Exam 2018-08-31

Marco Kuhlmann

This exam consists of three parts:

1. **Part A** consists of 5 items, each worth 3 points. These items test your understanding of the basic methods that are covered in the course. They require only compact answers, such as a short text, calculation, or diagram.

   Collected wildcards are valid for this part of the exam. The numbering of the questions corresponds to the numbering of the wildcards.

2. **Part B** consists of 3 items, each worth 6 points. These items test your understanding of the more advanced methods that are covered in the course. They require detailed and coherent answers with correct terminology.

3. **Part C** consists of 1 item worth 9 points. This item is an essay question that tests your understanding of the following article:

   Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.* Science 356(6334): 183–186, 2017.

Grade requirements 729G17: For grade G (Pass), you need at least 12 points in Part A. For grade VG (Pass with distinction), you additionally need at least 14 points in Part B, or at least 7 points in Part C.

Grade requirements TDP030: For grade 3, you need at least 12 points in Part A. For higher grades, you additionally need points in the other parts: for grade 4, at least 9 points in Part B; for grade 5, at least 14 points in Part B, or at least 7 points in Part C.

Note that surplus points in one part do not raise your score in another part.

Good luck!

Part A

01 Text classification **(3 points)**

a) The evaluation of a text classifier produced the following confusion matrix. The marked cell gives the number of times the system classified a document as class C whereas the gold-standard class for the document was A.

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Based on this confusion matrix, set up fractions for the following values:

i. precision with respect to class C ii. recall with respect to class B

b) Complete the decision rule for the Naive Bayes classifier under the assumption that the classifier uses log probabilities. Explain your notation.

$$\hat{c} = \arg\max_{c \in C} \ldots$$

c) A certain Naive Bayes classifier has a vocabulary consisting of 48,359 unique words. We have the following counts for a collection of movie reviews:

$$\#(\text{pokemon}, \text{pos}) = 0 \qquad \#(\bullet, \text{pos}) = 712480$$
$$\#(\text{pokemon}, \text{neg}) = 20 \qquad \#(\bullet, \text{neg}) = 636767$$

Here $\#(w, c)$ denotes the number of occurrences of the word w in documents with class c, and $\#(\bullet, c)$ denotes the total number of tokens in documents with class c. Estimate the following probabilities using maximum likelihood estimation with add-k smoothing, $k = 0.1$. Answer with fractions.

i. $P(\text{pokemon} \mid \text{pos})$ ii. $P(\text{pokemon} \mid \text{neg})$

02 Language modelling (3 points)

The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English, containing approximately 560 million tokens. In this corpus we have the following counts of unigrams and bigrams:

snow	*white*	*white snow*	*purple*	*purple snow*
38,186	256,091	122	11,218	0

a) Estimate the following probabilities using maximum likelihood estimation without smoothing. Answer with fractions.

i. $P(purple)$

ii. $P(snow \mid white)$

b) Explain why a bigram model trained on the COCA corpus using maximum likelihood estimation without smoothing may not be very useful. Use the following (preprocessed) sentence as an example to illustrate the problem.

a vibrant and colorful artist with a playful sense of humor known under the pseudonym purple snow , she was born in lithuania .

c) We use maximum likelihood estimation with add-k smoothing to train n-gram models on the COCA corpus, with $n \in \{1, \dots, 5\}$ and $k \in \{0, 0.1, 1\}$. The following table shows the entropy of each trained model on the training data. Which row corresponds to which k-value, and why? Answer with a short text.

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
$k = a$	7.3376	5.9834	6.7332	6.9556	7.0555
$k = b$	7.3376	3.4269	1.4290	0.5436	0.4171
$k = c$	7.3376	7.3837	8.4573	8.6577	8.7350

03 Part-of-speech tagging (3 points)

a) The following matrices specify (parts of) a hidden Markov model. The marked cell specifies the probability for the transition from BOS to AB.

	AB	PN	PP	VB	EOS
BOS	1/11	1/10	1/12	1/11	1/25
AB	1/11	1/11	1/11	1/10	1/14
PN	1/11	1/12	1/12	1/10	1/16
PP	1/13	1/11	1/12	1/14	1/18
VB	1/11	1/10	1/10	1/13	1/15

	she	got	up
AB	1/25	1/25	1/14
PN	1/13	1/25	1/25
PP	1/25	1/25	1/13
VB	1/25	1/14	1/19

Which probability does this model assign to the following tagged sentence (word/tag)? Answer with a fraction.

up/PP she/PN got/VB

b) One difference between greedy tagging with the multi-class perceptron and tagging with hidden Markov models based on the Viterbi algorithm is what features the two models have access to. Which of the following features can be expressed in a hidden Markov model? Answer with a short text.

i. current word

ii. word to the left of the current word

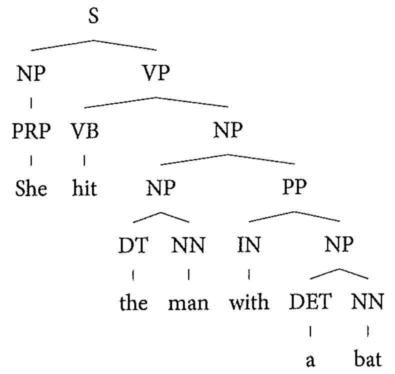
iii. word to the right of the current word

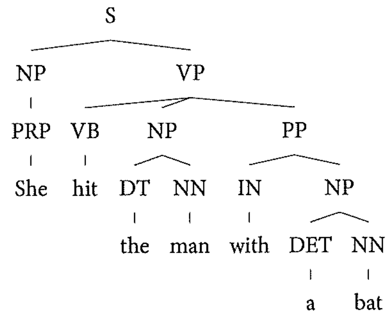
iv. part-of-speech tag of the word to the left of the current word

c) State at least two other differences between greedy tagging with the multi-class perceptron and tagging with hidden Markov models based on the Viterbi algorithm, apart from the aforementioned differences in what features the two models have access to.

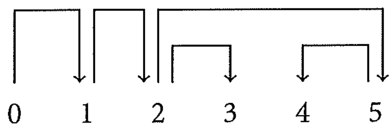
04 Syntactic analysis **(3 points)**

a) Below is a small phrase structure treebank, consisting of just two trees. Read off all rules whose left-hand sides are either NP or VP and estimate their rule probabilities using maximum likelihood estimation.





b) State a sequence of transitions that make an transition-based dependency parser produce the following dependency tree:



c) Explain how transition-based dependency parsers are trained. What data is required? Answer with a short text.

05 Semantic analysis (3 points)

a) Provide an example word pair for each of the following semantic relations:

i. synonym
ii. antonym
iii. hyponym
iv. hypernym

b) Here are three signatures (glosses and examples) from Wiktionary for different senses of the word *colour*:

1 The spectral composition of visible light. *Humans and birds can perceive colour.* **2** A particular set of visible spectral compositions, perceived or named as a class. *Most languages have names for the colours black, white, red, and green.* **3** Hue as opposed to achromatic colours (black, white, and grays). *He referred to the white flag as one 'drained of all colour'.*

Based on these signatures, which of the three senses of the word *colour* does the Lesk algorithm predict in the following sentence? Ignore the word *colour*, stop words, and punctuation.

As the large flag of blue colour was raised in a highly visible spot at the top of the mountain, a light rain began to fall.

c) We read off word vectors from the following co-occurrence matrix (target words correspond to rows, context words correspond to columns):

	HuSHa'	*Ha'DIbaH*
qa'vIn	5	1
qurgh	5	5
jonta'	1	0
Dargh	1	4

Sort the four words in decreasing degree of semantic similarity (most similar to least similar) to the word *jonta'*, assuming that semantic similarity is measured as the angle between word vectors.

Part B

06 **Levenshtein distance** **(6 points)**

The following matrix shows the values computed by the Wagner–Fischer algorithm for finding the Levenshtein distance between the two words *student* and *teacher*. Note that the matrix is missing a value (marked cell).

t	7	6	5	5	5	6	6	7
n	6	5	4	4	5	5	6	6
e	5	4	3	4	4	5	5	6
d	4	3	3		4	5	6	7
u	3	2	2	3	4	5	6	7
t	2	1	2	3	4	5	6	7
s	1	1	2	3	4	5	6	7
#	0	1	2	3	4	5	6	7
	#	t	e	a	c	h	e	r

a) Define the concept of the Levenshtein distance between two words. The definition should be understandable even to readers who have not taken this course.

b) Calculate the value for the marked cell. Explain your calculation. Show that you have understood the Wagner–Fischer algorithm.

c) It is much more likely for a user to mistype the word *student* as *stusent* than as *stulent*; this is because the keys for the letters *d* and *s* are much closer to each other on the keyboard than the keys for the letters *d* and *l*. Explain how the Wagner–Fischer algorithm could be adapted to take this information into account.

07 Viterbi algorithm (6 points)

The following matrices specify a hidden Markov model in terms of costs (negative log probabilities). The marked cell gives the transition cost from BOS to AB.

	AB	PN	PP	VB	EOS
BOS	11	10	12	11	25
AB	11	11	11	10	14
PN	11	12	12	10	16
PP	13	11	12	14	18
VB	11	10	10	13	15

	she	got	up
AB	25	25	14
PN	13	25	25
PP	25	25	13
VB	25	14	19

When using the Viterbi algorithm to calculate the least expensive (most probable) tag sequence for the sentence 'she got up' according to this model, we get the following matrix. Note that the matrix is missing three values (marked cells).

		she	got	up	
BOS	0				
AB		36	59	72	
PN		23	60	82	
PP		A	60	B	
VB		36	47	78	
EOS					C

a) Calculate the missing values.

b) Starting in cell C, list the backpointers for the matrix and state the least expensive (most probable) tag sequence for the sentence.

c) Let m denote the number of tags in the hidden Markov model and let n denote the length of the input sentence. The memory required by the Viterbi algorithm is in $O(mn)$. The runtime of the Viterbi algorithm is in $O(m^2n)$. Explain what these statements mean and why they hold.

08 Named entity tagging (6 points)

Named entity tagging is the task of identifying entities such as persons, organisations, and locations in running text. One way to approach this task is to use the same techniques as in part-of-speech tagging. However, a complicating factor is that named entities can span more than one word. Consider the following sentence:

Alfred Nobel was an inventor from Sweden.

In this example, while the unigram 'Sweden' corresponds to one named entity of type 'location' (LOC), we would also like to identify the bigram 'Alfred Nobel' as a mention of *one* named entity, of type 'person' (PER).

To solve this problem, we can use the so-called IOB tagging. In this scheme we introduce a special 'part-of-speech' tag for the beginning (B) and inside (I) of each entity type, as well as one tag for words outside (O) any entity. Here is the example sentence represented with IOB tags:

$\text{Alfred}_{\text{B-PER}}$ $\text{Nobel}_{\text{I-PER}}$ was_{O} an_{O} $\text{inventor}_{\text{O}}$ from_{O} $\text{Sweden}_{\text{B-LOC}}$

a) Tag the following (tokenised) sentence with IOB tags for the entity types LOC (location), ORG (organisation), and PER (person).

   American Airlines , a unit of AMR Corp. , immediately matched the move , spokesman Tim Wagner said .

b) Named entity taggers often use *gazetteers*. Explain what a gazetteer is and how it can be integrated into a named entity tagger based on the multi-class perceptron.

c) In addition to gazetteers, named entity taggers often use *word shape features*, which represent the abstract letter pattern of a word. Give a concrete example of such a feature, and explain why word-shape features are especially useful in named entity tagging, compared to part-of-speech tagging.

Part C

09 **The limitations of word embeddings** **(9 points)**

This item is an essay question that tests your understanding of the following article:

Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. *Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.* Science 356(6334):183–186, 2017.

a) Summarise the paper by Caliskan et al. in your own words.

b) Caliskan et al. investigate whether word vectors embed knowledge of the real-world properties associated with the words, or, as they put it, 'whether there is an algorithm that can extract of predict the property, given the vector'. Explain their approach using the example illustrated in Fig. 1 below (reproduced from the article).

Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

c) Caliskan et el. write: 'Our work has implications for AI and machine learning because of the concern that these technologies may perpetuate cultural stereotypes.' Explain this statement. Would you agree with the authors' concern?