תאריך : 12.3.2020

מבחן מועד ב' בקורס עיבוד שפה טבעית (67658)

2019/2020 שנת לימודים התש"ף

משך המבחן : שעתיים

מרצה הקורס : ד"ר עמרי אבנד

השימוש בכל חומר עזר אסור, נא כתבו על הכריכה על אילו שאלות עניתם.

חלק א' (80 נקודות): ענו על <u>בדיוק שתי שאלות</u> מבין שאלות 1-3

<u>שאלה 1 (40 נק'):</u>

א. (10 נק') הגדירו פורמלית מהו מודל שפה מרקובי מסדר שני (Trigram Markov Language Model).

ב. (15 נק') הגדירו פורמלית מהי שיטת ההחלקה קנזר-ניי (Kneser-Ney) עבור מודל שפה מרקובי מסדר ראשון (Bigram Markov Language Model). רשמו בצורה מדויקת את כל הנוסחאות המעורבות בשיטה.

ג. (7 נק') מהי המגבלה של מודלי שפה מרקוביים (Markov Language Models) המודגמת ע"י המשפט *colorless green ideas sleep furiously*? נמקו.

ד. (8 נק') האם מודל שפה המבוסס על RNN (recurrent neural network) יסבול מהמגבלה שציינתם בסעיף ג'? נמקו.

<u>שאלה 2 (40 נק'):</u>

א. (5 נק') הסבירו מהי בעיית דלילות הנתונים (sparsity) בהקשר של מודל מרקובי חבוי (HMM). לאיזה אפקט <u>לא רצוי</u> היא עשויה לגרום?

ב. (5 נק') האם הבעיה מסעיף א' תקפה באותה המידה הן להסתברויות הפליטה (emission probabilities) והן להסתברויות המעברים (transition probabilities) או שהיא תקפה <u>במיוחד</u> לאחד מהם?

ג. (10 נק') הסבירו מהי שיטת ההחלקה של HMM באמצעות פסאודו-מילים (pseudowords). לאיזה אפקט <u>רצוי</u> ולאיזה אפקט <u>לא רצוי</u> היא עשויה לגרום?

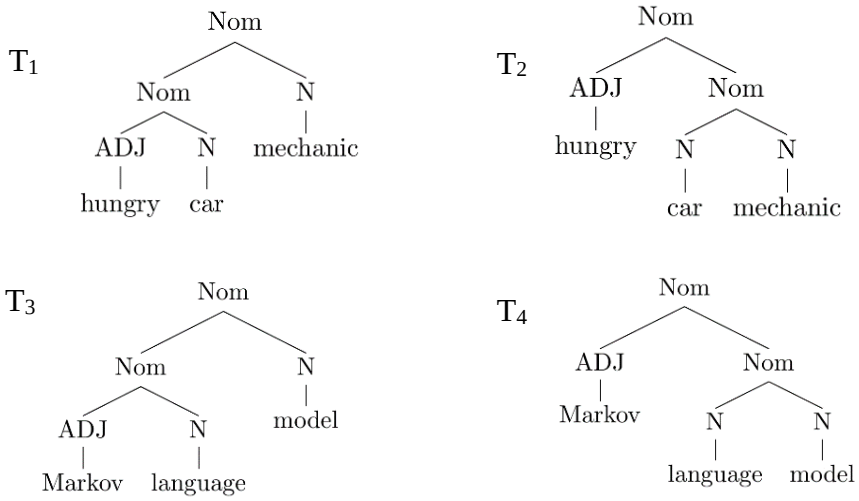
ד. (10 נק') נניח כעת שנשנה את מודל ה-Bigram HMM עבור משימת תיוג חלקי דיבר (POS tagging) כך שהסתברויות המעברים יהיו בהכרח אחידות (uniform). כלומר כל הסתברויות המעברים יהיו שוות לערך אחד קבוע. רשמו בצורה מדויקת ומפורטת כיצד שינוי זה ישפיע על התיוגים שהמודל ייצר? נמקו.

ה. (10 נק') נניח כעת שנשנה את המודל מסעיף ד' כך שהסתברויות הפליטה יהיו בהכרח אחידות. כלומר כל הסתברויות הפליטה יהיו שוות לערך אחד קבוע. רשמו בצורה מדויקת ומפורטת כיצד שינוי זה ישפיע על התיוגים שהמודל ייצר? נמקו.

שאלה 3 (40 נק׳):

א. (10 נק׳) הגדירו פורמלית את מודל Probabilistic Context-free Grammar (PCFG).

נתונים זוגות הניתוחים (parses) האפשריים הבאים עבור הביטוי "hungry car mechanic" ועבור הביטוי "Markov language model":



ב. (15 נק׳) הראו שההסתברויות שמודל PCFG (ללא לקסיקליזציה) יעניק לניתוחים אלו בהכרח מקיימות את תנאי (i) **או** את תנאי (ii). הניחו שהסתברויות כל הניתוחים חיוביות.

i. $Pr(T_1) \geq Pr(T_2)$ וגם $Pr(T_3) \geq Pr(T_4)$

ii. $Pr(T_1) \leq Pr(T_2)$ וגם $Pr(T_3) \leq Pr(T_4)$

ג. (5 נק׳) הסבירו מדוע התכונה שהוכחתם בסעיף ב׳ איננה רצויה.

ד. (10 נק׳) כיצד ניתן לשנות את המודל מסעיף ב׳ כך שהתכונה שהגדרתם בסעיף ב׳ לא תתקיים? נמקו.

חלק ב׳ (20 נקודות): ענו על <u>בדיוק שאלה אחת</u> מבין שאלות 4-5

שאלה 4 (20 נק׳):

א. (10 נק׳) הגדירו מהו beam search בהקשר של transition-based parsing.

ב. (10 נק׳) מהו היתרון האפשרי משימוש ב-beam search בניתוח תלויות (dependency parsing) על פני שימוש בגישה חמדנית (greedy)?

שאלה 5 (20 נק׳):

א. (10 נק׳) בהקשר של שיכוני מילים (word embeddings), מהו ההבדל בין מודלים המבוססים על ספירות (count-based models) ומודלים המבוססים על חיזוי (prediction-based models)?

ב. (10 נק׳) מהי גישת ה-distant supervision ל-relation extraction? רשמו אלגוריתם סכמתי (schematized algorithm) לאימון מודל ל-relation extraction תחת גישה זו.

בהצלחה!