLE algorithm.

- Draw the graph in its original form, mark the initially selected edges.
- For each contracting step, draw the contracted graph, and label the edges with the resulting negative weights. Mark the newly selected edge in the contracted graph.
- Draw the original graph with the newly selected edge instead of the original edge.

For the solution of the task, 3 drawings are required.

## 2 Named Entity Recognition

1. In the lecture you have learned about BIO and IO encoding for labeling data sets in named entity recognition. Describe a reason for choosing BIO encoding over IO encoding.

2. Define the terms Entity Recognition and Relation Detection in your own words.

Which two subtasks can Relation Extraction be divided into?

3. Give the following text, write down rules to find all marked Named Entities. Marked entities represent persons(green), locations(blue), and titles(red). You don't need to use a specific syntax. You may assume that the text is fully pre processed. This means you can use POS tags, parse trees, lemmata, etc. as features in your rules.

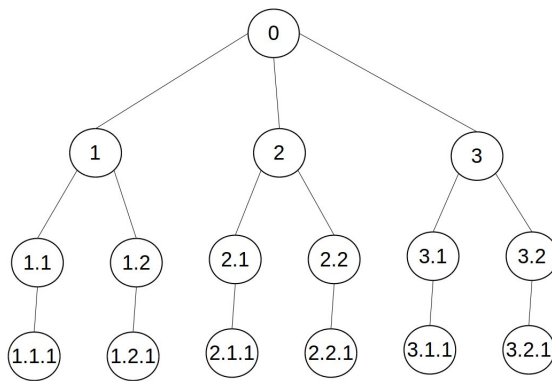
An example rule could look as follows:

NN replied to NN → both NNs are names

`gimli` and `legolas` are walking to `mordor`. »my father `gloin`, the ruler of `erebor`, really enjoyed the last season of '`game of thrones`'«, `gimli` grummeled. » i also enjoyed the part, where `dumbledore` revealed `luke skywalker` to be his father.«, `legolas` replied to `gimli`. they are joined by `aragorn`. »who do you think you are, `aragorn`?«, asks `gimli`. »i was sent to you by `harrys` owl.«, he replied.

### 3 Non-Structured Hierarchical Classification

1. Explain in your own words the difference between flat and hierarchical classification.
2. What is the potential requirement for classifiers in non-structured hierarchical classification when using a local classifier per parent node?
3. Given are the following hierarchy of classifiers and the following training dataset (with a local classifier in each node):



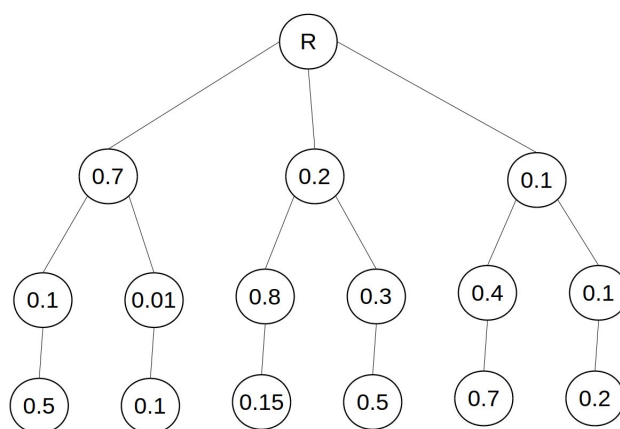
Nr.	Features	Label	S1	S2	S3
1	...	2.2			
2	...	2.1			
3	...	3.2			
4	...	1.2			
5	...	1.1			
6	...	2.2.1			
7	...	2.1.1			
8	...	1.1.1			
9	...	3.1			
10	...	2			
11	...	3			
12	...	1			
13	...	3.2.1			

Consider the classifier 2.2 of the hierarchy. For the following training strategies (S1, S2, S3), indicate which training examples will be used as positive or negative examples, or not used at all.

- S1: Exclusive Policy
- S2: Less-Inclusive Policy
- S3: Siblings Policy

Complete the table by using check marks for positive, crosses for negative, and horizontal lines for non-observed examples.

4. For a given word, the class indicators in the following hierarchy generate the values that are written inside the nodes.



Describe possible problems in the choice of label nodes for the following classification strategies:

- Global Maximum
- Maximum of each layer (top down)

What could an improved strategy look like?

## 4 Coreference Resolution

1. Define the goal of *Coreference Resolution* in your own words.
2. How would you generate training data for the binary classifier in the tournament model out of labeled (coref, non-coref) mention pairs  $(m_i, m_j)$ ?
3. In the most primitive form of this model, there will always be a returned winner. Why is this a problem and how could you solve it?
4. The most primitive tournament type is to let every candidate fight against every other. Which type of tournament would be much more efficient?
5. Calculate all training instances for coreference resolution using the method of Soon et al. using the following sentence:

$[John]_1^1$  told  $[Bill]_2^2$  that  $[his]_3^1$   $[mother]_4^3$  was angry.  
 $[She]_5^3$  wanted to know where  $[he]_6^1$  had been.

The MentionID is subscript and the EntityID is superscript.