NAME:

**CS-5340/6340, Natural Language Processing**
**Final Exam – SOLUTIONS, Fall 2006**

1. (15 pts) Indicate whether each sequence of part-of-speech tags would be accepted by the grammar below.

| **Grammar** |
|---|
| S → NP VP |
| NP → art NP1 |
| NP → NP2 |
| NP1 → adj NP1 |
| NP1 → NP2 |
| NP2 → noun NP2 |
| NP2 → noun NP3 |
| NP2 → noun |
| NP3 → prep NP |
| NP3 → prep NP NP3 |
| VP → modal VP1 |
| VP → aux VP1 |
| VP1 → adv VP1 |
| VP1 → verb VP2 |
| VP1 → verb |
| VP2 → verb VP2 |
| VP2 → verb |
| VP2 → adv |

(a) noun prep noun modal verb
*yes*

(b) noun prep noun prep noun modal verb
*yes*

(c) art noun modal verb
*yes*

(d) art noun prep noun modal verb
*yes*

(e) art noun prep noun modal verb adv
*yes*

(f) art noun prep noun modal verb adv adv
*no*

(g) noun prep noun modal adv adv verb
*yes*

(h) noun prep noun verb adv
*no*

(i) noun noun noun prep noun noun aux verb
*yes*

(j) noun noun prep noun aux adv verb adv
*yes*

(k) noun noun prep noun aux aux adv verb
*no*

(l) noun noun prep noun modal aux adv verb
*no*

(m) art noun noun prep art noun noun aux verb
*yes*

(n) adj noun noun prep art noun aux verb
*no*

(o) art noun prep art adj adj noun noun aux verb verb
*yes*

2. (10 pts) For each underlined verb below, specify the conceptual dependency (CD) primitive that would best represent its meaning in the sentence. (You do not need to generate a CD representation of the entire sentence – just name the CD primitive that you would use.)

   (a) Scott sneezes whenever he hears the word "parser".
   *EXPEL*

   (b) Curtis concocted an evil plan to destroy the world.
   *MBUILD*

   (c) Alex raced to class so he wouldn't be late for the exam.
   *PTRANS*

   (d) Carlos shared his debugging knowledge with Matthew.
   *MTRANS*

   (e) Seth inherited the family farm from his great-great-great grandfather.
   *ATRANS*

   (f) Rex smelled smoke in the hallway.
   *ATTEND*

   (g) John swung the baseball bat at the pitch.
   *PROPEL*

   (h) Eli wiggled his toes to make the child laugh.
   *MOVE*

   (i) Jason learned a great deal from the science book.
   *MTRANS*

   (j) The eagle flew 200 miles to its nesting grounds.
   *PTRANS*

3. (15 pts) Suppose you want to use Yarowsky's word sense disambiguation algorithm to disambiguate between 3 different senses of the word "star": the ASTRONOMY sense (e.g., *a star in the sky*), the CELEBRITY sense (e.g., *Tom Cruise is a star*), and the SHAPE sense (e.g., *the snowflake was a 7-pointed star*). Assume that a thesaurus lists the following words for each category:

| ASTRONOMY | CELEBRITY | SHAPE |
|---|---|---|
| galaxy | famous | angle |
| moon | important | circle |
| planet | money | cluster |
| sky | performance | geometry |
| universe | reputation | polygon |

Use the 15 sentences in the box below as your "text corpus". This corpus contains exactly 200 words. The words from the thesaurus appear in boldface.

The **universe** may have been similar in composition to a neutron star.
The children drop the **circle**, square, and triangle shapes into a box.
**Geometry** tells you a lot about the **planet** and its distance from the star.
Tatum O'Neal, known for her **performance** in Paper **Moon**, got married today.
It passes bright star Spica in the eastern **sky**.
The five-pointed star is one of the most **important** symbols in history.
The Milky Way gets its name from a Greek myth about the goddess Hera who sprayed milk across the **sky**.
The hour **angle** is measured along the equator between the HC of the star and the observer's meridian.
Appearing in movies can make you **famous** and give you power.
First you get the **money**, then you get the star power.
Hubble found an **important planet** that orbits its star every 10 hours.
The **famous** star **cluster** omega Cen is the largest in the **galaxy**.
The vertices of a star are sorted by **angle**.
Broadway blitzed the showbiz world by announcing that movie star Tom Cruise will appear in a **performance** of Grease.
Volume rendering can be performed more quickly than **polygon** rendering but has the **reputation** for being slower.

Compute the Salience value of each word below for the category given. You should assume that the context window for a word spans the entire sentence containing the word but does not cross sentence boundaries. All calculations should be case insensitive (i.e., "case", "Case", and "CASE" should all be treated as the same word). Ignore punctuation marks. *Please leave your answers in fractional form and show all of your work!*

- salience("power"), with respect to CELEBRITY

  $P(power \mid CELEBRITY) = \frac{2}{8}$
  $P(power) = \frac{2}{200}$
  Salience $= \frac{\frac{2}{8}}{\frac{2}{200}} = \frac{200}{8} = 25$

- salience("power"), with respect to ASTRONOMY

  $P(power \mid ASTRONOMY) = \frac{0}{7}$
  $P(power) = \frac{2}{200}$
  Salience $= \frac{\frac{0}{7}}{\frac{2}{200}} = 0$

- salience("star"), with respect to CELEBRITY

  $P(star \mid CELEBRITY) = \frac{5}{8}$
  $P(star) = \frac{10}{200}$
  Salience $= \frac{\frac{5}{8}}{\frac{10}{200}} = \frac{1000}{80} = \frac{25}{2}$

- salience("star"), with respect to ASTRONOMY

  $P(star \mid ASTRONOMY) = \frac{5}{7}$
  $P(star) = \frac{10}{200}$
  Salience $= \frac{\frac{5}{7}}{\frac{10}{200}} = \frac{1000}{70} = \frac{100}{7}$

- salience("star"), with respect to SHAPE

  $P(star \mid SHAPE) = \frac{4}{6}$
  $P(star) = \frac{10}{200}$
  Salience $= \frac{\frac{4}{6}}{\frac{10}{200}} = \frac{800}{60} = \frac{40}{3}$

4. (12 pts)

(a) Write a script that would represent the typical experience of traveling on a commercial airplane.

*One possible script would be:*
*1. Go to airport.*
*2. Wait in ticketing line.*
*3. Get boarding pass and check luggage.*
*4. Go to gate.*
*5. Wait at gate until boarding begins.*
*6. Board plane.*
*7. Put on seat belt.*
*8. Sit in cramped seat during flight.*
*9. Depart plane.*
*10. Go to baggage claim.*
*11. Get baggage.*
*12. Go to destination.*

(b) List 3 roles that would be relevant to an airplane travel script.

*Some possible answers are:*
*pilot, flight attendants, ticket taker, passengers, security agents*

(c) List 3 props that would be relevant to an airplane travel script.

*Some possible answers are:*
*airplane, seat, seat belt, luggage, ticket, boarding pass, oxygen mask*

(d) List 3 settings that would be relevant to an airplane travel script.

*Some possible answers are:*
*parking garage, origin airport, security area (metal detectors), inside of airplane, destination airport*

5. (8 pts) For each problem below, state whether it is best characterized as an **information retrieval** task, an **information extraction** task, or a **named entity recognition** task. You can assume that each problem would be applied to an on-line archive of newspaper articles.

   (a) Identifying mentions of buildings or property that were damaged in a hurricane.

   *Information Extraction*

   (b) Identifying articles written about the Utah Jazz.

   *Information Retrieval*

   (c) Identifying mentions of colleges and universities.

   *Named Entity Recognition*

   (d) Identifying the names of companies that were acquired by another company.

   *Information Extraction*

   (e) Identifying references to currency (i.e., money amounts).

   *Named Entity Recognition*

   (f) Identifying the names of people who won a lawsuit.

   *Information Extraction*

   (g) Identifying articles that review Chinese restaurants.

   *Information Retrieval*

   (h) Identifying the names of cities.

   *Named Entity Recognition*

6. (6 pts) Consider the story below, which is originally from the San Jose Mercury news, "News of the Weird", but slightly modified for the purposes of this assignment. :)

** Police charged Gregory Rosa with a string of vending machine robberies in January when he (1) fled from police inexplicably after they spotted him loitering around a vending machine and (2) later tried to post his $400 bail in coins.*

** Karen Lee Joachimmi, a 20-year-old woman, was arrested in Florida for robbery of a Howard Johnson's motel. She was armed with only an electric chain saw, which was not plugged in.*

** A man walked into a Burger King in Ypsilanti, Michigan at 7:50am, flashed a gun and demanded cash. The clerk in the fast food restaurant said he couldn't open the cash register himself without a food order. When the man ordered onion rings, the clerk said onion rings weren't available for breakfast. The man, frustrated, walked away. Burger King called the clerk and got a good description of the man, who was later arrested.*

(a) Give a relative pronoun that appears in the story.

*Possible answers are: which, who*

(b) Give a reflexive pronoun that appears in the story.

*himself*

(c) Give a personal pronoun that appears in the story.

*Possible answers are: he, they, him, she, he*

(d) Given an example of metonymy that appears in the story.

*Burger King called*

(e) Give a definite noun phrase that has an antecedent in the story (show both the definite NP and its antecedent).

*Possible answers are: the man/a man, the clerk/the clerk, the fast food restaurant/ a Burger King*

(f) Give an appositive that appears in the story.

*Karen Lee Joachimmi, a 20-year-old woman*

7. (12 pts) The table below shows a sentence, its correct part-of-speech tags (TRUTH), and part-of-speech tags assigned to it by an initial state annotator (INIT).

| | Bo | Zo | plans | to | buy | Mary | a | hat | and | take | Mary |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TRUTH | *n* | *n* | *verb* | *inf* | *verb* | *n* | *art* | *n* | *conj* | *verb* | *n* |
| INIT | *n* | *n* | *n* | *prep* | *verb* | *n* | *art* | *n* | *conj* | *n* | *n* |
| | to | Alta | because | the | ski | runs | will | open | after | the | storm |
| TRUTH | *prep* | *n* | *conj* | *art* | *n* | *n* | *mod* | *verb* | *prep* | *art* | *n* |
| INIT | *prep* | *n* | *conj* | *art* | *n* | *verb* | *n* | *verb* | *prep* | *art* | *verb* |

(a) Consider a transformation-based learning (TBL) system that uses only this template:

Change tag X to tag Y if the previous word has tag Z.

Show all rules that would be generated from this template that would fix at least one POS tagging error.

*Change tag n to verb if the previous word has tag n*
*Change tag prep to inf if the previous word has tag n*
*Change tag n to verb if the previous word has tag conj*
*Change tag verb to n if the previous word has tag n*
*Change tag n to mod if the pervious word has tag verb*
*Change tag verb to n if the pervious word has tag art*

(b) Consider a transformation-based learning (TBL) system that uses only one template:

Change tag X to tag Y if the previous word is Z.

Show all rules that would be generated from this template that would fix at least one POS tagging error.

*Change tag n to verb if the previous word is Zo*
*Change tag prep to inf if the previous word is plans*
*Change tag n to verb if the previous word is and*
*Change tag verb to n if the previous word is ski*
*Change tag n to mod if the pervious word is runs*
*Change tag verb to n if the pervious word is the*

8. (6 pts) Consider a statistical speech recognition system that uses a bigram language model. Given some speech input, it has produced three hypotheses that it is trying to rank:

   (a) *I saw to dogs.*

   (b) *I saw two dogs.*

   (c) *I saw too dogs.*

   Show the specific bigram language model equation that would be used to estimate the likelihood of each hypothesized sentence above. Make sure to show the equation for each sentence separately. (Only the language model is relevant here, not the acoustic model.)

P(I saw to dogs) $= P(I \mid \phi) * P(saw \mid I) * P(to \mid saw) * P(dogs \mid to)$

P(I saw two dogs) $= P(I \mid \phi) * P(saw \mid I) * P(two \mid saw) * P(dogs \mid two)$

P(I saw too dogs) $= P(I \mid \phi) * P(saw \mid I) * P(too \mid saw) * P(dogs \mid too)$

9. (4 pts) In statistical approaches to NLP, *smoothing* techniques can help to generate better statistics for which of the following situations? Answer *yes* if smoothing will likely help, answer *no* if smoothing will not necessarily help. (No further explanation is necessary.)

   (a) words that did not occur in a large training corpus
   *yes*

   (b) words that occurred exactly once in a large training corpus
   *yes*

   (c) words that occurred exactly twice in a large training corpus
   *yes*

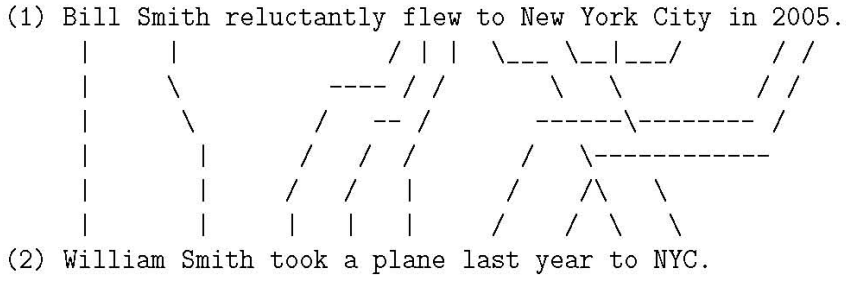
   (d) words that occurred many times in a large training corpus
   *no*

10. (12 pts total) Short-answer questions.

(a) (3 pts) What would be a good baseline method to use when evaluating a text summarization system designed to summarize news articles?

*A good baseline system would be to use the "position-based" method and just use the first N sentences of the article as a summary for that article.*

(b) (3 pts) Consider the following two sentences:




Pretend that sentence (1) and sentence (2) are translations of each other in two different natural languages. Show how the words in these sentences should be aligned by a perfect word alignment algorithm.

*The alignments are drawn above, with my lovely character graphics. :)*

(c) (3 pts) In a typical question-answering system, give an example of a general type of question that would be handled better if the system includes a good named entity recognizer, or argue that there are no such questions. *Briefly* explain your answer.

*Who, when, and where questions are general types of questions that can benefit enormously from a good NER system. In many cases, the answer to the question can be narrowed down quite a bit by limiting the candidates to be entities of the appropriate type based on the question word (Who ⇒ people, organizations, or geopolitical entities; When ⇒ times and dates; Where ⇒ locations). More sophisticated question typing can also look at the noun following the question word, for example "How many..." questions are usually asking for a number.*

(d) (3 pts) In a typical question-answering system, give an example of a general type of question that would <u>not</u> necessarily be handled better if the system includes a good named entity recognizer, or argue that there are no such questions. *Briefly* explain your answer.

*Why, What, and How questions are examples of question types that often do not benefit very much from an NER system. These questions are usually looking for answers that are complex noun phrases, clauses, or even entire sentences. These more complex answer types are typically not identified by NER systems.*

**IMPORTANT: Question #11 is for CS-6340 students ONLY!**

11. (16 pts total) This question relates to the Collins & Singer bootstrapping method for named entity recognition. The predicate Contains($w$) is satisfied if a sequence of words contains the word $w$.

**TABLE 1**

| NP | CONTEXT | CLASS |
|---|---|---|
| Michael Jordan | basketball player | PERSON |
| Apple Computer | a computer company | COMPANY |
| Jeff Jordan | chief executive officer | PERSON |
| Jeff Jordan | president of PayPal | PERSON |
| Jeff Citron | chairman and CEO of Vonage | PERSON |
| Jeff Citron | a CEO | PERSON |
| Jordan | small country | LOCATION |
| Apple | Gwyneth's child | PERSON |

(a) (8 pts) Using only the Contains($w$) predicate, fill in the table below with all of the rules that would be generated from the NPs in TABLE 1 and compute the probability of each rule. Leave the probabilities in fractional form!

| NP Rule | Probability |
|---|---|
| If Contains("Michael") → PERSON | 1/1 |
| If Contains("Jordan") → PERSON | 3/4 |
| If Contains("Apple") → COMPANY | 1/2 |
| If Contains("Computer") → COMPANY | 1/1 |
| If Contains("Jeff") → PERSON | 4/4 |
| If Contains("Citron") → PERSON | 2/2 |
| If Contains("Jordan") → LOCATION | 1/4 |
| If Contains("Apple") → PERSON | 1/2 |

(b) (3 pts) Using only the NP rules in your previous table that have a probability of 1.0, apply those rules to the instances in TABLE 2 and fill in TABLE 2 with the class that would be assigned to each instance. If no class would be assigned to an example, simply put *none*.

**TABLE 2**

| NP | CONTEXT | CLASS |
|---|---|---|
| Jeff Jones | CEO | PERSON |
| Citron Inc | car company | PERSON |
| River Jordan Inc | internet corporation | NONE |
| Maine Apple Orchard | family farm | NONE |
| Dell Computer Inc | computer manufacturer | COMPANY |

(c) (5 pts) Using only the Contains($w$) predicate, fill in the table below with all of the **Context Rules** that would be generated from the instances in TABLE 2 and compute the probability of each rule based on the instances in **both** TABLE 1 and TABLE 2. Leave the probabilities in fractional form!

| **Context Rule** | **Probability** |
|---|---|
| If Contains("CEO") → PERSON | 3/3 |
| If Contains("car") → PERSON | 1/1 |
| If Contains("company") → PERSON | 1/2 |
| If Contains("computer") → COMPANY | 2/2 |
| If Contains("manufacturer") → COMPANY | 1/1 |