CSCI 5832 Spring 2010 Final

Name: _____

On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work. _____.

1. **(5 points) True of False:** It's been a long year.
2. **(5 points)** Given a sample of text denoted $w_1 ... w_n$, characterize the computation being performed by the following formula. You can either use a nice succinct name, or describe what it's doing.

$$\prod_{i=1}^{n} P(w_i)$$

It's computing the probability of the word sequence using a unigram language model (or a 1st order Markov assumption).

3. **(5 points) True or False:** The computation in question 2 could be accurately described as a bag-of-words model. True
4. **(5 points)** The ***hypernym*** and ***hyponym*** relations in WordNet hold between which of the following notions (pick one):
   - **a)** word forms
   - **b)** lemmas
   - **c)** synsets ←
   - **d)** llamas
5. **(15 points)** Describe how you could apply a naïve Bayes classifier to the problem of word-sense disambiguation. For the purposes of this question assume that you're dealing with the problem of disambiguating instances of a single word type (e.g, a word like ***plant***). Be sure to specify the usual three parts (what's the model, how to train the model, how to apply the model).

The naïve Bayes model for this would be to argmax P(sense|word) by arxmax P(word|sense)P(sense). To implement this you would need a tagged corpus. The prior is just the proportion for each sense out of the total tagged. P(word|sense) is really the word in some context. We can model the context as a bag of words in a window around the sense in question. To do this divide the corpus

up into sub-collections based on sense; extract windows around each tagged sense; and then train a unigram language model on each sub-corpus. To tag a new instance, extract its window pass to each language model and multiply by the priors. Return the sense with the highest probability.

6. **(5 points)** Statistical approaches to Named-Entity Recognition often use a word-by-word IOB-style feature encoding. In such an approach, it is possible for words to have identical, or nearly identical, sets of features, yet have different target labels (e.g. the vectors for *Gulf* and *Mexico* in the attached example). How do sequence classifiers trained with IOB-style features overcome this apparent problem?

When training our sequence classifiers we include features from the +/- n words around the word being tagged. With this added context the feature vectors are no longer so similar. For example, "gulf" has an "O" label prior to it, while Mexico has a "I_LOC" label.

7. **(15 Points)** Imagine your job was to build a system to populate a database of sentiment related facts about particular car models extracted from car reviews. Describe three distinct sentiment-related sub-problems that would have to be solved in order to extract the primary bit of sentiment-based information being expressed here. (You don't have to solve them).

   Unlike the Lexus, whose V-6 engine occasionally took its sweet time accelerating, the Prius was one peppy little car.

First we need to know what entity the sentiment is about. In this case we can infer sentiment about both the Lexus and the Prius.

Second we need to figure out what aspect of the entity is being evaluated. In this case, it is "acceleration".

Finally, we need to figure out the polarity of the sentiment. In this case, negative for the Lexus and positive for the Prius.

The final result is something like {lexus, acceleration, -} and {prius, acceleration, +} for this instance.

8. Consider the following question and two candidate answers retrieved from relevant documents in the context of a generic factoid-based question answering system.

Who created the first effective polio vaccine?

> *1. Becton Dickinson created the first disposable syringe for use with the mass administration of the first effective polio vaccine.*
>
> *2. The first effective polio vaccine was created in 1952 by Jonas Salk at the University of Pittsburgh.*

a. (5 points) Describe **3** kinds of surface features that would have resulted in their being retrieved and highly ranked in the context of this question.

Both share 1) the main verb "create", 2) entities of the right answer type for a "who" question, and 3) a 5-gram matching a 5-gram in the question ("the first effective polio vaccine")

b. (5 points) Describe an NLP technique that could be used by Q/A systems to prefer the correct passage (the second one) over the first.

Semantic role labeling would allow us to prefer the right answer. The "theme" of create in the question is that 5-gram. The answer to the question should be the "agent" of a sentence with that as the "theme". Passage 2 meets that criteria (even in the passive form). The "theme" of create in the first passage is "syringe" hence the "agent" there is a false positive.

9. **(5 points)** One way to view the problem of machine translation is as a function that finds the best target translation that balances faithfulness with fluency. Give the basic equation for the noisy channel statistical model of MT and explain how faithfulness and fluency are accounted for in it.

The basic model is based on argmaxing P(F|E)P(E). In this formulation P(E) addresses fluency in the target language and P(F|E) addresses how well the target matches the source original.

10. **(10 point)** What problems might arise in applying the standard phrase-based statistical approach to MT to morphologically complex languages? How might these problems be addressed?

In morphologically complex languages, morphology is used to a greater degree to mark up or signal important information. This is as opposed to doing it with say word order or function words as markers. The result of this is that individual words get more complex, and sentences or utterances can be quite simple. This leads to sparseness problems for phrase-to-phrase models. One way around it is to do morphological processing to break complex words down into their parts and then allowing those morphemes to play a role in alignment and phrase discovery.

11. Consider the small sentence-aligned corpus involving Tentauran (top lines) and Barcturan (bottom lines) on the attached last page. Assume we're going to use this corpus to build a phrase-based statistical MT system based on the usual model to **translate from Tentauran to Barcturan**. The first step in such a process is to perform an EM-based word alignment.
    a) **(5 points)** Give the standard Bayesian formulation for this ***specific*** MT problem.

We want to argmax P(B|T) by argmaxing P(T|B)P(B)

b) **(15 points)** Using the usual simplified starting assumptions (1-1 word alignment; equi-probable start probabilities) perform **1 iteration** of EM; give me the ***probability of the alignment*** shown here for the first sentence pair (answer in the form of a fraction is fine; use a calculator if you like).

Start by noting that given the Bayesian formulation we need an alignment from B to T. That is, we need probs like P(lalok | wat). Not the other way around. Next recall that the probability of an alignment is the product of the word alignments that comprise it. So the probability of this alignment is just P(lalok | wat)*P(farok | jjat). To get those we'll do one round of EM. To do that we need a starting model. Since the word alignments are equiprobable to start, each alignment is equally probable. The first sentence pair has two alignments so they're 1/2 and 1/2. The second sentence pair has 6 alignments so they each have a probability of 1/6. We'll use these numbers to discount the counts for each possible word alignment.

We need P(lalok|wat) and P(farok | jjat) meaning that we need the word translation tables for "wat" and "jjat" filled in. Let's start with raw counts for "wat".

Lalok	Farok	Mok	Nok	
3	1	2	2	wat

This table encodes how many pieces of evidence (counts) that there are for each target word. "Lalok" gets 1 count from 1 alignment for the first sentence. And 2 from the second sentence pair (there are 6 alignments there, two of which align wat with lalok).

But we're not done since we have to discount those counts based on the probabilities of the alignments from which they came. Recall that's either 1/2 or 1/6. Doing that gets us...

Lalok	Farok	Mok	Nok	
1*1/2 + 2 * 1/6	1* 1/2	2 * 1/6	2 *1/6	wat

Re-expressing things arithmetically gets us the following counts.

Lalok	Farok	Mok	Nok	
5/6	3/6	2/6	2/6	wat

And normalizing that gets us the following probabilities.

Lalok	Farok	Mok	Nok	
5/12	3/12	2/12	2/12	wat

So the P(lalok | wat) is 5/12.

If we do the same thing for jjat we get the following table.

Lalok	Farok	Mok	Nok	
1/2	1/2			jjat

So in the end. The Probability of that alignment we're after is P(lalok | wat)*P(farok | jjat). Or 5/12 * 1/2, or 5/24.

lalok farok **wat jjat**	**lalok mok nok** **wat nnat gat**

As the oil spill in the Gulf of Mexico moves into its third week with no clear end in sight, the legal and environmental consequences of the disaster could match those faced by Exxon Mobil.

Features Word/POS/Chunk/Capitalization				Label
in	IN	B_PP	Lower	O
the	DT	B_NP	Lower	O
Gulf	NNP	I_NP	Cap	B_LOC
of	IN	I_NP	Lower	I_LOC
Mexico	NNP	I_NP	Cap	I_LOC
moves	VBZ	BVP	Lower	O