Optional tests

Marco Kuhlmann

01 Text classification (3 points)

a) Here are confusion matrices from the evaluation of three text classifiers. In each 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	12	9	6
B	6	15	3
C	9	3	18

(classifier 1)

	A	B	C
A	18	6	3
B	3	12	9
C	6	9	15

(classifier 2)

	A	B	C
A	15	6	6
B	9	12	3
C	3	9	18

(classifier 3)

Which of the three classifiers has the highest

i. recall with respect to class `B`?

ii. precision with respect to class `A`?

b) Use Maximum Likelihood estimation with add-one smoothing to estimate the class probabilities and word probabilities of a Naive Bayes text classifier on the following document collection. Assume that the vocabulary consists of all words in the collection. Answer with fractions.

	document	class
1	Stockholm Oslo	S
2	Copenhagen Stockholm	D
3	Stockholm Copenhagen	S
4	Copenhagen	D

c) Based on the estimated probabilities, which class does the classifier predict for the two-word document 'Stockholm Oslo'? Show that you have understood the Naive Bayes classification rule.

Sample answers:

a) (i) classifier 1, (ii) classifier 2

b) Estimated probabilities:

$$P(\mathtt{S}) = 2/4 \quad P(\text{Stockholm} \mid \mathtt{S}) = 3/7 \quad P(\text{Oslo} \mid \mathtt{S}) = 2/7 \quad P(\text{Copenhagen} \mid \mathtt{S}) = 2/7$$
$$P(\mathtt{D}) = 2/4 \quad P(\text{Stockholm} \mid \mathtt{D}) = 2/6 \quad P(\text{Oslo} \mid \mathtt{D}) = 1/6 \quad P(\text{Copenhagen} \mid \mathtt{S}) = 3/6$$

c) The system first computes class-specific scores:

$$\text{score}(\mathtt{S}) = P(\mathtt{S}) \cdot P(\text{Stockholm} \mid \mathtt{S}) \cdot P(\text{Oslo} \mid \mathtt{S}) = \frac{2}{4} \cdot \frac{3}{7} \cdot \frac{2}{7} = \frac{3}{49}$$
$$\text{score}(\mathtt{D}) = P(\mathtt{D}) \cdot P(\text{Stockholm} \mid \mathtt{D}) \cdot P(\text{Oslo} \mid \mathtt{D}) = \frac{2}{4} \cdot \frac{2}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

The system then predicts the class with the highest score, here: `S`.

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 containing concrete numbers. You do not have to simplify the fractions.

i. $P(\mathit{white})$ ii. $P(\mathit{snow} \mid \mathit{white})$

b) Estimate the following probabilities using maximum likelihood estimation with additive smoothing, $k = 0.01$. Assume that the vocabulary consists of 1,254,193 unique words. Answer with fractions containing concrete numbers. You do not have to simplify the fractions.

i. $P(\mathit{snow})$ ii. $P(\mathit{snow} \mid \mathit{purple})$

c) We use maximum likelihood estimation with add-k smoothing to train n-gram models on the COCA corpus, with $n \in \{1, \ldots, 5\}$ 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.

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

Sample answers:

a) Maximum likelihood estimation without smoothing:

$$P(\mathit{white}) = \frac{256{,}091}{560 \cdot 10^6} \qquad\qquad P(\mathit{snow} \mid \mathit{white}) = \frac{122}{256{,}091}$$

b) Maximum likelihood estimation with add-k smoothing, $k = 0.01$:

$$P(\mathit{snow}) = \frac{38{,}186 + 0.01}{560 \cdot 10^6 + 0.01 \cdot 1{,}254{,}193}$$

$$P(\mathit{snow} \mid \mathit{purple}) = \frac{0 + 0.01}{11{,}218 + 0.01 \cdot 1{,}254{,}193}$$

c) The k values for the three rows are $k = 0.1$, $k = 1$, and $k = 0$. On the one hand, higher values of n *decrease* the entropy, as predictions are conditioned on longer and longer contexts. On the other hand, higher values of k *increase* the entropy, as larger and larger shares of the total probability mass are redistributed to n-grams not observed training. The latter effect increases with increasing values of n, as there are more and more unobserved n-grams.

03 Part-of-speech tagging **(3 points)**

a) The evaluation of a part-of-speech tagger produced the confusion matrix shown to the left below. The marked cell gives the number of times the system tagged a word as a verb (`VB`) whereas the gold standard specified it as a noun (`NN`).

	NN	JJ	VB
NN	58	6	1
JJ	5	11	2
VB	0	7	43

	NN	JJ	VB
NN			
JJ			
VB			

Fill the confusion matrix to the right with numbers in such a way that accuracy is as in the left matrix, but precision on adjectives (`JJ`) is 100%.

b) Suppose that you know the probability P of the following tagged sentence under some hidden Markov model:

Kim	hates	broccoli	but	loves	parsnips
PROPN	VERB	NOUN	CONJ	VERB	NOUN

Replacing the tag CONJ with X yields a different tagged sentence with a new probability P'. Fill in the probabilities needed to compute P' from P.

$$P' = P \cdot \frac{\boxed{} \cdot \boxed{} \cdot \boxed{}}{\boxed{} \cdot \boxed{} \cdot \boxed{}}$$

c) State at least three substantial differences between tagging with the hidden Markov model (based on the Viterbi algorithm) and tagging with the multi-class perceptron (based on the greedy left-to-right algorithm).

Sample answers:

a) There are many different solutions; here is one:

	NN	JJ	VB
NN	58	0	7
JJ	5	11	2
VB	7	0	43

b) probability of the new tagged sentence:

$$P' = P \cdot \frac{P(\mathrm{X} \mid \mathrm{NOUN}) \cdot P(\mathrm{but} \mid \mathrm{X}) \cdot P(\mathrm{VERB} \mid \mathrm{X})}{P(\mathrm{CONJ} \mid \mathrm{NOUN}) \cdot P(\mathrm{but} \mid \mathrm{CONJ}) \cdot P(\mathrm{VERB} \mid \mathrm{CONJ})}$$

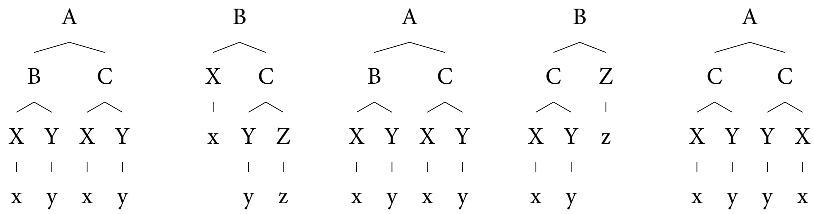
c) see the comparison on the lecture slides (page 46)

04 Syntactic analysis **(3 points)**

a) You sum up all rule probabilities in a certain probabilistic context-free grammar. Which (zero or more) of the following values can you *not* get as a result?

i. 0.42 ii. 1 iii. 4.2 iv. 42

b) Below is a small phrase structure treebank. Read off all rules whose left-hand sides are either A or B and estimate their rule probabilities using maximum likelihood estimation (no smoothing).



c) Draw the dependency tree generated by a transition-based dependency parser after executing the following sequence of transitions:

`SH SH SH SH RA SH SH RA RA LA RA`

Sample answers:

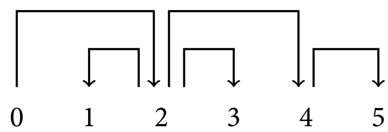
a) (i) and (iii)

b) All rules whose left-hand sides are either A or B:

$$\mathrm{A} \rightarrow \mathrm{B}\ \mathrm{C}\ \tfrac{2}{3} \qquad \mathrm{A} \rightarrow \mathrm{C}\ \mathrm{C}\ \tfrac{1}{3}$$

$$\mathrm{B} \rightarrow \mathrm{X}\ \mathrm{Y}\ \tfrac{2}{4} \qquad \mathrm{B} \rightarrow \mathrm{X}\ \mathrm{C}\ \tfrac{1}{4} \qquad \mathrm{B} \rightarrow \mathrm{C}\ \mathrm{Z}\ \tfrac{1}{4}$$

c) Generated dependency tree:



05 Semantic analysis (3 points)

a) Choose the correct semantic relation: synonym, antonym, hyponym, hypernym?

buy	is a/an … of	purchase
chair	is a/an … of	furniture
automobile	is a/an … of	car
bird	is a/an … of	eagle
synonym	is a/an … of	antonym

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 have seen that semantic similarity can be quantified in terms of cosine similarity of word vectors in a high-dimensional vector space constructed from co-occurrence counts. Based on this idea, which (zero or more) of the following words do you expect to end up close to the word *apple* in such a vector space?

i. mango ii. computer iii. swim iv. apples

Sample answers:

i. Semantic relations:
   - *buy* is a synonym of *purchase*
   - *chair* is a hyponym of *furniture*
   - *automobile* is a synonym of *car*
   - *bird* is a hypernym of *eagle*
   - *synonym* is an antonym to *antonym*

ii. sense 1 (match with *visible* and *light*)

iii. mango (another fruit), computer (as in the brand name), apples (inflection form)