תאריך: 16.2.2021

מבחן מועד א' בקורס עיבוד שפה טבעית (67658)

2020/2021 שנת לימודים התשפ"א

משך המבחן: מ-16.2.2021 ב-9:00 בבוקר עד 19.2.2021 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.nlp.course@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 ולסרוק או לרשום אותם בתוכנה לעריכת טקסט.
- אנא הגישו את המבחן מבעוד מועד. לא יינתנו הארכות.
- **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 (30 נקודות):

בשאלה זו נעסוק בבעיית ה-Named Entity Recognition (NER). בבעיה זו יש לחזות רצפי מילים שמשתייכים לקטגוריות Person או Location (שמות אחרים לא יסומנו). לדוגמא עבור המשפט "John lives in New York City", הפרדיקציה הנכונה היא:

(1) $[\text{John}]_{\text{Person}}$ lives in $[\text{New York City}]_{\text{Location}}$

נפתור בעיה זו באמצעות Condition Random Field (CRF) מסדר ראשון (bigram), כך שאוסף התגיות שלו הוא:
L={None, Begin-Person, Continue-Person, Begin-Location, Continue-Location}

למשל, הפרדיקציה הנכונה עבור "John lives in New York City" היא:

(2) $\text{John}_{\text{Begin-Person}}$ $\text{lives}_{\text{None}}$ in_{None} $\text{New}_{\text{Begin-Location}}$ $\text{York}_{\text{Continue-Location}}$ $\text{City}_{\text{Continue-Location}}$

א. (5 נק') כתבו פסאודו-קוד המקבל משפט מתויג ע"פ נוטציה (1) ומחזיר משפט מתויג ע"פ נוטציה (2).
ב. (5 נק') הראו שישנם רצפים של תגיות ע"פ נוטציה (2) שלא יכולים להתקבל כפלט לפרוצדורה שתיארתם בסעיף א'. תארו באופן מלא את רצפי התגיות ע"פ נוטציה (2) שלא יכולים להתקבל כפלט בסעיף א'.

בהינתן משפט $x_1,x_2,\ldots,x_n$, נגדיר את רצפי התגיות שלא יכולים להתקבל כפלט בסעיף א' כ-"רצפי תגיות לא חוקיים" ונסמנם ב-$A(x_1,\ldots,x_n)$, ואת רצפי התגיות שיכולים להתקבל כפלט בסעיף א' כ-"רצפי תגיות חוקיים" ונסמנם ב-$B(x_1,\ldots,x_n)$. האיחוד של שתי הקבוצות הוא L^n.

ג. (5 נק') הראו שכל מודל CRF עבור בעיה זו בהכרח ייתן הסתברות חיובית גם לרצפי תגיות לא חוקיים.
ד. (15 נק') הציעו אלגוריתם יעיל המקבל מודל CRF מאומן לבעיה זו, כלומר כולל פונקציית תכונות (feature vector) ווקטור משקולות, ומשפט נתון $x_1,\ldots,x_n$ ומחזיר את סכום ההסתברויות ע"פ המודל הנתון של רצפי תגיות חוקיים, כלומר את:

$$\sum_{y_1,..y_n \in B(x_1,..x_n)} P(y_1,..y_n|x_1,..x_n)$$

שאלה 2 (15 נקודות):

בשאלה זו נעסוק בבעיית ה-Unlabeled Dependency Parsing באמצעות graph-based methods. נסמן את משפט הקלט כ-$x_1,\ldots,x_n$ ואת הפרדיקציה עבור המשפט (העץ) ב-T.

נגדיר את *האורך הכולל בתווים* של parse tree באופן הבא: עבור כל צלע, האורך שלה בתווים יוגדר להיות מספר התווים הכולל (ללא white space) שמופיעים בין שתי המילים שהצלע מחברת. פורמלית, אם הצלע היא בין x_i ל-x_j ו-$j>i$, אז אורך הצלע בתווים יקבע להיות סכום מספר התווים במילים $x_{i+1},x_{i+2},\ldots,x_{j-1}$. אם $j<i$, אז אורך הצלע בתווים

יקבע להיות סכום מספר התווים במילים $x_{j+1},..,x_{i-1}$. האורך הכולל בתווים של עץ T יוגדר כסכום אורכי הצלעות בתווים עבור הצלעות שב-T.

א. (7 נק') האם ניתן לשלב את האורך הכולל בתווים של עץ כ-feature (תכונית) ב-edge-factored model או שמא נדרש מודל מסדר גבוה יותר? נמקו את תשובתכם.

ב. (8 נק') תארו תכונית של T שלא ניתן לייצג באמצעות edge factored model, אבל כן ניתן לייצג אותה עם grandchild model. הסבירו את תשובתכם.

<u>שאלה 3 (25 נקודות):</u>

נתונים שלושה מודלי שפה (language models) המבוססים על recurrent neural networks. נסמן את פונקציה ההסתברות שהם מתארים ע"י P_{M1}, P_{M2}, P_{M3}. נרצה לבנות מודל שפה חדש המשלב את היתרונות של שלושת המודלים ע"י אינטרפולציה ביניהם.

נגדיר מודל שפה חדש M ע"י:

$$P_M(x_n|x_1,\dots,x_{n-1}) = \lambda_1 P_{M1}(x_n|x_1,\dots,x_{n-1}) + \lambda_2 P_{M2}(x_n|x_1,\dots,x_{n-1}) + \lambda_3 P_{M3}(x_n|x_1,\dots,x_{n-1})$$

כאשר $\lambda_1, \lambda_2, \lambda_3$ הם מספרים ממשיים.

א. (5 נק') מהו התנאי שצריך להתקיים על $\lambda_1, \lambda_2, \lambda_3$ על מנת ש-P_M אכן יגדיר התפלגות חוקית? הוכיחו את טענתכם.

ב. (10 נק') נניח שנרצה להפעיל את M על טקסטים הלקוחים מאתרי חדשות. נתון קורפוס שנלקח מאתרי חדשות C. הציעו שיטה לקבוע את הערכים של $\lambda_1, \lambda_2, \lambda_3$ על מנת להתאים את M לעבודה על אתרי חדשות. רשמו את כל הנוסחאות המשמשות אתכם.

הערה: אין צורך להסביר כיצד פותרים את בעיית האופטימיזציה, אלא רק להגדיר מהי בעית האופטימיזציה אותה יש לפתור.

ג. (10 נק') נניח שידוע שהמודל M1 עובד טוב יותר מ-M2 ו-M3 בחיזוי (פרדיקציה) המילה x_n כאשר x_{n-1} מתחילה באות a, והמודל M2 עובד טוב יותר מ-M1 ו-M3 כאשר x_{n-1} מתחילה באות b. לא ידוע על מקרים בהם M3 עובד טוב יותר מ-M1 ו-M2. כיצד ניתן לשנות את השיטה שהצעתם בסעיף ב' על מנת לקחת בחשבון מידע זה?

שאלה 4 (15 נקודות):

נתבונן במודל Maximum Entropy Markov Model (MEMM) מסדר שני, כלומר מודל המקיים:

$$P(y_1,..y_n|x_1,..x_n) = \prod_{i=1...n} P(y_i|x_1,..x_n,y_{i-1},y_{i-2})$$

כאשר y_{-1} ו-y_0 שווים תמיד ל-START.

א. (5 נק') רשמו נוסחא מפורשת ל- $P(y_i|x_1,..x_n,y_{i-1},y_{i-2})$. מהם הארגומנטים שמקבלת פונקציית התכוניות (feature function)?

ב. (5 נק') כתבו נוסחא עבור הגרדיינט של ה-conditional log-likelihood בהינתן training data של N דוגמאות: $\{(x_1^{(i)},\dots,x_{n(i)}^{(i)},y_1^{(i)}\dots y_{n(i)}^{(i)})\}_{i=1}^{N}$.

ג. (5 נק') האם ניתן לחשב את הגרדיינט בזמן פולינומי בגודל הקלט? נמקו את תשובתכם.

שאלה 5 (15 נקודות):

נעסוק בבעיה של Text Classification, כלומר בהינתן טקסט נרצה לסווג אותו לאחת מ-3 קטגוריות. נסמן את הקטגוריות ב-l_1, l_2, l_3.

א. (7 נק') הגדירו באופן מדויק log-linear classifier עבור הבעיה הזו שמשתמש בתכוניות bag of words בינאריות. רשמו כיצד מחושבת פונקציית התכוניות (feature function), כיצד מוגדר המודל ההסתברותי, וכיצד מבוצע היסק (inference) במודל.

ב. (8 נק') נניח שנתון log-linear classifier כפי שמוגדר בסעיף א' שאומן על קטעי טקסט באנגלית מהעיתונות. נרצה להפעיל אותו על תחום (domain) חדש שבו טקסטים מכילים הרבה טעויות איות (spelling errors) או איות לא סטנדרטי (כמו "u 2" במקום "you too"). נניח שנתון הרבה טקסט לא מתויג מהתחום החדש, אך לא נתון טקסט מתויג מתחום זה כלל.

הציעו כיצד ניתן לעשות שימוש ב-word embeddings על מנת לשנות את ה-classifier מסעיף א' כך שיתמודד טוב יותר עם טעויות האיות והאיות הלא סטנדרטי.

בהצלחה!

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: _____