תאריך: 17.1.2022

מבחן מועד א' בקורס עיבוד שפה טבעית (67658)

2021/2022 שנת לימודים התשפ"ב

משך המבחן: מ-17.1.2022 ב-9:00 בבוקר עד 21.1.2022 ב-9:00 בבוקר

מרצה הקורס: פרופ' עמרי אבנד

<u>נהלי הבחינה:</u>

- כל חומר עזר מותר בשימוש, אבל קיים איסור מוחלט על דיון עם אנשים נוספים בנוגע למבחן או לקורס, למעט עם סגל הקורס. זה כולל איסור על התייעצות עם סטודנטים בקורס אבל גם עם כל אדם אחר, וזה כולל גם דיונים טכניים או שאלות הנוגעות להבנת הכתוב.
- כל השאלות לגבי המבחן, באם ישנן, צריכות להיות מופנות לסגל הקורס בלבד.
- קיים איסור מוחלט לפרסם בזמן המבחן כל מידע הנוגע לקורס או למבחן, כולל בפורומים, מדיה חברתית או כל אמצעי אחר.
- הודעות מטעמנו לגבי המבחן יפורסמו בפורום ההודעות של הקורס.
- צוות הקורס יהיה זמינים לשאלות באימייל של הקורס (huji.nlp2021@gmail.com בכתובת).
- בדומה למבחן רגיל, נענה אך ורק על שאלות הבהרה.
- עליכם להגיש את המבחן דרך ה-Moodle כקובץ pdf. ניתן לכתוב את התשובות בכתב יד ברור ולסרוק או לרשום אותן בתוכנה לעריכת טקסט.
- אנא הגישו את המבחן מבעוד מועד. לא יינתנו הארכות.
- חלק מהשאלות הן שאלות פתוחות שעשויות להיות עליהן מספר תשובות נכונות. כל תשובה מנומקת והגיונית תחשב כנכונה.
- **יש לענות על כל השאלות במבחן.**
- **עליכם להגיש הצהרה חתומה (מצורפת בסוף קובץ זה) יחד עם המבחן. אם אין ברשותכם מדפסת, אתם מוזמנים להעתיק את נוסח ההצהרה בכתב יד, לחתום ולצרף. מבחן ללא הצהרה חתומה לא ייבדק. אנא מלאו את ההצהרה <u>לפני</u> שאתם מתחילים לפתור על המבחן.**

שאלה 1 (15 נקודות):

נניח שנרצה לייצר training data עבור לימוד היחס place_of_birth ועבור היחס workplace באמצעות distant supervision שניתן ע״י Wikipedia infoboxes.

היחס place_of_birth מוגדר להיות יחס בין שתי יישויות, המתקיים אם היישות הראשונה היא אדם והשניה היא היישוב בו הוא/היא נולד/ה. למשל היחס מתקיים בין (David Ben Gurion, Plonsk). היחס workplace מוגדר להיות יחס בין שתי יישויות המתקיים אם היישות הראשונה היא אדם והשניה היא מקום העבודה בו הוא/היא עובד/ת או עבד/ה בעבר. למשל היחס מתקיים בין (Tim Cook, Apple).

1. (3 נק׳) הסבירו את ההבדל בין הגישה המפוקחת (supervised) לגישת ה-distant supervision. נזכיר כי בגישה המפוקחת ה-training set כולל טקסטים ובהם מסומנים זוגות של יישויות המקיימות את היחסים (לדוגמא, משפט בו כתוב "**Tim Cook** is the CEO of **Apple**״ ומסומן כי במשפט היחס בין Tim Cook ל-Apple הוא מופע של היחס workplace).
2. (5 נק׳) נניח גם שיש לכם תקציב לתייג דוגמאות ולבצע supervised learning עבור אחד היחסים ולהסתפק ב-distant supervision עבור השני. הציעו שני שיקולים שכדאי לקחת בחשבון על מנת להעריך איזה מבין היחסים מתאים יותר (ואיזה פחות) לגישת ה-distant supervision. הערה: אין צורך לרשום איזו החלטה הייתם מקבלים, אלא רק את השיקולים שינחו אתכם.
3. (7 נק׳) נציע כעת גישה אלטרנטיבית עבור חילוץ יחסים עבור שני היחסים הנ״ל. נאמן Neural Language Model על כמות גדולה של טקסט מהרשת. נסמן את ההסתברות שהמודל נותן עבור משפט כ-P. כעת נאמר ששתי יישויות X ו-Y (כלומר שאלו שמות העצם הפרטיים המתאימים להם) מקיימות את היחס birthplace אם

$$P(\text{"X was born in Y"}) > \eta$$

עבור ערך η, מספר ממשי נתון. אילו יתרונות ואילו חסרונות יש לגישה זאת על פני גישת ה-distant supervision? הציגו לפחות 2 יתרונות ו-2 חסרונות. נמקו את תשובתכם.

שאלה 2 (20 נק׳)

נרצה להגדיר מתייג (tagger) עבור זיהוי וסיווג ביטויים מרובי מילים (multi-word expressions) בשפה האנגלית. ביטוי מרובה מילים יכול להיות מורכב ממספר כלשהו (גדול מ-1) של מילים רצופות, או משני מקטעים רציפים של מילים. פורמלית, בהינתן משפט $x_1,x_2,...,x_n$ ביטוי מרובה מילים הוא תת-סדרה של מילים במשפט מאחת משתי הצורות הבאות:

1. $x_i,...,x_j$ for $i<j$ (for example, "**Tel Aviv** is the financial capital of Israel")
2. $x_i,...,x_j$ and $x_k,...,x_r$ for $i\leq j$ and $j+1<k\leq r$ (for example, "John **made** his sister **laugh**")

ניתן גם להניח שביטויים מרובי מילים מסוג (2) אינם מקוננים זה בתוך זה. כלומר, אם נתון ביטוי מרובה מילים מסוג (2) $x_i,...,x_j$ ו- $x_k,...,x_r$ עבור $i\leq j<j+1<k\leq r$ אז לכל אינדקס i בין $j+1$ ל-$k-1$ מתקיים ש- x_i לא שייך לאף ביטוי מרובה מילים מסוג (2) (יכול להיות שהוא יהיה שייך לביטוי מרובה מילים מסוג (1)). נניח ש-L היא קבוצה סופית לא ריקה של תגיות אפשריות עבור ביטויים מרובי מילים. נפתור בעיה זו באמצעות Conditional Random Field מסדר ראשון (bigram CRF).

1. (5 נק׳) נרצה בשלב ראשון לפתח מתייג המסוגל לזהות ביטויים מרובי מילים מסוג (1) (לא בהכרח מזהה ביטויים מסוג (2)). הציעו אלגוריתם המבוסס על CRF עם אוסף התגיות:

$$\{None\} \cup \{Begin_\ell : \ell \in L\} \cup \{Continue_\ell : \ell \in L\}$$

כלומר, כל מילה יכולה לקבל אחת מ-$2|L| + 1$ התגיות האלו.

2. (15 נק׳) הרחיבו את אוסף התגיות בו השתמשתם בסעיף הקודם, והציעו אלגוריתם המבוסס על CRF המסוגל לזהות ביטויים מרובי מילים מסוג (1) ו-(2).

עבור כל אחד מהסעיפים (סעיף 2.1 וסעיף 2.2) כללו כחלק מתשובתכם פסאודוקוד המקבל משפט $x_1,x_2,...,x_n$ ובו אוסף מסומן של ביטויים מרובי מילים, וממיר אותו לפלט של מודל ה-CRF שהצעתם, ופסאודוקוד המקבל משפט $x_1,x_2,...,x_n$ עם פלט אפשרי של מודל ה-CRF שהצעתם $y_1,y_2,...,y_n$ וממיר אותו לאוסף הביטויים מרובי המילים במשפט. שימו לב שמודל ה-CRF עשוי להחזיר $y_1,y_2,...,y_n$ שלא מתאימים לתיוג חוקי של הקלט. במקרה זה, על הפסאודוקוד לפלוט שקרתה שגיאה.

הערה: בשני הסעיפים של שאלה זו, אין צורך להגדיר מהן התכוניות שהמודל משתמש בהן (ה-feature function).

<u>שאלה 3 (25 נק׳):</u>

בשאלה זאת נעסוק בבעיית ה-transition-based parsing **unlabeled**.

1. (7 נק׳) נתבונן ב-Arc-eager transition system, ונסיר ממנה את פעולת ה-REDUCE, כך שה-transitions שישארו הן SHIFT, LEFT-ARC, RIGHT-ARC. נקרא ל-transition-system המתקבלת Modified-arc-eager. האם אוסף העצים אותם יכולים ליצור סדרות של transitions ב-Arc-eager system זהה לאוסף העצים אותם יכולים ליצור סדרות של transitions ב-Modified-arc-eager? נמקו את תשובתכם.

2. (10 נק') נתבונן ב- transition-system הבאה (אותם הסימונים כפי שלמדנו בשיעור):

SHIFT (same)	move one item from the buffer to the stack: $(\Sigma,\ i\|B,\ A) \Rightarrow (\Sigma\|i,\ B,\ A)$
LEFT-ARC'$_k$	create arc $j \rightarrow i_k$ and remove the top k elements from the stack: $(\Sigma\|i_k\|i_{k-1}\|\ldots\|i_1,\ j\|B,\ A) \Rightarrow (\Sigma,\ j\|B,\ A \cup \{(j, i_k)\})$ Condition: $i_1, \ldots i_{k-1}$ have a head, and i_k does not have a head.
RIGHT-ARC'$_k$	create arc $i_k \rightarrow j$, remove $i_1, \ldots i_{k-1}$ and shift: $(\Sigma\|i_k\|i_{k-1}\|\ldots\|i_1\ j\|B,\ A) \Rightarrow (\Sigma\|i_k\|j,\ B,\ A \cup \{(i_k, j)\})$ Condition: $i_1, \ldots i_{k-1}$ have a head.

Initial configuration: same as arc-standard.
Terminal configuration (Σ does not have to be [0]):

$$c_t = (\Sigma, [], A)$$

נקרא למערכת זאת: Arc-normal. קיימת פעולת LEFT-ARC'$_k$ ו-RIGHT-ARC'$_k$ עבור כל k טבעי. האם קיימים עצים שניתן לייצר באמצעות ה-Arc-eager system ולא ניתן לייצר באמצעות ה-Arc-normal system הנתונה? האם קיימים עצים שניתן לייצר באמצעות ה-Arc-normal system שלא ניתן לייצר באמצעות Arc-eager? הוכיחו את תשובתכם.

3. (8 נק') כתבו פסאודוקוד ל-Oracle עבור ה-Arc-normal system (אין צורך להוכיח את נכונותו).

שאלה 4 (20 נק׳):

בשאלה זאת העוסקת ב-sequence classification נסמן את הקלט (רצף המילים) $x_1,x_2\ldots x_n$ ואת רצף התגיות $y_1,y_2,\ldots,y_n$.

1. (5 נק׳) נתבונן במודל MEMM מסדר ראשון (bigram). הוכיחו או הפריכו: y_1 ו-y_3 בלתי תלויים בהינתן y_2 ו-$x_1,x_2\ldots x_n$.

כעת נרצה לבנות מתייג (tagger) המסוגל לבצע תיוג של שתי משימות של sequence classification בו זמנית (למשל POS Tagging ו-Named Entity Recognition).

נמשיך לסמן את הקלט (רצף המילים) ב-$x_1,x_2\ldots x_n$ ונסמן את את שני רצפי התגיות ב-$y_1,y_2\ldots y_n$ ו-$z_1,z_2\ldots z_n$.

אוספי התגיות האפשריות הם L_1 עבור ה-y_i ו-L_2 עבור ה-z_i. נניח גם שלשניהם אותו הגודל (מספר התגיות האפשריות עבור שני הרצפים הוא זהה). נסמן ב-M את הגודל המשותף של L_1 ו-L_2.

נגדיר מודל הסתברותי דיסקרימינטיבי המגדיר את ההתפלגות המשותפת של רצפי התגיות בהינתן הקלט באופן הבא:

$$P(y_1,\ldots,y_n,z_1,\ldots,z_n|x_1,\ldots,x_n)=\prod_{i=1}^{n}P(y_i,z_i|y_{i-1},z_{i-1},x_1,\ldots,x_n)=$$
$$\prod_{i=1}^{n}\frac{e^{\phi(y_i,z_i,y_{i-1},z_{i-1},x_1,\ldots,x_n,i)\cdot w}}{\sum_{y'\in L_1,z'\in L_2}e^{\phi(y',z',y_{i-1},z_{i-1},x_1,\ldots,x_n,i)\cdot w}}$$

כאשר w הוא וקטור משקולות ממשי ממימד d, וכאשר ϕ היא פונקצית פיצ׳רים המחזירה וקטורי תכוניות ממשיים ממימד d.

2. (7 נק׳) הציעו אלגוריתם הפותר את בעיית האופטימיזציה הבאה (בהינתן $x_1,x_2\ldots x_n$):

$$\max_{y_1,\ldots,y_n,z_1,\ldots,z_n}P(y_1,\ldots,y_n,z_1,\ldots,z_n|x_1,\ldots,x_n)$$

מהו זמן הריצה של האלגוריתם שהצעתם כפונקציה של M ושל n (אורך המשפט)?

3. (8 נק׳) נניח כעת שפונקצית הפיצ׳רים ϕ מתפרקת לסכום של שתי פונקציות תכוניות פשוטות יותר באופן הבא:

$$\phi(y_i,z_i,y_{i-1},z_{i-1},x_1,\ldots,x_n,i)=\phi_1(y_i,y_{i-1},x_1,\ldots,x_n,i)+\phi_2(z_i,z_{i-1},x_1,\ldots,x_n,i)$$

האם תוכלו להציע אלגוריתם יעיל יותר (כפונקציה של M) עבור הבעיה בסעיף הקודם?

שאלה 5 (20 נק׳):

1. (6 נק׳) נתבונן במודל שפה מרקובי מסדר ראשון (Bigram language model). נאמן את המודל באמצעות אומד נראות מירבית (Maximum Likelihood Estimation) על אוסף של משפטים, אבל לפני שנעשה זאת נסיר מהקורפוסים האלו את הסימון STOP המסמן סוף משפט, כך שהמשפטים יסתיימו במילה כלשהי ולאו דווקא ב-STOP. הראו שבמקרה כזה סכום ההסתברויות שמודל השפה המשוערך נותן עבור כל המחרוזות באורך סופי הוא גדול מ-1 (כלומר, המודל המשוערך איננו מודל שפה חוקי).
2. (6 נק׳) תנו דוגמא לשני משפטים בעברית או באנגלית, אחד נכון דקדוקית והשני אינו נכון דקדוקית, כך שמודל שפה מרקובי מסדר שני (trigram) ייתן הסתברות גבוהה למשפט הלא-נכון דקדוקית, או הסתברות נמוכה למשפט הנכון דקדוקית.
3. (8 נק׳) נניח כעת שנתון לנו קורפוס אימון עם עצים תחביריים (syntactically parsed training corpus). העצים הם עצי תלויות (dependency trees). הציעו דרך להגדיר מודל שפה חדש המתייחס לבעיה שהצגתם בסעיף הקודם. על המודל החדש להצליח הן לשמור על perplexity סביר (דומה לזה של מודל trigram) והן לתת לזוג הדוגמאות שנתתם בסעיף הקודם הסתברויות הגיוניות, כלומר הסתברות גבוהה למשפט הדקדוקי והסתברות נמוכה למשפט הלא-דקדוקי. הסבירו מדוע סביר שהמודל שהצעתם יקיים את התכונות הללו.

בהצלחה!

(ראו הצהרה בעמוד הבא)

הצהרה

אני הח"מ _____ (ת.ז. _____) מצהיר/ה כי לאורך זמן המבחן בקורס "עיבוד שפה טבעית" תשפ"א לא יצרתי קשר בנושא הקורס או המבחן עם אף אדם למעט סגל הקורס, ולא פרסמתי כל מידע בנוגע למבחן או לקורס בכל דרך שהיא.

תאריך: _____

חתימה: _____