University of Toronto, Department of Computer Science

CSC 2501/485F—Computational Linguistics, Fall 2015

Assignment 3

Due date: 13:10, Wednesday 4 November 2015, on CDF.

This assignment is worth 12% of your final grade.

- Fill out both sides of the assignment cover sheet, and staple together all answer sheets (in order) with the cover sheet (sparse side up) on the front. (Don't turn in a copy of this handout.)
- Please type your answers in no less than 12pt font; diagrams and tree structures may be drawn with software or neatly by hand.
- What you turn in must be your own work. You may not work with anyone else on any of the problems in this assignment, except for discussion in very general terms. If you need assistance, contact the instructor or TA.
- Any clarifications to the problems will be posted on the course website announcements page. You will be responsible for taking into account in your solutions any information that is posted there, or discussed in class, so you should check the page regularly between now and the due date.

Copyright © 2015 Frank Rudzicz, Graeme Hirst, and Suzanne Stevenson. All rights reserved.

1. Verb complements and gap features [20 marks]

Gap features enable us to make sure that the arguments to a verb are associated with the right syntactic positions to get the right thematic roles from the verb. For example, in the grammar for the passive construction that we saw in class, we can associate the NP subject with the object position through the "gap" feature, so that, in the semantics, the subject NP can be interpreted as the Theme of the verb.

There are other situations in which NPs are interpreted as if they occurred in positions in which they do not occur in the overt sequence of words. Consider the following sentences:

i. The student preferred to sleep.
ii. The student persuaded the teacher to sleep.
iii. The student promised the teacher to sleep.
iv. The student expected the teacher to sleep.

For example, in (i), *the student* is the Agent of *preferred*, as we would expect by it being in the subject position of *preferred*; but interestingly in (i), *the student* is also the Agent of *sleep*, even though it does not appear to occur in the subject position of *sleep*.

In answering the three questions on the next page, assume the following:

- Only Agent and Theme roles will be used in this question.
- Each NP **position** (subject or object) can be associated with only one thematic role from a verb.
- A subject is the Agent and an object is the Theme of the verb. (This is the mapping of thematic roles to positions that we saw in class.)
- An embedded clause (i.e., a complement that is a verb phrase or sentence) acts as a noun phrase and receives the Theme role from the verb taking it as a complement.

A. (3 marks) For each of the sentences (ii)-(iv) above, state each thematic role that the main verb of the sentence gives to other constituents of the sentence.

Hint: The treatment of the NP *the teacher* is different in each sentence.

B. (14 marks) Devise a context-free grammar augmented with features for the above types of sentences using the verbs *preferred*, *persuaded*, *promised*, and *expected*. Be sure to give the necessary lexical entries for these four verbs, as well as *sleep*.

Use only the features necessary to ensure the appropriate interpretation of NPs with respect to semantic roles.

C. (3 marks) Draw a parse tree for sentence (ii) above (*The student persuaded the teacher to sleep*). Annotate each constituent with the features assigned by your grammar.

2. Corpus-based disambiguation [20 marks]

In this problem, you will apply and analyze Ratnaparkhi's unsupervised method for resolving PP attachment, using a small corpus.

Note: You do not have to compute all the possible statistics based on the corpus in order to complete this problem. Compute only the statistics needed to answer the questions below.

Note: You may solve this problem by hand or by writing a short program; the former is probably faster, but the files `pp-corpus` and `wordlist` are available on the course web page if you want them.

A. (5 marks) Apply Ratnaparkhi's extraction heuristic to extract all "unambiguous" tuples from the corpus `pp-corpus` (page 6 of this handout). Please note the following:

- The corpus `pp-corpus` displays one sentence per line of text, except for a couple of long sentences that continue on a second line (indicated by indentation).
- Since `pp-corpus` is not PoS tagged, the file `wordlist` gives you the part-of-speech for each word. There are no words with more than one possible PoS tag.
- The sentences have been "morphologically processed" and "chunked", and thus the extraction heuristic should apply directly to the sequence of words as they appear in the corpus.
- In your extraction procedure, use $K = 2$.
- Don't extract $n2$, only $[n, p]$ and $[v, p]$ pairs.

Present your results as a list (one per line) of each "unambiguous" $[n, p]$ and $[v, p]$ pair, in the order that they **first** would be extracted, and with their total count. For example, for this two-line corpus:

```
antelopes merrily run onto sidewalks
antelopes on sidewalks awkwardly run onto grass
```

your answer would be as follows:

```
2  run onto
1  antelopes on
```

B. (5 marks) Apply formula (1) in Section 4 of Ratnaparkhi's paper, using the estimations in Section 4.1 and 4.2.1 (bigram counts) of the required probabilities, to disambiguate the following PP attachment, and state what is the preferred attachment.

```
swam whales onto
```

Show your work: Make clear what the component probabilities are and exactly how you are estimating them. (This means that even if you extracted the tuples incorrectly in part A, you can still get credit for knowing how to apply the formulas.)

C. (10 marks) Ratnaparkhi's formulas in Sections 4.1 and 4.2.1 include backoff specifications. In spite of this, the probabilities of verb attachment for the following two examples are both 0.

```
placed seals onto
advanced whales for
```

i. State precisely, for each of these examples, why the verb attachment probability is 0, even though there are backoff formulas stated in 4.1 and 4.2.1.

ii. If, instead of using the estimates in 4.2.1, we used the smoothed estimates in 4.2.2, would this solve the problem(s) (i.e., in either or both cases, would we then get a non-zero probability)?

iii. Under exactly what condition(s) will the backoff formulas be used in an attachment probability estimate that is not zero?

```
wordlist:

N   fish                          Adv  carefully
N   whales                        Adv  happily
N   seals
N   otters                        P   by
N   we                            P   onto
                                  P   toward
V   swam                          P   beside
V   hurried                       P   for
V   placed                        P   among
V   advanced
V   love
V   admire
```

otters hurried
whales hurried carefully for otters among fish by otters
seals placed seals onto seals
whales hurried whales beside fish by fish for whales
seals placed seals among whales onto seals
otters we admire placed otters onto whales
otters we admire swam carefully onto whales
otters placed seals onto otters
otters among seals advanced toward whales
otters placed seals onto otters
seals among whales placed seals onto seals onto otters onto whales
whales placed fish beside whales onto whales by whales
seals among fish swam by otters
fish beside seals onto whales swam carefully by fish by whales beside whales beside fish
otters hurried for whales
otters we admire hurried otters onto otters for fish
seals placed happily whales onto fish beside whales
seals swam carefully onto otters
fish by seals placed whales onto seals onto fish beside otters
seals swam by fish
otters hurried for seals
otters among seals hurried fish beside seals onto otters among whales for whales beside whales
otters onto fish hurried whales for fish
seals among whales hurried happily
whales hurried happily fish for fish
seals swam
otters onto otters hurried carefully whales for whales
fish advanced beside fish
whales advanced
seals we love placed otters onto fish
seals hurried carefully whales for otters onto otters
whales by otters hurried for otters
otters swam onto whales
fish swam by whales
fish swam
otters swam carefully by fish
seals we love hurried for fish by whales by whales
seals placed whales beside whales beside fish onto seals
whales advanced otters among whales toward whales
whales advanced toward fish

3. Playing with WordNet [10 marks]

Read section 2.5 of Bird *et al* (pages 67–73 of the printed edition) for a quick introduction to the NLTK interface to WordNet, and try out some of the examples. For a quick overview of the content of WordNet, you might also find Princeton's Web interface to be useful: `wordnetweb.princeton.edu/perl/webwn`

A. (4 marks) In WordNet, a leaf is a synset that has no hyponyms, and the depth of any synset is defined to be the length (in edges) of the shortest path from the root to that synset. Write a short program to find the depth of the shallowest and deepest leaves in the noun hierarchy of WordNet. Print out the depths and an example of a leaf at each of these depths. Finally, compute and print out the ratio of leaves to all synsets in the noun hierarchy.

B. (6 marks) Do exercise 28 of section 2.8 of Bird *et al*, using the `path_similarity` function. Compare your ranking with the human norms established by Miller and Charles and suggest an explanation for any notable differences.

Hint: There are two `path_similarity` functions. One is a member of the WordNet module; it takes two synsets as arguments. The other (shown on page 72 of the book) is a member of the class `synset`; it takes one synset and returns the distance to `self`.

Hint: You are given pairs of *words*, but semantic similarity is defined on *senses* or synsets. Informally, we say that the semantic similarity of two words is that of their two closest senses. For example, the pair *bank–money* is relatively close because of the financial senses of *bank* and *money*, even though money has nothing to do with river banks.

Hint: Save typing; the word-pairs are available in the online version of the NLTK book and also on the course web page, `www.cs.utoronto.ca/~gh/2501/`.