Final Exam
Natural Language Processing

June 5, 2017

Remember to fill in your name on all pages
Good luck

Problem 1 (1 point). For each of the following Jeopardy-style questions, find the best answer.

1. He developed a new theory of language, in which natural language was seen as a structure of mutually linked elements, similar or opposed to each other.
2. He wrote the "Translation" memorandum that brought the idea of Machine Translation to general notice and inspired many projects.
3. He wrote the phrase "Colorless green ideas sleep furiously" in his book as an example of a sentence that is grammatically correct, but semantically nonsensical.
4. The quote "One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know." belongs to him.

Hint: the possible answers are Noam Chomsky, Ferdinand de Saussure, Andrew D. Booth, Warren Weaver, Groucho Marx and Leonard Bloomfield.

Problem 2 (2.5 points). Assume you are given the task to pick up the most probable correction for the word *steming* among two possible candidates: *stemming* and *stewing* by using a Noisy channel model. For this, you will have to compute $P(x|w) * P(w)$ with $x =$ *steming* and w either *stemming* or *stewing*.

The language model used is a Good-Turing model based on the following corpus (make sure you include punctuation signs in your calculations).

Stew is the word obtained from stewing when stemming is performed. Brais is the word obtained from braising when you perform stemming.

The channel model is build based on the following list of typical errors.

accommodations: accomodations	*arrow: arow*	*asymmetrical: assymmetrical*
browse: bbrose	*bomb: bown*	*condemn: conden*
foremost: formost	*incremented: increented*	*recommend: recommened*
snowing: shoving	*sweetest: sweetes*	*swivel: swival*
want: mant	*whip: wip*	*wind: win*
woods: woodes	*wrong: wrang*	

Problem 3 (2 points). Suppose we have the following short movie reviews, each labeled with a genre, either **comedy** or **action**.

1. *fun, couple, love, love* (**comedy**)
2. *fast, furious, shoot* (**action**)
3. *couple, fly, fast, fun, fun* (**comedy**)
4. *furious, shoot, shoot, fun* (**action**)
5. *fly, fast, shoot, love* (**action**)

Using a Binarized Multinomial Naive Bayes approach with Laplace smoothing, how would we classify the following review: *fun, love, fun, shoot*? (use $\alpha = 2$ in your calculations)

Problem 4 (2.5 points). Let $G = (V, C, \Sigma, \mathrm{S}, L, R, P)$ be a probabilistic context free grammar with $V = \{\mathrm{S}, \mathrm{NP}, \mathrm{VP}, \mathrm{PP}\}$, $C = \{\mathrm{N}, \mathrm{V}, \mathrm{P}\}$, $\Sigma = \{\text{people, fish, tanks, rods, with}\}$ and L, R given below, with their corresponding weights.

The set R of productions:

$\mathrm{S} \to \mathrm{NP\ VP}$	0.9
$\mathrm{S} \to \mathrm{VP}$	0.1
$\mathrm{VP} \to \mathrm{V\ NP}$	0.5
$\mathrm{VP} \to \mathrm{V}$	0.1
$\mathrm{VP} \to \mathrm{V\ NP\ PP}$	0.3
$\mathrm{VP} \to \mathrm{V\ PP}$	0.1
$\mathrm{NP} \to \mathrm{NP\ NP}$	0.1
$\mathrm{NP} \to \mathrm{NP\ PP}$	0.2
$\mathrm{NP} \to \mathrm{N}$	0.7
$\mathrm{PP} \to \mathrm{P\ NP}$	1.0

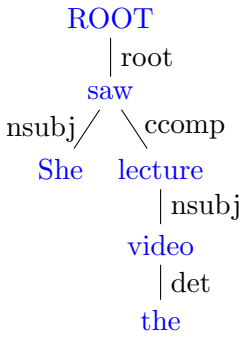
The lexicon L:

$\mathrm{N} \to$ people	0.5
$\mathrm{N} \to$ fish	0.2
$\mathrm{N} \to$ tanks	0.2
$\mathrm{N} \to$ rods	0.1
$\mathrm{V} \to$ people	0.1
$\mathrm{V} \to$ fish	0.6
$\mathrm{V} \to$ tanks	0.3
$\mathrm{P} \to$ with	1.0

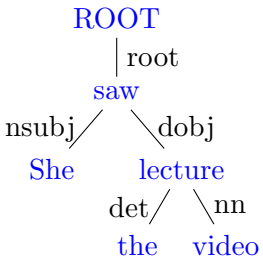
Apply the extended Cocke-Younger-Kasami (CYK) algorithm and find out whether *fish people fish tanks* is a sentence accepted by the grammar G (argument your answer). In the affirmative case, draw the derivation tree of highest probability.

Problem 5 (1 point). Compute the accuracy (both labeled and unlabeled) for the following parsing of the sentence *She saw the video lecture*.

Parsed:



Gold:



Problem 6 (1 point). For each of the following types of relations, find two words w_1 and w_2 that are in that relation.

- w_1 and w_2 are homophones,
- w_1 and w_2 are synonyms,
- w_1 and w_2 are antonyms,
- w_1 is a hyponym of w_2,
- w_1 is an instance of w_2.

DEFINITION 1. No definition is necessary for this exercise.

DEFINITION 2. Let w be a word that appears c times in the corpus, N_c = the count of things we've seen c times, and N = the total number of tokens in the corpus. Then,

$$c^*(w) = \begin{cases} \frac{(c+1)N_{c+1}}{N_c}, & \text{if } c > 0 \\ N_1, & \text{if } c = 0 \end{cases}$$

$P^*_{GT}(w) = \frac{c^*(w)}{N}$

Computing error probability:

$del[x, y]$: count(xy typed as x) $sub[x, y]$: count(x typed as y)

$ins[x, y]$: count(x typed as xy) $trans[x, y]$: count(xy typed as yx)

$$P(x \mid w) = \begin{cases} \frac{del[w_{i-1},w_i]}{count(w_{i-1}w_i)}, & \text{if } x = w_1 \dots w_{i-1}w_{i+1} \dots w_n \text{ and } w = w_1 \dots w_{i-1}w_iw_{i+1} \dots w_n \\ \frac{ins[w_{i-1},x_i]}{count(w_{i-1})}, & \text{if } x = w_1 \dots w_{i-1}x_iw_i \dots w_n \text{ and } w = w_1 \dots w_{i-1}w_i \dots w_n \\ \frac{sub[w_i,x_i]}{count(w_i)}, & \text{if } x = w_1 \dots w_{i-1}x_iw_{i+1} \dots w_n \text{ and } w = w_1 \dots w_{i-1}w_iw_{i+1} \dots w_n \\ \frac{trans[w_i,w_{i+1}]}{count(w_iw_{i+1})}, & \text{if } x = w_1 \dots w_{i+1}w_i \dots w_n \text{ and } w = w_1 \dots w_iw_{i+1} \dots w_n \end{cases}$$

DEFINITION 3. Let S be a sentence and $w_1, \dots, w_k$ its words after removing possible duplicates, C a class of possible labels and V the vocabulary. Then,

$$c_{BMNB}(S) = argmax_{c_j \in C} P(c_j) \prod_{i=1}^{k} P(w_i \mid c_j),$$

where the prior probability of the class is $P(c_j) = doccount(c_j)/N_{doc}$, and the probability of a word w_i given the class c_j, computed using Laplace smoothing is:

$$P(w_i \mid c_j) = \frac{n_i + \alpha}{n + \alpha|V|}$$

with n_i = the number of occurrences of w_i in $Text_j$, n = the number of tokens in $Text_j$, and $Text_j$ the text obtained by concatenating all documents with label c_j in which all duplicates are removed.

DEFINITION 4. No definition is necessary for this exercise.

DEFINITION 5. The accuracy of a dependency parser is computed as the percentage of correct dependencies with respect to the total number of dependencies.

DEFINITION 6. No definition is necessary for this exercise.