

תאריך: 25.2.2020

**מבחן מועד א' בקורס עיבוד שפה טבעית (67658)**

שנת לימודים התש"ף 2019/2020

משך המבחן: שעותיים

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 בדיוק שתי שאלות מבין שאלות 1-3**

**שאלה 1 (40 נקודות):**

- א. (8 נק') הגדירו בצורה פורמלית את Bigram Maximum Entropy Markov Model (Bigram MEMM) עבור תיוג חלקי דיבר (POS tagging). יש להגדיר בתשובתכם את כל הסימונים בהם אתם משתמשים ואת המודל ההסתברותי.
- ב. (2 נק') מהם הפרמטרים של המודל שהגדרתם בסעיף א'?
- ג. (10 נק') רשמו את בעיית האופטימיזציה אותה יש לפתור ע"מ לשערך את הפרמטרים של המודל מסעיף א', וכיצד ניתן לפתור את הבעיה. רשמו את כל הנוסחאות המעורבות בהסבר שלכם.
- ד. (20 נק') נניח כעת שנתון מודל Bigram MEMM מאומן (כלומר, עם ערך מסוים לכל אחד מהפרמטרים). כתבו פסאודוקוד למתודה המקבלת כקלט את המודל המאומן, אינדקס  $j$  (מספר טבעי), סדרה של מילים  $x_1, \dots, x_n$  (tokens) כך ש- $j > n$ , וקבוצה סופית של כל חלקי הדיבר  $L$ , כך שכל מילה  $x_i$  מקבלת תגית  $y_i \in L$ . המתודה מחזירה את ההסתברות  $P(y_1, y_j | x_1, \dots, x_n)$  לכל ערך אפשרי של  $y_1$  ו- $y_j$ .
- הערה:** על זמן הריצה של המתודה להיות פולינומיאלי בגודל הקלט.

**שאלה 2 (40 נקודות):**

- א. (10 נק') הסבירו באופן איכותני מהו היתרון האפשרי משימוש ב-arc eager לעומת arc standard. נמקו את תשובתכם.
- ב. (10 נק') הגדירו באופן פורמלי מהו מצב או קונפיגורציה (state or configuration) בהקשר של transition-based parsing.
- ג. (10 נק') נניח כעת שידוע שבעצי תלויות מתקיים שאין שתי צלעות שחולקות מקור  $(t_1, t_2)$  ו- $(t_1, t_3)$  שיש להן את אותה התווית (label). נקרא לתכונה זאת תכונה (\*). כיצד ניתן לבצע מודיפיקציה ל-transition based parser כך שכל עץ אותו הוא פולט מקיים את התכונה (\*)?
- ד. (5 נק') הסבירו מדוע יהיה קשה יותר לבנות מודיפיקציה ל-MST parser כך שהפלט שלו בהכרח יקיים את התכונה (\*).
- ה. (5 נק') הסבירו כיצד בכל זאת ניתן לבצע מודיפיקציה ל-graph-based parser כך שהפלט שלו בהכרח יקיים את התכונה (\*), באמצעות שימוש בתכונות מסדר גבוה (higher-order features).

### שאלה 3 (40 נקודות):

- א. (10 נק') הגדירו מהו מודל הסתברותי דיסקרימינטיבי (discriminative probabilistic model) ומהו אומד נראות מרבית (maximum likelihood estimator) במקרה זה.
- ב. (10 נק') נתון מודל דיסקרימינטיבי עם וקטור פרמטרים  $w \in \mathbb{R}^d$  עבור קלאסיפיקציה בינארית, שבהינתן קלט  $x$  מחזיר התפלגות על  $\{0,1\}$ . נסמן ב- $f_w(x)$  את ההסתברות  $P(y=1|x; w)$ , ולכן  $P(y=0|x; w)=1-f_w(x)$ . שיטה נפוצה לאימון מודלים דיסקרימינטיביים היא ע"י מזעור ה-cross-entropy loss שמוגדר עבור אוסף דוגמאות  $\{(x_i, y_i)\}_{i=1 \dots N}$  כ:

$$L(w) = \frac{1}{N} \sum_{i=1 \dots N} -[y_i \log(f_w(x_i)) + (1 - y_i) \log(1 - f_w(x_i))]$$

- הראו שוקטור הפרמטרים הממזער את ה-cross-entropy loss הוא אומד הנראות המרבית במקרה זה.
- ג. (10 נק') נניח כעת שפתרון בעיית האופטימיזציה בסעיף ב' הובילה לוקטור  $w$  המקבל loss נמוך על דוגמאות האימון (training data), אבל הביצועים שלו על דוגמאות אחרות חלשים. הציעו מודיפיקציה לבעיית האופטימיזציה מסעיף ב' שעשויה לפתור את הבעיה. נמקו את תשובתכם.
- ד. (10 נק') הסבירו מהו מודל שפה מבוסס תווים (character-level language model). האם במקרה של מודל שפה מסוג זה תעדיפו להשתמש במודל שפה מבוסס RNN (recurrent neural network) או במודל שפה מרקובי (Markov language model)? נמקו את תשובתכם.

### חלק ב' (20 נקודות): ענו על בדיוק שאלה אחת מבין שאלות 4-5

#### שאלה 4 (20 נקודות):

- א. (10 נק') הסבירו במדויק מהי לקסיקליזציה במודל PCFG.
- ב. (10 נק') מהו היתרון והחיסרון בשימוש בלקסיקליזציה (lexicalization) במודל מסוג PCFG. נמקו את תשובתכם.

#### שאלה 5 (20 נקודות):

- א. (10 נק') הסבירו מהי הנחת הלוקאליות בשפה טבעית (locality) בהקשר של תרגום אוטומטי והסבירו מדוע מודל ההתאמה (alignment model) המבוסס על IBM model 1 איננו מתחשב בתכונת הלוקאליות.
- ב. (10 נק') הגדירו את מדד ההערכה (evaluation) לתרגום אוטומטי BLEU.

**בהצלחה!**