Research Group Social Computing
Department of Informatics
Technical University of Munich

Eexam
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Natural Language Processing

Exam:	IN2361 / Graded Electronic Exercise	**Date:**	Wednesday 1st July, 2020
Examiner:	Georg Groh	**Time:**	13:45 – 14:45

	P 1	P 2	P 3	P 4	P 5	P 6	P 7
I							

Working instructions

- This exam consists of **10 pages** with a total of **7 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 70 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - please see the latest version of the GEE fact sheet ("fact sheet of the graded electronic exercise july 1st" in section "exam") on the Moodle page of IN2361 (content also communicated to you via email
- IMPORTANT: **do not write that much in the answer boxes that the scrollbar appears**. Only what is finally visible after editing that box is done can be graded. Reason: TUM-Exam uses an IMAGE-BASED processing chain. If your edits "seem gone" after finishing edit for a box (this seems to sometimes happen with MacOs and Preview: try "export as PDF" and use a different file-name.
- Working Period: 13:45-14:45 (60 minutes), Submission Period: 13:45-15:00 (60+15 minutes), Upload Period: 15:00-15:30 (an extra 30 minutes
- Please ignore the "Left room from / to" and "Early submission at" box below

Left room from __________ to __________ / Early submission at __________

Problem 1 Language Models (10 credits)

0

1.1 (0 P) Just to be sure: Write your first (given) name, your last (family) name, and your matriculation number (just as a sanity check).

0
1
2
3
4
5
6
7
8

1.2 (8 P) A person claims: „modern neural language models trained on large text corpora really are a manifestation of true artifical intelligence“. As an NLP expert, state two arguments supporting that claim and two arguments against that claim! Do NOT state MORE THAN TWO arguments each!!!

0
1
2

1.3 (2 P) Provide the mathematical expression for a a tri-gram approximation for $P(w_1 w_2 w_3 w_4 w_5)$

Problem 2 Simple sentiment analysis with Naïve Bayes classifiers (10 credits)

2.1 (2 P) MAP solution for smoothing: why do we propose a Dirichlet prior $p(\theta_c|\alpha)$ in the expression for the posterior $p(\theta_c|D)$?

$$
\begin{aligned}
p(\theta_c|D) &\propto p(D|\theta_c)p(\theta_c|\alpha) && (1)\\
&\propto \prod_{v=1}^{V} \theta_{vc}^{N_{vc}}\theta_{vc}^{\alpha_v-1} && (2)\\
&\propto \prod_{v=1}^{V} \theta_{vc}^{N_{vc}+\alpha_v-1} && (3)\\
&= Dir(\theta|(\alpha_1 + N_{1c}, \alpha_2 + N_{2c}, \dots, \alpha_V + N_{Vc}) && (4)
\end{aligned}
$$

2.2 (2 P) While the simple MLE solution is $\theta_{vc}^{MLE} = \frac{N_{vc}}{N_c}$, the MAP solution using a Dirichlet prior $p(\theta_c|\alpha)$ is $\theta_{vc}^{MAP} = \frac{N_{vc}+\alpha_v-1}{N_c+(\sum_{v=1}^{V}\alpha_v)-V}$. What are the values for α_v that we have to choose to get the Lapalace smoothing and why?

2.3 (1 P) Why is the Naive Assumption called "naive"?

$$
p(\mathcal{D}|\Theta) = \prod_{n=1}^{N}\prod_{c=1}^{C}\prod_{v=1}^{V} p(x_v^{(n)}|\theta_{vc})^{\mathbb{1}(y^{(n)}=c)} \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y^{(n)}=c')} \quad (5)
$$

2.4 (1 P) What is the meaning of the parameters π_c?

0
1
2
3
4

2.5 (4 P) Using all of the modern arsenal of NLP: how would you solve the problem of sentiment classification of product reviews?

Problem 3 Constituency Parsing and Chunking (10 credits)

3.1 (5 P) The CYK algorithm requires the grammar to be in CNF which requires eliminating unit productions ($A \rightarrow B$) from the grammar. Augment the standard CYK algorithm so that it can also handle unit productions!

0 1 2 3 4 5

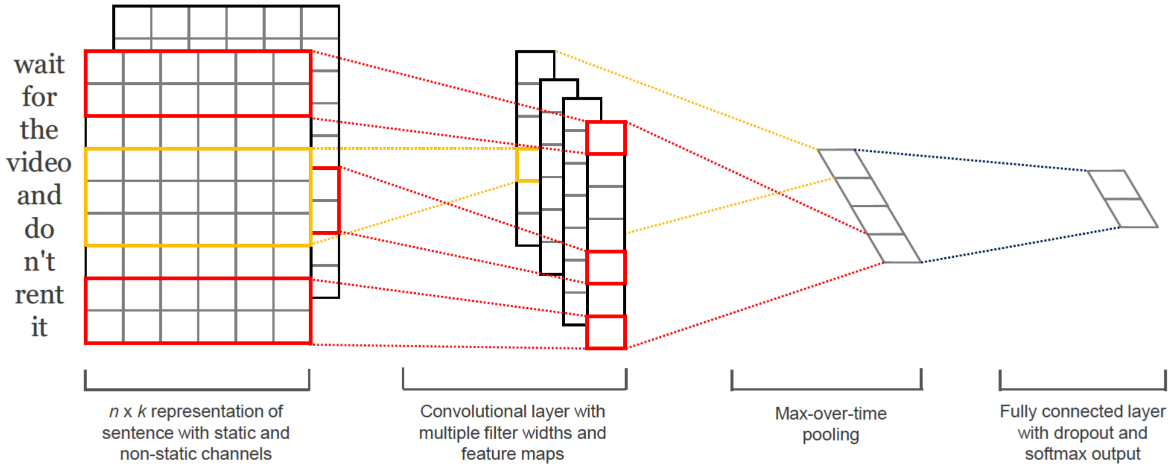
```
funcion CKY-Parse(words, grammar) returns table

for j <- from 1 to LENGTH(words) do
    for all {A | A -> words[j] ∈ grammar}
        table[j-1, j] <- table[j-1, j] ∪ A
    for i <- from j-2 downto 0 do
        for k <- i+1 to j-1 do
            for all {A | A -> BC} ∈ grammar and B ∈ table[i,k] and C ∈ table[k, j]}
                table[i, j] <- table[i, j] ∪ A
        // handling unit production
```

3.2 (5 P) For a simple supervised chunking task we want to chunk noun phrases and verb phrases only. When using BIO tagging, what are the classes we require?

0 1 2 3 4 5

Problem 4 Convolutional NN for sentence classification tasks (paper Kim, Y. (2014). Convolutional neural networks for sentence classification) (10 credits)



4.1 (5 P) Explain the motivation for using two channels in the input!

4.2 (5 P) What is the motivation for using several feature maps in the convolutional layer?

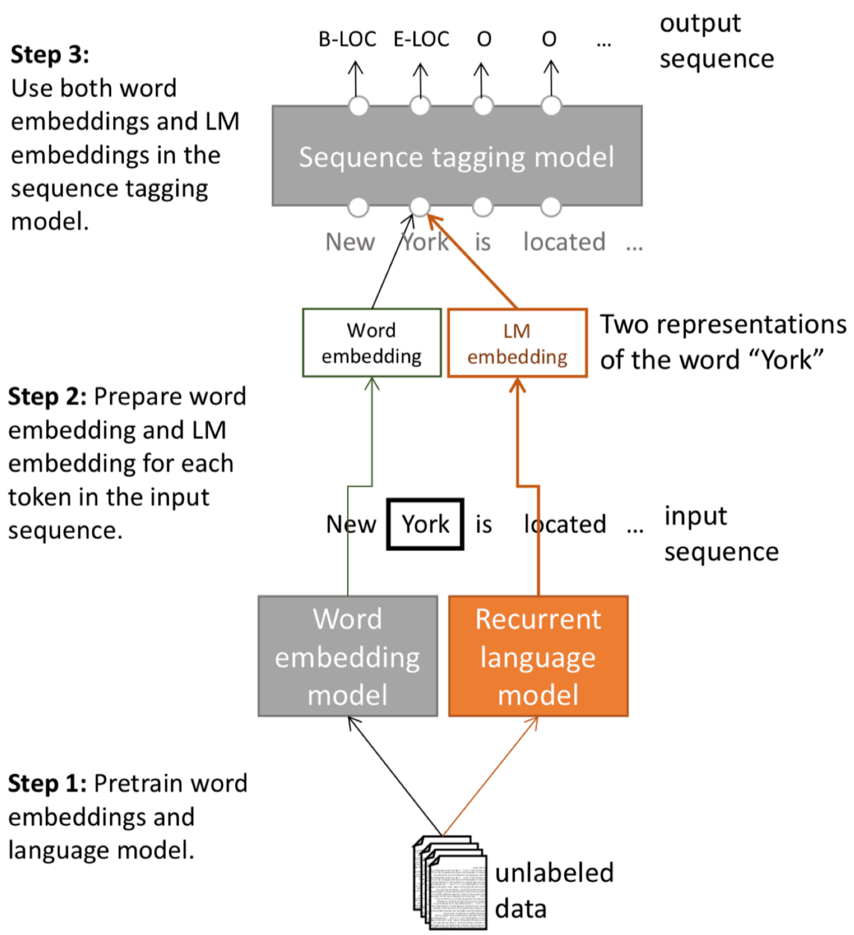
Problem 5 Modern contextual embeddings (10 credits)

5.1 (2 P) You want to use pretrained embeddings for some NLP task. What is the technical difference in the data that you get and the way you use it, technically, when (a) downloading pretrained GloVe or (b) downloading pretrained BERT?

0
1
2

5.2 (2 P) Peters at al (2017) "Semi-supervised sequence tagging with bidirectional language models" (Tag-LM paper, "pre-ELMo") (see image 5.2): Motivate why they are using the "LM embeddings" from the "Recurrent language model"!

0
1
2



0 1 2

5.3 (2 P) Key motivations to go from RNN-LM-based contextual embeddings (as in ELMo) to BERT: Why is bi-directionality desirable for sentence-level tasks?

0 1 2

5.4 (2 P) Key motivations to go from RNN-LM-based contextual embeddings (as in ELMo) to BERT: Why is bi-directionality a problem in RNN-LM-based approaches? How do you address that in BERT?

0 1 2

5.5 (2 P) Key motivations to go from RNN-LM-based contextual embeddings (as in ELMo) to BERT: If you would want to stick to the RNN-LM idea for contextual embeddings, how could you cope with the bi-directionality problem?

Problem 6 Transformer models (Vaswani et al: Attention is All You Need (2017) (10 credits)

6.1 (2 P) Additive attention may outperform dot-product attention for large dimension. What is a reason for that? How is the problem solved in the transformer model?

0 1 2

6.2 (2 P) The input width of a Transformer model is fixed. How do you handle shorter inputs?

0 1 2

6.3 (2 P) What is the motivation for the residual connections in the model?

0 1 2

6.4 (2 P) Why are trigonometric functions used for the positional encodings?

0 1 2

6.5 (2 P) How does the model ensure that each attention head attends in a different way?

0 1 2

Problem 7 **Alice and Bob (10 credits)**

0 1 2 3 4 5

7.1 (5 P) Alice and Bob are active on social media and they're friends on Facebook. Alice likes to write public text posts about various things on her Facebook account. She has been very active over the last years and has literally written thousands of posts. Bob likes anonymous type social media websites, where identities are not known. Recently he has stumbled upon an intriguing post there that he believes is written by Alice. Given only her Facebook text posts, how can Bob check his hypothesis with NLP?

0 1 2 3 4 5

7.2 (5 P) How can Alice use NLP techniques in order to solve the reverse problem and hide her identity from possible adversarial attacks like in part a)? More concretely, how can she construct a system that given her text post, the system would output another text post with the same meaning, but hides her identity and style of writing?