Question 1 (25 points total)

Q1 10

the following corpus:

```
add the flour then stir
next put flour and sugar into the bowl
stir the sugar into the butter
```

1a (10 points): Write down a word-context counts table for the words flour and sugar using a context of one word.

context

		sugar	flour	put	next	add	the	then	stir	into	butter	and	bowl	P(w
word	flour	0	0	1	0	0	1	1	0	0	0	1	0	4/8
	sugar	0	0	0	0	0	1	0	0	2	0	1	0	4/8
p(context)		0	0	1/8	0	0	2/8	1/8	0	2/8	0	2/8	0	

(OR)

context

		the	then	put	and	into	p(w)
Word	flour	1	1	1	1	0	1/2
	sugar	1	1	0	0	2	1/2
p(context)		1/4	1/4	1/8	1/8	1/4	

Total: 8

1b (10 points): Use the table you wrote in question 1a to calculate a positive pointwise tual information (PPMI) table for flour and sugar.

Q2 10

wing information may be helpful in answering this question:

x	$\log_2(x)$
1/8	-3
1/4	-2
1/2	-1
1	0
2	1
4	2
8	3

PPMI table

context

word		sugar	flour	put	next	add	the	then	stir	into	butter	and	bowl
	flour	0	0	1	0	0	0	1	0	0	0	0	0
	Sugar	0	0	0	0	0	0	0	0	1	0	0	0

PMI (flour, put) = $\log \frac{1/8}{1/8 \times 4/8} = \log 2 = 1 \Rightarrow$ PPMI = 1

PMI (flour, the) = $\log \frac{1/8}{2/8 \times 4/8} = \log 1 = 0 \Rightarrow$ PPMI = 0

PMI (flour, then) = $\log \frac{1/8}{1/8 \times 4/8} = \log 2 = 1 \Rightarrow$ PPMI = 1

PMI (flour, and) = $\log \frac{1/8}{2/8 \times 4/8} = \log 1 = 0 \Rightarrow$ PPMI = 0

PMI (sugar, the) = $\log \frac{1/8}{2/8 \times 4/8} = \log 1 = 0 \Rightarrow$ PPMI = 0

PMI (sugar, into) = $\log \frac{2/8}{2/8 \times 4/8} = \log 2 = 1 \Rightarrow$ PPMI = 1

PMI (sugar, and) = $\log \frac{1/8}{2/8 \times 4/8} = \log 1 = 0 \Rightarrow$ PPMI = 0

Q3 5

1c (5 points): Using the results in question 1b, calculate the cosine of the PPMI vectors (i.e. the rows of the PPMI table) for <u>flour</u> and <u>sugar</u>.

cosine = 0

Question 2 (25 points total)

Q4 10 points)

Recall that according to IBM Model 1, the probability of translating a sentence $\mathbf{e} = e_1, \ldots, e_n$ to $\mathbf{f} = f_1, \ldots, f_m$ according to alignment a which maps every f_i to some e_j is

$$p(\mathbf{f}, a|\mathbf{e}) \propto \prod_{i=1}^{m} t(f_i|e_{a(i)})$$

Further recall that, as part of the EM algorithm for estimating $p(\mathbf{f}, a|\mathbf{e})$ from data, when collecting fractional counts to estimate the table of word translation probabilities t, one must calculate, for each possible alignment a of a sentence pair $(\mathbf{f}, \mathbf{e})$, the probability of that alignment, i.e. $p(a|\mathbf{f}, \mathbf{e})$, using the previous estimate of t.

Rewrite $p(a|\mathbf{f}, \mathbf{e})$ in terms of $t(f|e)$. Give an explanation for each step of the derivation. An explanation can be, e.g. "definition of conditional probability," "Bayes' law," "law of total probability," or "definition of IBM Model 1."

The first step is given to you below:

$$p(a|\mathbf{f}, \mathbf{e}) = \frac{p(a, \mathbf{f}, \mathbf{e})}{p(\mathbf{f}, \mathbf{e})}; \text{ definition of conditional probability}$$

$$= \frac{p(a,f|e) \cdot \cancel{p(e)}}{p(f|e)\ \cancel{p(e)}} \quad ; \text{defn. of conditional probability}$$

$$\propto \frac{\prod_{i=1}^{m} t(f_i|e_{a(i)})}{p(f|e)} \quad ; \text{defn. of IBM model 1}$$

$$= \frac{\prod_{i=1}^{m} t(f_i|e_{a(i)})}{\sum_a p(f,a|e)} \quad ; \text{law of total probability}$$

$$= \frac{\prod_{i=1}^{m} t(f_i|e_{a(i)})}{\sum_a \prod_{i=1}^{m} t(f_i|e_{a(i)})}$$

2b (10 points)

Q5 10

$p(a|\mathbf{f}, \mathbf{e})$ for each alignment of the sentence pair ("das buch", "the book"). You may use fractions s to represent your answer. Assume the four alignments below (labeled A, B, C, D) are the only possible alignments for that sentence pair. Use the following t table:

	das	ein	buch	haus
the	3/5	0	1/5	1/5
a	0	3/5	0	2/5
book	2/5	0	3/5	0
house	3/10	3/10	0	2/5

A	das	buch
the	x	x
book		

$p(A|\mathbf{f}, \mathbf{e}) =$ -----

$\frac{3/25}{4/5} = \frac{3}{\cancel{25}_5} \times \frac{\cancel{5}}{4} = \boxed{\frac{3}{20}}$

$\frac{3}{5} \times \frac{1}{5} = \frac{3}{25} = p(f, A | e)$

B	das	buch
the		
book	x	x

$p(B|\mathbf{f}, \mathbf{e}) =$ -----

$\frac{6/25}{4/5} = \frac{\overset{3}{\cancel{6}}}{\cancel{25}_5} \times \frac{\cancel{5}}{\cancel{4}_2} = \boxed{\frac{3}{10}}$

$\frac{2}{5} \times \frac{3}{5} = \frac{6}{25} = p(f, B | e)$

C	das	buch
the	x	
book		x

$p(C|\mathbf{f}, \mathbf{e}) =$ -----

$\frac{9/25}{4/5} = \frac{9}{\cancel{25}_5} \times \frac{\cancel{5}}{4} = \boxed{\frac{9}{20}}$

$\frac{3}{5} \times \frac{3}{5} = \frac{9}{25} = p(f, C | e)$

D	das	buch
the		x
book	x	

$p(D|\mathbf{f}, \mathbf{e}) =$ -----

$\frac{2/25}{4/5} = \frac{\cancel{2}}{\cancel{25}_5} \times \frac{\cancel{5}}{\cancel{4}_2} = \boxed{\frac{1}{10}}$

$\frac{2}{5} \times \frac{1}{5} = \frac{2}{25} = p(f, D | e)$

$$\sum_{a = A, B, C, D} p(f, a | e) = \frac{3}{25} + \frac{6}{25} + \frac{9}{25} + \frac{2}{25}$$

$$= \frac{20}{25} = \frac{4}{5}$$

2c (5 points)

Q6 5 hat in question 2a we observed that $p(\mathbf{f}, a|\mathbf{e}) \propto \prod_{i=1}^{m} t(f_i|e_{a(i)})$. The exact definition (i.e. with an sign instead of a proportion sign) is

$$p(\mathbf{f}, a|\mathbf{e}) = \frac{\epsilon}{(n+1)^m} \prod_{i=1}^{m} t(f_i|e_{a(i)})$$

where $\frac{\epsilon}{(n+1)^m}$ is a term that ensures the model is probabilistic.[1] Using your answer to question 2a, show why this term is not needed to calculate $p(a|\mathbf{f}, \mathbf{e})$.

- Because

$$p(a|f,e) = \frac{p(f,a|e)}{p(f|e)}$$

$$= \frac{p(f,a|e)}{\sum_a p(f,a|e)}$$; defn. of total probability

- Since numerator & denominator will have the constant $\frac{\epsilon}{(n+1)^m}$, they will get cancelled

- Numerator & denominator are both proportional to $\frac{\epsilon}{(n+1)^m}$

[1] $\epsilon = p(m|\mathbf{e})$ but that's not important for this question.

Q7 2

Question 3 **Multiple Choice (2 points each; 30 total): For each question circle all answers that apply. Zero, one, or more than one may apply in each case.** ***If zero answers apply, you must explicitly note this or the question will be treated as unanswered.***

3a **In a 5-gram feed-forward neural LM with one hidden layer, a hidden dimension of 250, vocabulary embeddings of size 500, and a vocabulary of 50,000 (including special symbols), what is the size of the embedding-to-hidden weight matrix?**

1. 2000x250
2. 500x250
3. 250x250
4. 250x50,000
5. 2500x250
6. [✓ 6.] 25,000x500
7. 1000x50,000

3b **In the LM described above in Question 3a, what is the size of the hidden-to-output weight matrix?**

1. 2000x250
2. 500x250
3. 250x250
4. [✓ 4.] 250x50,000
5. 2500x250
6. 25,000x500
7. 1000x50,000

Q8 10

In each of the next three questions (3c,3d,3e), consider the following reference (gold) English translation of a foreign-language sentence and three potential machine translation outputs:

reference	we bought the sturdy boat for seven dollars
translation A	we bought the boat for seven dollars
translation B	the sturdy boat for many dollars did we bought it indeed
translation C	the strong boat was purchased for seven dollars by us

3c If calculating BLEU for a corpus containing just this sentence, which translation(s) would result in a score of 0?

1. translation A
2. translation B
[✓ 3. translation C]

3d If calculating BLEU for a corpus containing just this sentence, which translation(s) would incur a brevity penalty?

[✓ 1. translation A]
2. translation B
3. translation C

3e Which translation would be rewarded by METEOR more than BLEU?

1. translation A
2. translation B
[✓ 3. translation C]

3f carrot is a __ of vegetable

1. hypernym
[✓ 2. hyponym]
3. holonym
4. meronym

3g hand is a __ of finger

1. hypernym
2. hyponym
[✓ 3. holonym]
4. meronym

Q9 2

3h Which of the following is a description of a translation of a Chinese sentence generated by a Chinese-to-English MT engine that would be scored as having high adequacy but low fluency by monolingual English-speaking human evaluators?

1. A sequence of random words that are not fluent English and have nothing to do with the input

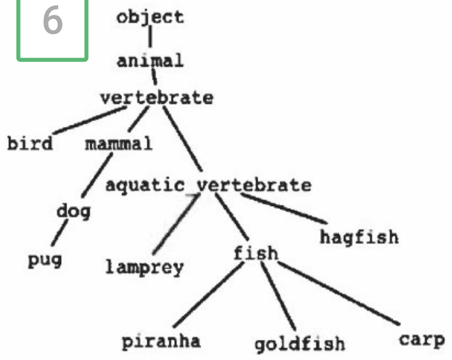
2. A perfectly phrased English sentence that has nothing to do with the Chinese input

[✓ 3.] A sequence of phrases that capture much of the meaning of the Chinese original but do not have proper subject/verb agreement and use prepositions improperly

4. A perfectly phrased Italian sentence that perfectly captures the meaning of the Chinese input

Questions 3i, 3j, 3k, and 3l refer to the hypernym tree and table of counts below:

Q10 6



animal	30
bird	5
dog	20
fish	2
pug	15
lamprey	5
hagfish	5
piranha	3
goldfish	5
carp	10

3i What is pathlen(pug, piranha)?

[✓ 1. 7]

2. 6

3. 40

4. 5

5. 2

3j What is simpath(fish, piranha)?

1. 1/4

[✓ 2. 1/2]

3. 1

4. 2

5. 4

3k What is the Resnik similarity of "lamprey" and "goldfish"?

1. -log(.7)

2. -log(.5)

[✓ 3. -log(.3)]

4. 0

5. -log(.2)

$$-\log P(aq\text{-}vert) = -\log \frac{5+2+5+3+5+10}{100}$$

$$= -\log 0.3$$

3l What is the lowest common subsumer (LCS) of "bird" and "hagfish"?

Q11 8

(.7)

[✓ 2. vertebrate]

3. 5

4. 0

Questions 3m, 3n refer to the following feature vectors $\mathbf{f} = [f_1, f_2, f_3]$ for a list of two translation hypotheses (h_1 and h_2) for a single sentence:

hyp	f_1	f_2	f_3
h_1	0	-1	2
h_2	1	3	-1

3m Assume the weight vector $\mathbf{w}$ is $\mathbf{w} = [1, 1, 1]$ for a log-linear model $m = \mathbf{f} \cdot \mathbf{w}$. Which hypothesis has the higher model score?

1. h_1

[✓ 2. h_2]

3. They have the same score

3n If we run MERT starting from the weight vector $\mathbf{w} = [1, 1, 1]$ and seek a new value for w_1 (the weight for feature f_1), at what value of w_1 would the model scores for both hypotheses be equal?

1. -2

[✓ 2. -1]

3. 0

4. 1

5. 2

3o In the sentence that begins "European countries, *especially* France, England, and Spain..." from the pattern "X, especially Y," we know that there is a relation between "European countries" and "France." This method of extracting relations is called...

1. a supervised machine learning method

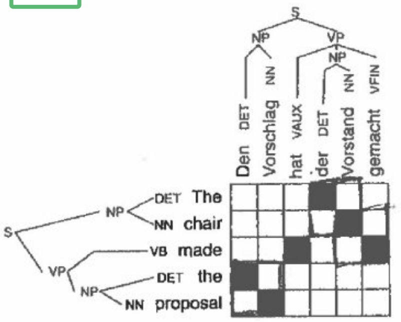
2. an unsupervised machine learning method

3. a semi-supervised machine learning method

[✓ 4. a pattern/rule-based method]

Question 4 Short Answers (5 points each, 20 points total)

Q12 4a 1



Consider the above word-aligned sentence pair with syntactic annotation.

1. Considering alignment constraints, list all legal phrase pairs that can be extracted from this sentence pair (do not consider the syntactic annotation in this part and do not extract any hierarchical phrase translation rules). The - der, chair - Vorstand, made - hat, made - gemacht, the - Den, proposal - Vorschlag, The chair - der Vorstand, the proposal - Den Vorschlag, chair made - Vorstand germacht, The chair made - der Vorstand gemacht, The chair made - hat der Vorstand gemacht, The chair made the - hat der Vorstand gemacht, The chair made the proposal - hat der Vorstand gemacht, The chair made the proposal - Vorschlag hat der Vorstand gemacht, The chair made the proposal - Den Vorschlag hat der Vorstand germacht, The chair made - Vorstand gemacht, The chair made the - Vorstand gemacht, The chair made the - Vorstand gemacht, The chair made the - der Vorstand gemacht

2. Which of the above phrase pairs, if any, would not be allowed if you additionally considered syntactic annotation constraints?

4b

Q13 4

Entity	Category
new york	GPE
pennsylvania	GPE
bill gates	PER
justin bieber	PER
michele buck	PER

Entity	Category
hershey	CMP
microsoft	CMP
york peppermint patty	PRD
never say never	MOV
red sox	SPR

1. Using the named entity lexicon above, annotate the following two-sentence corpus with BIO tags for named entity recognition (place the tag above each word).

B-PER	I-PER	O	B-GPE	I-GPE	B-CMP	O
bill	gates	gave	new	york	microsoft	stock

O	O	O	B-PER	I-PER	O	O	B-MOV	I-MOV	I-MOV	
for	tickets	to	justin	bieber	's	movie	never	say	never	.

O	O	O	O	O	B-PRD	I-PRD	I-PRD
i	would	never	buy	a	york	peppermint	patty

O	O	O	O	O	B-GPE	B-CMP	O
because	though	she	lives	in	pennsylvania	hershey	president

B-PER	I-PER	O	O	B-SPR	I-SPR	O	
michele	buck	is	a	red	sox	fan	.

2. Would the binary feature "previous word is tagged with a label containing GPE" help a named entity recognizer (that does not have access to the lexicon you used) predict CMP entities in this data? Why or why not?

Yes, we have 2 such examples in this data

1. york microsoft
2. pennsylvania hershey

3. Could the feature described above be used by a recognizer implemented as a simple logistic regression classifier, as a recognizer implemented as a CRF, both kinds, or neither kind (you do not need to justify your answer)?

Only as a simple logistic regression classifier

Q14 0

4c You are using an n-gram feed-forward neural network language model with a three-word vocabulary to generate a word given context. After multiplying the hidden vector by the hidden-to-output matrix and adding in the output bias terms, you obtain the preliminary output vector $[x, y, z]$.

1. Use the softmax activation function to convert the preliminary vector into a vector of probabilities of each of the three words. You may use summation notation, variables, etc. as long as your answer is clear.

2. Give two reasons why the softmax activation function is a good function to use for transforming $[x, y, z]$ into probabilities.

4d

Q15 2

almost there

Sentence	Centauri	Arcturan
1	wiwok farok axok stok	totat jjat quat cat
2	farok lalok ororok sprok lalok izok enemok	wat jjat bichat wat dat vat eneat

1. You are given a corpus of (Centauri, Arcturan) translation pairs, a portion of which is shown above. Using just that portion, explain how you are able to reason that (farok, jjat) is a valid word translation pair.

2. Using the entire corpus you obtain a partial alignment for one of the above sentences, shown below. Using just this partial alignment, explain how you are able to complete the alignment by reasoning that (enemok, eneat) is a valid word translation pair.

	wat	jjat	bichat	wat	dat	vat	eneat
farok		x					
lalok	x						
ororok			x				
sprok					x		
lalok				x			
izok						x	
enemok							