Midterm

Computational Linguistics II
April 4, 2012

Name: _____
by writing my name I swear by the honor code

Read all of the following information before starting the exam:

- Show all work, clearly and in order, if you want to get full credit. I reserve the right to take off points if I cannot see how you arrived at your answer (even if your final answer is correct).
- You have two and a half hours to complete this exam (but you shouldn't need it).
- This exam is open book, open notes, but closed Internet. You may use a calculator, but you shouldn't need to.
- Justify your answers algebraically whenever possible to ensure full credit. Be sure to have units for all answers that call for them.
- Circle or otherwise indicate your final answers.
- Please keep your written answers brief; be clear and to the point. I will take points off for incorrect or irrelevant statements.
- Attempt to answer all questions.
- Good luck!

1 Machine Translation (25 points)

1.1 Number of Phrase Pairs? (12 points)

Consider these word alignment examples.

1. Alignment 1

	f_1	f_2	f_3
e_1	X		
e_2		X	
e_3			X

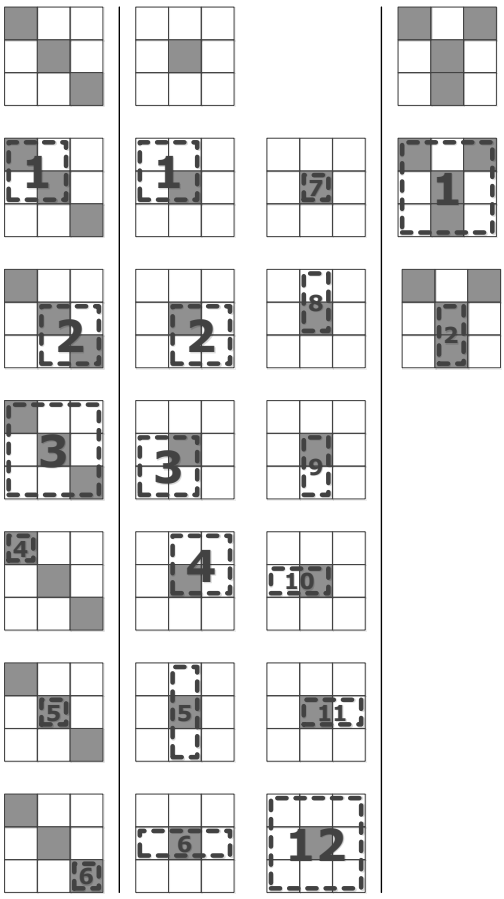
2. Alignment 2

	f_1	f_2	f_3
e_1			
e_2		X	
e_3			

3. Alignment 3

	f_1	f_2	f_3
e_1	X		X
e_2		X	
e_3		X	

For each, which and how many phrase pairs can be extracted (assume that you cannot have "gapped" phrases)? It's fine to draw squares or list the phrase pairs, but do be sure to give a total number. What do these examples suggest for the relationship between number of alignment points and the number of extracted phrase pairs?



In general, the fewer alignment points, the more phrase pairs are possible. However in the extreme, this is not true, as a sentence without alignments cannot have phrase pairs.

1.2 Out of Vocabulary Words (13 points)

Consider the problem of out of vocabulary (OOV) translations. For concreteness, let's assume that we want to do Chinese to English translation.

- When might this be a problem? Is this a problem that could be solved by larger datasets alone?
- At a high level, what techniques might you employ to handle these problems? Consider ways to introduce new words into your systems and to make better use of words that are in your system.
- How might OOV words interfere / work with your translation model and your language model? What changes would you have to make during the decoding process?

1. Named entities and nonce words
2. Transliteration is one option. Another option is to find a source word that appears in similar contexts and use that source word instead.
3. If you're adding new target words, those won't appear in your language model. So you'll want to make sure that your language model doesn't give zero scores to new words.

2 Log Linear Models (25 points)

Take V to be the set of possible words (e.g. "the", "cat", "dog", ...). Take V' to be the set of all words in V **plus** their reverses (e.g. "the", "eht", "cat", "tac", "dog", "god"). You can assume that there are no palindromes in v (e.g. "eye"). Nathan L. Pedant generates $(x, y) : x \in V, y \in V'$ pairs as follows:

- With probability $\frac{1}{2}$ he chooses y to be identical to x
- With probability $\frac{1}{3}$ he chooses y to be the reverse of x
- With probability $\frac{1}{6}$ he chooses y to be some string that is neither x nor the reverse of x

Create a log-linear distribution (i.e. supply features f and weights θ) of the form:

$$p(y|x, \vec{\theta}) = \frac{\exp \sum_i \theta_i f_i(x, y)}{\sum_{y'} \exp \sum_i \theta_i f_i(x, y')} \tag{1}$$

that models Nathan's process perfectly. You'll get full points for a model with only two parameters (e.g. θ_1 and θ_2 only), nearly full points for a three parameter model, and fewer points for the more parameters you need.

For convenience, let's say that $|V| = N$. Define the following features:

1. $f_1(x, y) \equiv 1 \iff x == y$
2. $f_2(x, y) \equiv 1 \iff \text{reverse}(x) == y$

Thus, we know that:

$$\frac{\exp \theta_1}{N - 2 + \exp \theta_1 + \exp \theta_2} = \frac{1}{2} \tag{2}$$

$$\frac{\exp \theta_2}{N - 2 + \exp \theta_1 + \exp \theta_2} = \frac{1}{3} \tag{3}$$

$$\tag{4}$$

Solving this system of linear equations gives $\exp \theta_1 = 3N - 6$ and $\exp \theta_2 = 2N - 4$. Thus, we can create the probability distribution with these parameters:

1. $\theta_1 \equiv \log(3N - 6)$
2. $\theta_2 \equiv \log(2N - 4)$

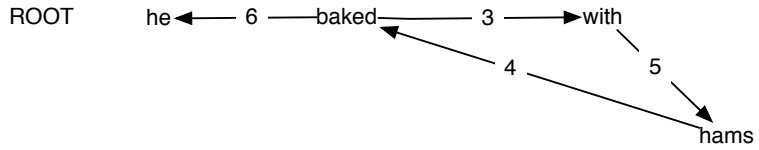
3 Dependency Parsing (25 points)

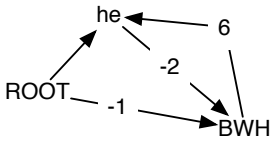
Consider the following arc-factored dependency parsing problem. Read rows as weights "from" that word; that is, the weight for the edge going from "baked" to "he" is 6, but the weight of the edge from "he" to "baked" is 0.

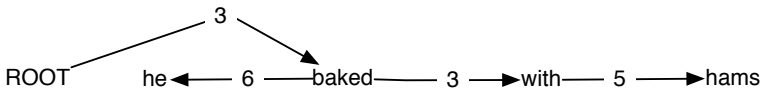
	ROOT	he	baked	with	hams
ROOT		0	3	0	3
he	$-\infty$		0	1	1
baked	$-\infty$	6		3	0
with	$-\infty$	0	0		5
hams	$-\infty$	3	4	2	

Run the Chu-Liu-Edmonds algorithm to determine the maximum spanning tree of the sentence and give the score of that sentence. Be sure to show all the steps in the algorithm (e.g. if you contract a cycle, be sure to show the edges present in the new graph). Be sure to give your final answer as a **directed** tree.

First, we greedily choose the best incoming edge for each node. This creates a cycle between baked, with, and hams. So we contract those three nodes into "BWH" and recompute the weights. Recall that the weights for edges to "BHW" must also subtract the cost of breaking the cyle. Now the best edge into "BWH" is from "ROOT" $(3 - 4 = -1)$, so we take that edge, and we now have a tree. We're done and can reconstruct the full tree.







4 Hypothesis Tests (25 points)

Hoshi the exolinguist is studying a new language that they've encountered, Klingon. Klingon actively marks volition in the language via suffixes, and Hoshi wants to understand the relationship between volition and pronouns. Pronouns in Klingon are marked by verb prefixes, which specialize for person, number, and ergativity. Some examples of volition markers in Klingon:

word	morphemes	morpheme translation	English translation
vIleghnIS	vI legh nIS	I see (volition = need)	I need to see (him / her)
Heghqang	Hegh qang	die (volition = willing)	He is willing to die
cheHoHvIp	che HoH vIp	you-us kill (volition = afraid)	You (plural) are afraid to kill us

Hoshi uncovered the following counts:

	-vIp (afraid)	-nIS (need)	-rup (prepared)
jI- (I, no object)	1	4	0
bI- (you, no object)	6	4	0
0 (they / he / she / it, no object)	3	2	5

Hoshi hypothesizes that the "-rup" suffix only is applied to subjects that lack animacy and that the "-vIp" suffix is only applied to subjects that lack honor (and thus would not be associated with first person pronouns).

1. Conduct a χ^2 test, providing the value of the χ^2 statistic, the expected counts, and the degrees of freedom. Does it reject the null? You may find the table of significance cutoffs on the next page helpful.

   Expected counts:

   2 2 1
   4 4 2
   4 4 2

   $$\chi^2 = (1-2)^2/2+(4-2)^2/2+(0-1)^2/1+(6-4)^2/4+(4-4)^2/4+(0-2)^2/2+(3-4)^2/4+(2-4)^2/4+(5-2)^2/2 = 12.25 \quad (5)$$

   Which with $df = (3-1)(3-1) = 4$, rejects the null at 0.05.

2. Assume that the null is rejected. Does that back up Hoshi's hypothesis?

   No. That the null is rejected only confirms that there are not independent but doesn't state the nature of the dependence.

3. Is it appropriate to use the χ^2 test here? Why or why not? What more do you need to know about how the data were collected?

   There are expected counts that are less than five, which means that the assumptions are not met. Moreover, we would need to know that the sentences were representative and that there weren't confounding effects that might cause correlations that aren't real.

df	$P = 0.05$	$P = 0.01$	$P = 0.001$
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59