Research Group Social Computing
Department of Informatics
Technical University of Munich

Personal sticker

S5062

Compliance to the code of conduct
I hereby assure that I solve and submit this exam myself under my own name by only using the allowed tools listed below.

____
Signature or full name if no pen input available

Natural Language Processing

Exam: IN2361 / Graded Electronic Exercise

Date: Wednesday 24th February, 2021

Examiner: Georg Groh

Time: 17:00 – 18:00

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

Working instructions

- This exam consists of **12 pages** with a total of **6 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 60 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - please see the latest version of the GEE fact sheet on the Moodle page of IN2361
- 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: 17:00-18:00 (60 minutes), Submission Period: 17:00-18:15 (60+15 minutes), (additional) Upload Period (no more changes possible): 18:15-18:45 (an extra 30 minutes)
- Please ignore the "Left room from / to" and "Early submission at" box below

Left room from __________ to __________ / Early submission at __________

□ Exam empty

Problem 1 Regular Expressions, Heaps' Law, 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 □

1.2 (3 P) Describe / list the patterns that may be detected by the regular expression
`o+u?h|a+h+|hm+`
(1 P) What is the possible purpose of this RegEx?

0 □ 1 □ 2 □

1.3 (2 P) You fit Heaps' law $|V| = kN^{\beta}$ to two different documents and you get the following values:

1. $\beta = 0.99, k = 1$
2. $\beta = 0.71, k = 80$

Provide a reasonable suggestion of the nature of the documents!

0 □ 1 □ 2 □

1.4 (2 P) You fit Heaps' law $|V| = kN^{\beta}$ to the Java source code of the JDK class library. Provide a meaningful guess for the value of β you would get and give a short reasoning!

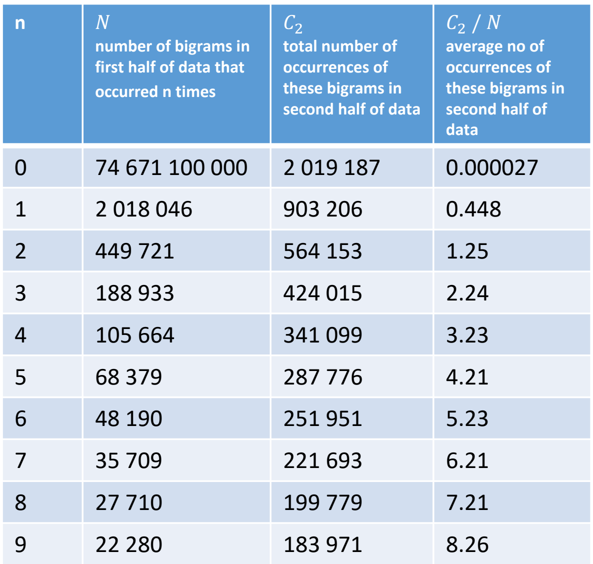
0 □ 1 □ 2 □

1.5 (2 P) What are the the mathematical consequences in terms of conditional independence when applying a tri-gram approximation for a language model for four word sequences $P(w_1 w_2 w_3 w_4)$ (using no beginning- or end-of-sequence markers)?

Problem 2 Language Models (ctd.), Naive Bayes vs. Logistic Regression, Embeddings (10 credits)

2.1 (3 P) Language models: Absolute Discounting: Church&Gale 1991: How would the second column of the table ("N number of bigrams in first half of data that occurred n times") look like for the *second* half of the data?

□ 0 □ 1 □ 2 □ 3



n	N number of bigrams in first half of data that occurred n times	C_2 total number of occurrences of these bigrams in second half of data	C_2 / N average no of occurrences of these bigrams in second half of data
0	74 671 100 000	2 019 187	0.000027
1	2 018 046	903 206	0.448
2	449 721	564 153	1.25
3	188 933	424 015	2.24
4	105 664	341 099	3.23
5	68 379	287 776	4.21
6	48 190	251 951	5.23
7	35 709	221 693	6.21
8	27 710	199 779	7.21
9	22 280	183 971	8.26

2.2 (3 P) Naive Bayes vs. Logistic Regression: How is $P(x, y|\theta)$ mathematically decomposed for a generative classifier, and how is it decomposed for a discriminative classifier? (In your answer, you can e.g. write θ as "theta").

□ 0 □ 1 □ 2 □ 3

2.3 (2 P) What is an advantage of Logistic Regression compared to Naive Bayes? State and explain *one* advantage (not more than one)!

□ 0 □ 1 □ 2

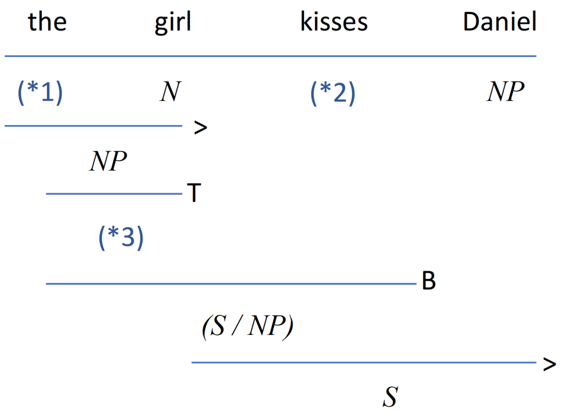
□ Page empty

0 □
1 □
2 □

2.4 (2 P) Your new PhD student suggests to train Word2Vec Skip-Gram embeddings with replacing the inner product based similarity $\mathbf{t} \cdot \mathbf{c}$ in $P(+|t, c) = 1/(1 + exp(-\mathbf{t} \cdot \mathbf{c}))$ with a similarity measure based on the Jensen-Shannon-Divergence between the vectors $\mathbf{t}/\sum_i t_i$ and $\mathbf{c}/\sum_i c_i$. Provide one reasonable counter-argument for doing that! (Not more than one)!

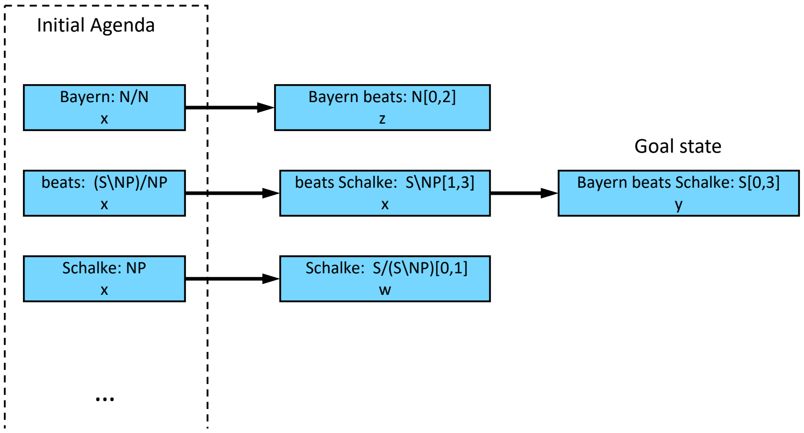
Problem 3 CCG parsing, Constituency Grammars (10 credits)

3.1 (3 P) In the following incomplete CCG parse, provide the missing three categories!

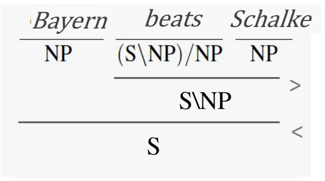


□ 0
□ 1
□ 2
□ 3

3.2 (4 P) CCG parsing with the A* algorithm: provide expressions for w, x, y and z in terms of the a_i, b_i and c_i, assuming that $a_1 < a_2 < a_3 < a_4$, $b_1 < b_2$ and $c_1 < c_2 < c_3$!



$-\log_{10} P$		
Bayern	beats	Schalke
N/N: a_1	(S\NP)/NP: b_1	NP: c_1
NP: a_2	N: b_2	N/N: c_2
S/S: a_3		S/(S\NP) c_3
S\S: a_4		



□ 0
□ 1
□ 2
□ 3
□ 4

□ 0
□ 1
□ 2
□ 3

3.3 (3 P) Explain the motivation for subcatagorization of verb-phrases, especially for training machine-learning-based constituency parsers?

Problem 4 GloVe embeddings, LSTM neurons (10 credits)

4.1 (3 P) Motivation for GloVe embeddings: Given the following table, sort the numbers a_1, a_2, a_3, a_4 in ascending order (e.g. $a_2 = a_4 < a_1 < a_2$)!

□ 0 □ 1 □ 2 □ 3

	$k = guitar$	$k = engine$	$k = wheel$	$k = saddle$
$P(k\|car)/P(k\|bike)$	a_1	a_2	a_3	a_4

4.2 (2 P) Motivation for GloVe embeddings: aside from symmetry or group-homomorphism considerations, motivate why setting $u_i^T v_k = logP(i|k)$ also makes intuitive sense!

□ 0 □ 1 □ 2

4.3 (3 P) Assuming you had never heard of contextual embeddings (such as BERT), how can static embeddings (GloVe, Word2Vec etc.) deal with different word senses? Why is that not really practical?

□ 0 □ 1 □ 2 □ 3

4.4 (2 P) LSTM neurons: why do we use the Hadamard product in connection with the gates and not the inner product or the outer product?

□ 0 □ 1 □ 2

Problem 5 Modern neural models (10 credits)

5.1 (2 P) Some systems use hybrid combinations of word-based approaches and character-based approaches for neural machine translation. Provide one pro-argument for these approaches and provide one counter-argument for these approaches! Do not provide more than one argument each!

□ 0 □ □ 1 □ □ 2

5.2 (2 P) What is the problem with just sticking to the standard word-based softmax architectures when vocabularies become large?

□ 0 □ □ 1 □ □ 2

5.3 (3 P) Another idea to deal with the problems associated with large vocabularies is combining standard word-based softmax architectures with Pointer Networks. What do pointer networks and basic attention have in common?

□ 0 □ □ 1 □ □ 2 □ □ 3

5.4 (3 P) Paper "Attention is all you need" (Vaswani et al, 2017): Multi-Head Attention is defined as

□ 0 □ □ 1 □ □ 2 □ □ 3

$$Multihead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \tag{5.1}$$

$$where \;\; head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5.2}$$

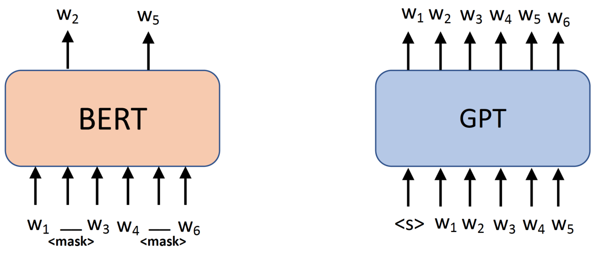
Someone states: "This is very similar to CNN approaches as in the paper Kim, Y. (2014) "Convolutional neural networks for sentence classification".

Provide one argument supporting the statement and one counter argument! Do not provide more than one argument each!

Problem 6 BERT and GPT (10 credits)

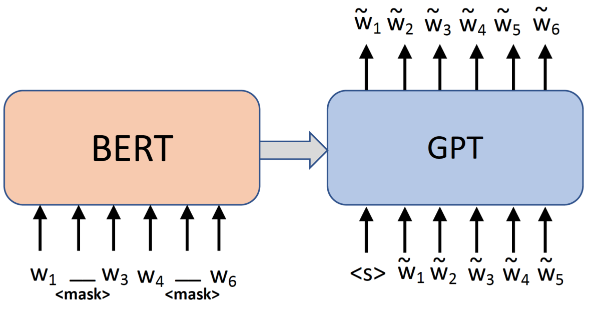
6.1 (3 P) Explain why BERT is not immediately usable for generation (e.g. as a language model)!

□ 0
□ 1
□ 2
□ 3


w_2 w_5
BERT
w_1 __ w_3 w_4 __ w_6
<mask> <mask>
w_1 w_2 w_3 w_4 w_5 w_6
GPT
<s> w_1 w_2 w_3 w_4 w_5

6.2 (3 P) Someone suggests using BERT as an encoder and GPT as a decoder in a seq-to-seq architecture. The decoder would attend to the encoder in a similar way as in the original Transformer model (Vaswani et al. 2017). Motivate the usefulness of this architecture for seq-to-seq tasks!

□ 0
□ 1
□ 2
□ 3


$\tilde{w}_1$ $\tilde{w}_2$ $\tilde{w}_3$ $\tilde{w}_4$ $\tilde{w}_5$ $\tilde{w}_6$
BERT
GPT
w_1 __ w_3 w_4 __ w_6
<mask> <mask>
<s> $\tilde{w}_1$ $\tilde{w}_2$ $\tilde{w}_3$ $\tilde{w}_4$ $\tilde{w}_5$

□ 0
□ 1
□ 2
□ 3
□ 4

6.3 (4 P) How could such a system as suggested in the previous assignment be trained and used for abstractive summarization?