תאריך: 13.2.2018

מבחן מועד א' בקורס עיבוד שפה טבעית (67658)

2017/2018 שנת לימודים התשע"ח

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 נקודות):</u>

א. (30 נק') נתבונן במודל Bi-gram Conditional Random Field (CRF):

$$p(y_1 \cdots y_N | x_1 \cdots x_N) = \frac{\prod_{j=1}^{N} \mathrm{e}^{w \cdot f(y_{j-1}, y_j, j, x_1 \cdots x_N)}}{Z(x_1 \cdots x_N; w)}$$

הניחו ש-$y_1,\ldots,y_N$ מקבלים ערכים בקבוצת התוויות (labels) Y, וש-y_0=START.

כתבו פסאודו-קוד שמקבל כקלט פונקציית תכוניות f (feature function), וקטור משקולות w, ומשפט הנתון כסדרה של מילים $x_1,\ldots,x_N$, ומחזיר את ההתפלגות

$$p(y_1, y_2, y_3 | x_1 \cdots x_N)$$

לכל ערך אפשרי של y_1,y_2,y_3 ב-Y.

ב. (10 נק') הסבירו מהו ה-label bias של Maximum Entropy Markov Models (MEMM) לתיוג סדרות (sequence labeling). הסבר מדוע מודל CRF אינו סובל מה-label bias.

<u>שאלה 2 (40 נקודות):</u>

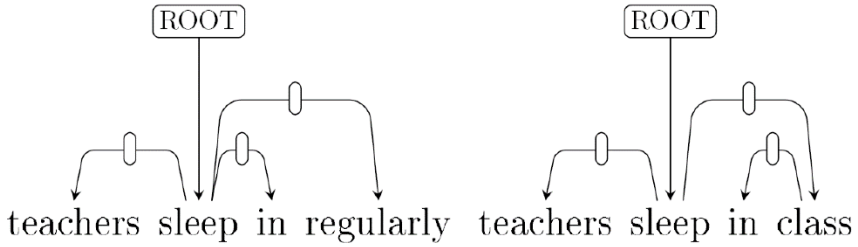
א. (20 נק') עבור מודל שפה בי-גרם (Bi-gram Language Model, מודל שפה מרקובי מסדר ראשון), הגדר את מודל השפה המתקבל ע"י אינטרפולציה לינארית שלו עם מודל Unigram (Linear Interpolation Backoff Model with a Unigram Model). בתשובתכם, רשמו את כל הנוסחאות הנחוצות לחישוב הפרמטרים של מודל האינטרפולציה, ואת בעיית האופטימיזציה המגדירה את מקדמי האינטרפולציה. **אין** צורך להסביר כיצד לפתור את בעיית האופטימיזציה.

ב. (10 נק') איזו הנחת אי-תלות מותנה מניח מודל שפה מרקובי (Markov Language Model)? הסבירו כיצד הנחה זו מופרת בשפה טבעית. לוו את ההסבר שלכם בדוגמא.

ג. (10 נק') כיצד מודל שפה המבוסס על Recurrent Neural Networks מתמודד עם הקושי שציינתם בסעיף ב'?

שאלה 3 (40 נקודות):

א. (10 נק') התבוננו בשני המשפטים הבאים ובעצי התלויות (dependency trees) הנכונים שלהם:



עבור כל אחד מהמשפטים, רשום את סדרת המעברים (sequence of transitions) אשר מניבה את עץ התלויות שלו, ע"פ מערכת המעברים arc standard (ראו נספח).

ב. (15 נק') רשמו פסאודו-קוד לפרסר תלויות מבוסס-מעברים חמדני (greedy transition-based dependency parser), המשתמש במערכת המעברים arc standard. הניחו שהפרסר הוא unlabeled, כלומר הוא פולט עצים מכוונים ללא תוויות על הצלעות.

על הפסאודו-קוד לקבל משפט (הנתון ע"י סדרה של מילים), ו-handle עבור transition classifier, ולפלוט עץ.

אין צורך להסביר כיצד לאמן את הפרסר.

ג. (15 נק') הניחו שה-transition classifier שבו משתמש הפרסר הוא מודל לוג-ליניארי (log-linear model). נסמן את פונקציית התכוניות (feature function) של ה-classifier ב- $\phi: C \to \{0,1\}^{|V|^2}$, כאשר |V| הוא מספר המילים השונות בשפה, ו-C היא קבוצת הקונפיגורציות האפשריות של הפרסר. קבוצת כל אחד מהממדים של הפלט של ϕ מתאים לזוג מילים (לדוגמא (the,dog)). בהינתן קונפיגורציה $c \in C$ שבה המילים w ו-w' נמצאות בראש המחסנית, ϕ מחזירה וקטור שבו הקואורדינטה שמתאימה ל-(w,w') שווה ל-1 ושאר הקואורדינטות שוות ל-0. במילים אחרות, ϕ מקודדת רק את זהות שתי המילים הנמצאות בראש המחסנית.

הנח שהפרסר פולט את העץ הנכון עבור המשפט "teachers sleep in regularly" והוכח שהוא לא פולט את העץ הנכון עבור "teachers sleep in class".

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

שאלה 4 (20 נקודות):

א. (10 נק') בהקשר של sentiment classification of polarity עבור מסמכים, הגדירו באופן פורמלי את המודל bag-of-words log-linear model עבור משימה זאת. **אין** צורך להסביר כיצד לאמן את המודל.

ב. (10 נק') באיזה קושי מודל זה נתקל בהתמודדות עם שלילה (negation)?

שאלה 5 (20 נקודות):

א. (10 נק') הגדירו באופן פורמלי את המודל ההסתברותי Probabilistic Context Free Grammar (PCFG). **אין** צורך להגדיר פורמלית מהו עץ גזירה בדקדוק חסר הקשר (Context Free Grammar).

ב. (10 נק') איזה קושי נוצר במודל PCFG עם השימוש ב-head lexicalization?