NAME:

CS-5340/6340, Solutions to Final Exam, Fall 2013

1. (18 pts) Consider the following short story:

> *John, Mary, and their two children drove to the zoo. They laughed at the crazy antics of the monkeys. One monkey was throwing bananas in the air. Another monkey was swimming in circles. Next, they watched snakes and lizards in the reptile house. Several snakes flicked their tongues against the glass, and a huge lizard sneezed in front of Tommy. The kids asked their parents to buy them food, which they ate near the elephant exhibit. Unexpectedly, one elephant grabbed a pumpkin, swallowed it whole, and then trumpeted loudly. Several nearby deer ran away when they heard the noise. Later in the afternoon, a zookeeper taught the kids that an angry camel may spit on you. That night, little Tommy dreamed about crazy monkeys, giant lizards, hungry elephants, and angry camels.*

For each conceptual dependency (CD) primitive ACT below, list <u>all</u> of the verbs in the story that would be represented by that CD primitive based on their use in the story. Each verb should be listed only once! Choose the CD primitive the best captures the meaning of the verb in the story above.

(a) INGEST

ate, *swallowed*

(b) EXPEL

sneezed, *spit*

(c) MOVE

flicked

(d) GRASP

grabbed

(e) PROPEL

throwing

(f) PTRANS

drove, *swimming*, *ran*

(g) ATRANS

buy

(h) MTRANS

asked, *taught*

(i) MBUILD

dreamed

(j) ATTEND

heard, *watched*

(k) SPEAK

laughed, *trumpeted*

2. (10 pts) For this question, you must define a case frame structure for a specific verb. The case frame should contain mappings between syntactic roles and thematic roles, but you do not need to include selectional restrictions. You do not need to fill in the case frame with the noun phrases in the sentences. Just create a case frame structure that would correctly assign thematic roles to all of the noun phrases in the sentences.

   (a) Create a *single* case frame structure for the verb "bought" that will assign the correct thematic role to all of the noun phrases in the sentences below.

   George bought a laptop for his daughter.
   Bill Gates bought 12 vacation homes with his credit card.
   Julie bought her son a new bicycle.

   BOUGHT (Active)
   Agent = subject
   Theme = direct object
   Recipient = PP(for)
   Recipient = indirect object
   Instrument = PP(with)

   (b) Create a *single* case frame structure for the verb "written" that will assign the correct thematic role to all of the noun phrases in the sentences below.

   His book was written with his wife.
   The document was written with a fountain pen by Mary Smith.
   The poem was written by Susan for her husband.

   WRITTEN (Passive)
   Theme = subject
   Co-Agent = PP(with)
   Instrument = PP(with)
   Agent = PP(by)
   Beneficiary = PP(for)

3. (16 pts) Consider the following 4 (short) documents:

D1: *natural language processing involves processing natural language texts*
D2: *spoken language understanding involves processing spoken natural language*
D3: *NLP involves ambiguity*
D4: *NLP has ambiguity*

(a) (10 pts) Create an inverted file representation containing all of the words in the 4 documents above.

natural: D1, D2
language: D1, D2
processing: D1, D2
involves: D1, D2, D3
texts: D1
spoken: D2
understanding: D2
NLP: D3, D4
ambiguity: D3, D4
has: D4

(b) (3 pts) Compute TF-IDF("spoken",D2), which is the TF-IDF weight that would be given to the term "spoken" for D2. Use log base 2. Show all your work!

TF("spoken",D2) = 2
DF("spoken") = 1
IDF("spoken") = $log_2(4/1) = log_2(4) = 2$
TF-IDF("spoken", D2) = 2 * 2 = 4

(c) (3 pts) Compute TF-IDF("NLP",D3), which is the TF-IDF weight that would be given to the term "NLP" for D3. Use log base 2. Show all your work!

TF("NLP",D3) = 1
DF("NLP") = 2
IDF("NLP") = $log_2(4/2) = log_2(2) = 1$
TF-IDF("NLP", D3) = 1 * 1 = 1

4. (10 pts) Consider the following two sentences (the first is a quote from Dr. Seuss):

S1: I meant what I said and I said what I meant
S2: he meant what he said yesterday

	and	I	he	meant	said	what	yesterday
S1	1	4	0	2	2	2	0
S2	0	0	2	1	1	1	1

(a) Assume that the S1 row of the table above is a feature vector representation of sentence S1. There are 7 unigram features represented by the 7 columns. Fill in frequency-based values for this feature vector representation in row S1.

(see table above)

(b) Assume that the S2 row of the table above is a feature vector representation of sentence S2. There are 7 unigram features represented by the 7 columns. Fill in the frequency values for this feature vector representation in row S2.

(see table above)

(c) Compute the Manhattan Distance between the S1 feature vector and the S2 feature vector. Show your work.

Manhattan Distance(S1,S2) = 1 + 4 + 2 + 1 + 1 + 1 + 1 = 11

(d) Compute the Jaccard Distance between the S1 feature vector and the S2 feature vector. Show your work.

Jaccard(S1,S2) = (0+0+0+1+1+1+0) / (1+4+2+2+2+2+1) = 3/14

5. (8 pts) Consider the following sentences:

Mary went to the store by her school.
John flew to Boston.
George swam in the river in Boston.
Susan moved to Utah.
Lee gave a donation to charity.
The boy went to a party in Idaho.

Assume that each prepositional phrase (PP) attaches to the closest preceding noun or verb. For example, "by her school" attaches to "store" and not "went".

Use the PP attachments in these sentences to compute the following probabilities. P(VERB | $prep_i$) is the probability that a PP with the preposition $prep_i$ attaches to a VERB. P(NOUN | $prep_i$) is the probability that a PP with the preposition $prep_i$ attaches to a NOUN. **Leave your answers in fractional form!**

(a) P(VERB | "in")

1/3

(b) P(NOUN | "in")

2/3

(c) P(VERB | "to")

4/5

(d) P(NOUN | "to")

1/5

6. (9 pts) Imagine that you have a text corpus of size 100 (i.e., it contains 100 word instances). This corpus contains 20 instances of "car", 10 instances of "gas", and 5 instances of "wheel". There are 4 documents that contain both the word "car" and the word "gas". [NOTE: additional clarifications were given in class indicating that the instances of car, gas, and wheel occurred in different documents.]

   (a) Compute P("gas")

   P("gas") = 10/100 = .10

   (b) Compute P("gas" | "car") , which is the probability that a document contains "gas" given that it contains "car".

   P("gas" | "car") = 4/20 = .20

   (c) Compute PMI("car", "gas"), where PMI is point-wise mutual information based on whether two terms occur in the same document. [NOTE: additional clarifications were given in class allowing N to be used to represent the number of documents in the corpus.]

   Both of the following answers were deemed acceptable:
   $log_2(\frac{\frac{4}{N}}{\frac{20}{N} * \frac{10}{N}})$

   or
   $log_2(\frac{\frac{4}{100}}{\frac{20}{100} * \frac{10}{100}}) = log_2(\frac{\frac{1}{25}}{\frac{1}{5} * \frac{1}{10}}) = log_2(2) = 1$

   (d) Does the PMI value that you computed indicate that "car" and "gas" are statistically dependent or independent?

   Having a log value greater than zero means that the terms are statistically dependent.

7. (24 pts) Give a short answer (1-2 sentences) for each question below.

   (a) Why are "factoid" questions usually easier for NLP systems to answer than other types of questions?

   Because factoid questions usually have a well-defined type of answer, which are often short phrases that can be recognized by named entity recognizers (e.g., people or location names)

   (b) Consider the sentence "Mary kicked the ball". Would a conceptual dependency (CD) representation of this sentence's meaning contain the same information as a case frame representation of this sentence's meaning? If not, explain what information would be explicitly captured by one representation but not the other.

   A CD representation would represent kicking as a PROPEL act along with a MOVE act with object=foot to represent the meaning of "kick" as propelling something forward by moving one's foot. A case frame representation would not explicitly contain any reference to a foot.

   (c) Give one reason why it may be more important to use smoothing for a bigram language model than for a unigram language model.

   In general, longer n-grams will be more sparse, which means that they will have lower frequencies, and many possible n-grams may never appear (i.e., have zero frequency) in a text corpus. Smoothing has a big impact for these cases.

   (d) Is processing spoken natural language different from processing written natural language, or are they essentially the same problem?

   Spoken language processing is substantially different from processing written text because spoken language contains disfluencies, has no punctuation, has no capitalization, etc.

   (e) Give one advantage of using stemming instead of morphological analysis.

   Stemming can be done without relying on a dictionary of root words and their parts-of-speech.

   (f) Give one advantage of using morphological analysis instead of stemming.

   Morphological analysis is generally more accurate than stemming because knowledge of the root forms and irregular words is used.

(g) Statistical machine translation (MT) systems often use a parallel text corpus. Is the parallel corpus used for the translation model component or the language model component? Briefly explain.

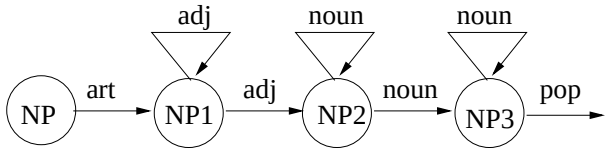
The translation model component uses the parallel corpus to obtain knowledge about common mappings between words in the source and target languages. The language model component only requires texts in the target language to help generate natural-sounding word sequences. So the language model could be generated from the target language texts in the parallel corpus, but the language model only needs texts in the target language.

(h) If I ask the question "Can you open the door?" as a request for you to open the door for me, is this a direct or indirect speech act? Briefly explain.

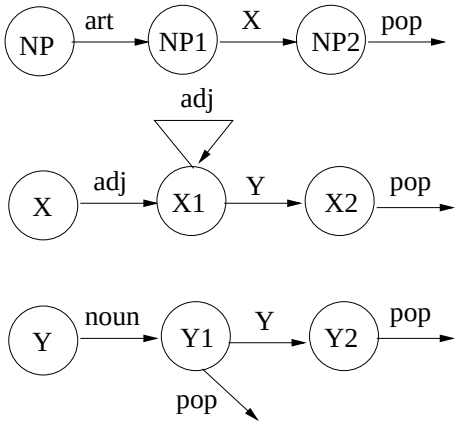
This is an indirect speech act because the literal interpretation of the question asks about your ability to open the door, but the intent of the question was a request for you to open the door. So the literal meaning is different from the intended meaning.

8. (5 pts) Indicate whether the two recursive transition networks (RTNs) below recognize exactly the same language, or whether they recognize different languages. If they recognize different languages, then give one example of a string that would be accepted by one RTN but not the other (and indicate which RTN would accept it).

#1



#2



The RTNs accept exactly the same language.

Question #9 is for CS-6340 students ONLY!

9. (10 pts) Indicate whether each sequence of part-of-speech tags would be accepted by the grammar below or not.

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) art noun modal verb

yes

(c) art noun prep noun modal verb

yes

(d) art noun prep noun modal verb adv

yes

(e) art noun prep noun modal verb adv adv

no

(f) noun prep noun modal adv adv verb

yes

(g) noun prep noun verb adv

no

(h) noun noun noun prep noun noun aux verb

yes

(i) noun noun prep noun aux adv verb adv

yes

(j) noun noun prep noun aux aux adv verb

no