עיבוד שפה טבעית תרגיל בית 5

יותם מנה, 204717862
גיא אזוב, 312567654

<u>שאלה 1:</u>

<u>סעיף a:</u>

1. משפטים עם מילים בעלות דוד משמעות שמית:

Ford makes the best cars.
המילה Ford במשפט מכוונת ליצרנית המכוניות, אף אותה מילה יכולה לשמש גם כשם של בן אדם.

I can get to JFK with a taxi.
המילה JFK מתייחסת לשדה התעופה בניו יורק, אך היא גם יכולה לשמש ככינוי שך נשיא ארה"ב לשעבר.

ii על ידי שימוש בפיצ'רים למעט המילה עצמה ניתן לתת לה הקשר בתוך המשפט, שיעזור לפתור את בעיית הדו-משמעות. בנוסף, מילים שהן named entities עלולות להיות נדירות באימון ולכן הפיצ'רים עוזרים בהכללה.

iii דוגמאות לפיצ'רים: (נניח כי המילה היא w_i, במיקום i במשפט)

- W_{i-1} is part of a named entity **and** w_i starts with a capital letter.
- W_{i-1} is a prepesition that refers to a location (e.g. at, inside, out of…)

<u>סעיף b:</u>

מימדים:

$\boldsymbol{e}^{(t)}: 1 \times (2w+1)D, \qquad W: (2w+1)D \times H, \qquad U: H \times C$

סיבוכיות חיזוי עבור משפט באורך T: (נניח גודל חלון w)

חישוב $\boldsymbol{e}^{(t)}$: מכפילים $(2w+1)$ פעמים את וקטור המילה (one hot) עם המטריצה E המכילה וקטורי קידוד באורך D – סה"כ $O((2w+1)D)$.

חישוב $\boldsymbol{h}^{(t)}$: $O((2w+1)DH+H)$.

חישוב $\boldsymbol{y}^{(t)}$: $O(HC)$

סה"כ עבור מילה בודדת במשפט: $O((2w+1)DH+HC)$
ולכן עבור משפט באורך T: $O(T((2w+1)DH+HC))$

<u>סעיף d:</u>
התוצאות הטובות ביותר עם דיוק של 83% F1:

2019-06-08 11:25:13,782:DEBUG: Token-level confusion matrix:

go\gu	PER	ORG	LOC	MISC	O
PER	2967	49	55	12	66
ORG	143	1638	120	53	138
LOC	60	94	1873	20	47
MISC	40	54	52	1009	113
O	58	54	15	26	42606

2019-06-08 11:25:13,782:DEBUG: Token-level scores:

label	acc	prec	rec	f1
PER	0.99	0.91	0.94	0.92
ORG	0.99	0.87	0.78	0.82
LOC	0.99	0.89	0.89	0.89
MISC	0.99	0.90	0.80	0.85
O	0.99	0.99	1.00	0.99
micro	0.99	0.98	0.98	0.98
macro	0.99	0.91	0.88	0.90
not-O	0.99	0.89	0.87	0.88

2019-06-08 11:25:13,782:INFO: Entity level P/R/F1: 0.82/0.85/0.83

ניתן להסיק מה-confusion matrix שהטעויות המשמעותיות הן:
True class: ORG, predicted class: PER,LOC,O
True class: MISC predicted class: O

באופן כללי רואים שהמודל הכי מתקשה עם זיהוי ORG לפי מספר השגיאות בשורה זאת בטבלה.

מגבלות המודל:

1. המודל לא משתמש בהמשכיות של תיוגים שביצע על מילים קודמות, כלומר הוא לא מתחשב בעובדה שרצף של מילים יכול להוות שם אחד של אותה יישות ולכן יש לתייג את כולן בצורה זהה. לדוגמה:

x : Jordan won the Samsung Tel Aviv marathon
y': LOC O O ORG ORG LOC O

האירוע "מרתון סמסונג תל אביב" מהווה ישות אחד, אך המודל תייג חלק מהמילים בצורה שונה.

2. במשפטים ארוכים המודל אינו מסוגל להתחשב בהקשר של מילה כדי לפתור דו משמעות. לדוגמה:

x : in his interview, Jordan said that he was very excited from his victory
y': O O O LOC O O O O O O O O O

ברור שהמשפט עוסק בראיון הניצחון של Jordan במרתון, אך המודל לא מסוגל להבחין בכך ולכן מתייג את השם ב LOC.

<u>שאלה 2:</u>

<u>סעיף a:</u>

i. מספר הפרמטרים של מודל ה RNN לעומת מודל החלון:

$W_x: D \times H$ (instead of $(2w+1)D \times H$ in window model)

$W_h: H \times H$ (instead of 0 in window model)

ii. סיבוכיות החישוב לחיזוי תיוגים של משפט באורך T:

a. חישוב $\boldsymbol{e}^{(t)}$: $O(D)$ - כאורך קידוד המילה

b. חישוב $\boldsymbol{h}^{(t)}$: $O(H^2 + HD + H)$ –

i. כפל וקטור בגודל H עם מטריצה בגודל HxH.

ii. כפל וקטור בגודל D עם מטריצה בגודל DxH.

iii. הוספת bias בגודל H.

c. חישוב $\hat{\boldsymbol{y}}^{(t)}$: $O(HC + C)$

i. כפל וקטור בגודל H עם מטריצה בגודל HxC.

ii. הוספת bias בגודל C.

d. סה"כ:

$$T(D + H^2 + HD + H + HC + C) = O\big(TH(H + D + C)\big)$$

<u>סעיף b:</u>

i. ניקח לדוגמה את השם הבא, כולל התיוגים שלו:

Tel/LOC Aviv/LOC Yaffo/LOC

עם התיוגים הבאים:

1. O O O
2. LOC LOC O

במקרה זה ערך ה cross entropy ירד מכיוון שמ-1 ל-2 חזינו נכון שני תיוגים יותר, ובנוסף בגלל שהפרדנו את Tel Aviv מ Yaffo, הגדלנו את מספר הישויות בו מ-1 ל-2. בכך ערך ה precision יורד, ערך ה recall לא משתנה ולכן סה"כ ערך ה F1 יורד.

ii. קשה לעשות אופטימיזציה ל F1 מכיוון שעל מנת לחשב אותו נדרשים התיוגים של המודל על כל הטקסט, ולכן אין קשה למקבל את התהליך.

<u>סעיף d:</u>

ללא masking ה loss וה gradient יתחשבו בתיוגים שהמודל ביצע על האפסים שהוספנו בסוף כל משפט וזה יבוא לידי ביטוי בשינוי הפרמטרים של המודל. השימוש ב masking מבטל את ההשפעות האלה ע"י איפוס הרכיבים הלא רלוונטיים ב loss וב gradient.

<u>סעיף g:</u>

מגבלה ראשונה של מודל ה RNN היא חוסר היכולת לתייג באותה צורה רצפים של מילים שכולן חלק מאותו שם, לדוגמה:

x : Ariel Sharon Park

y': PER LOC LOC

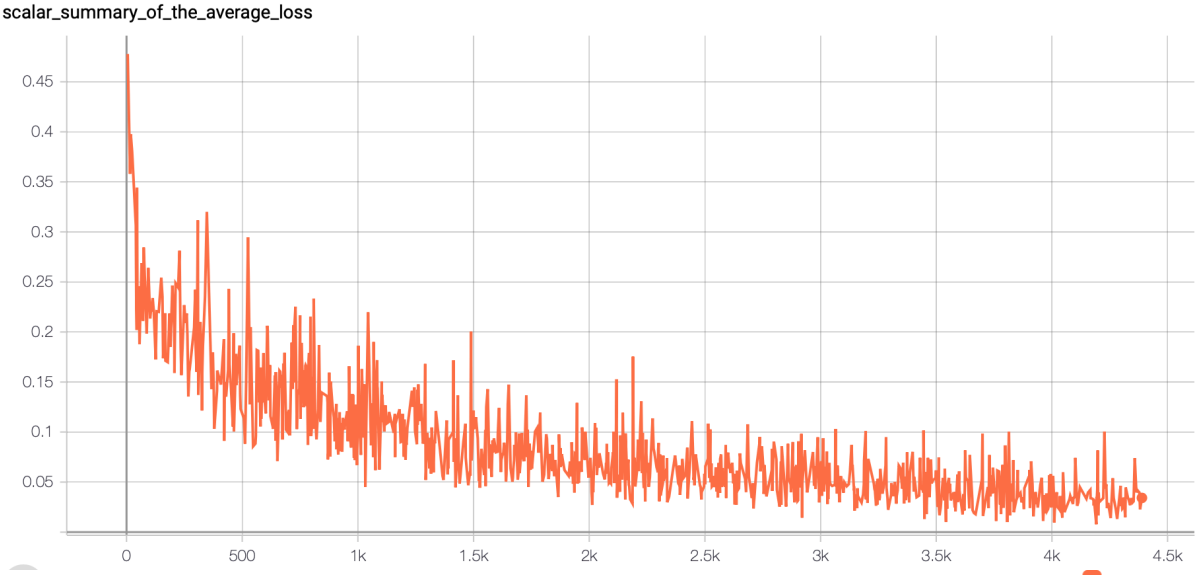
במשפט כתוב שם הפארק "פארק אריאל שרון" שכולו אמור להיות מתוייג LOC.

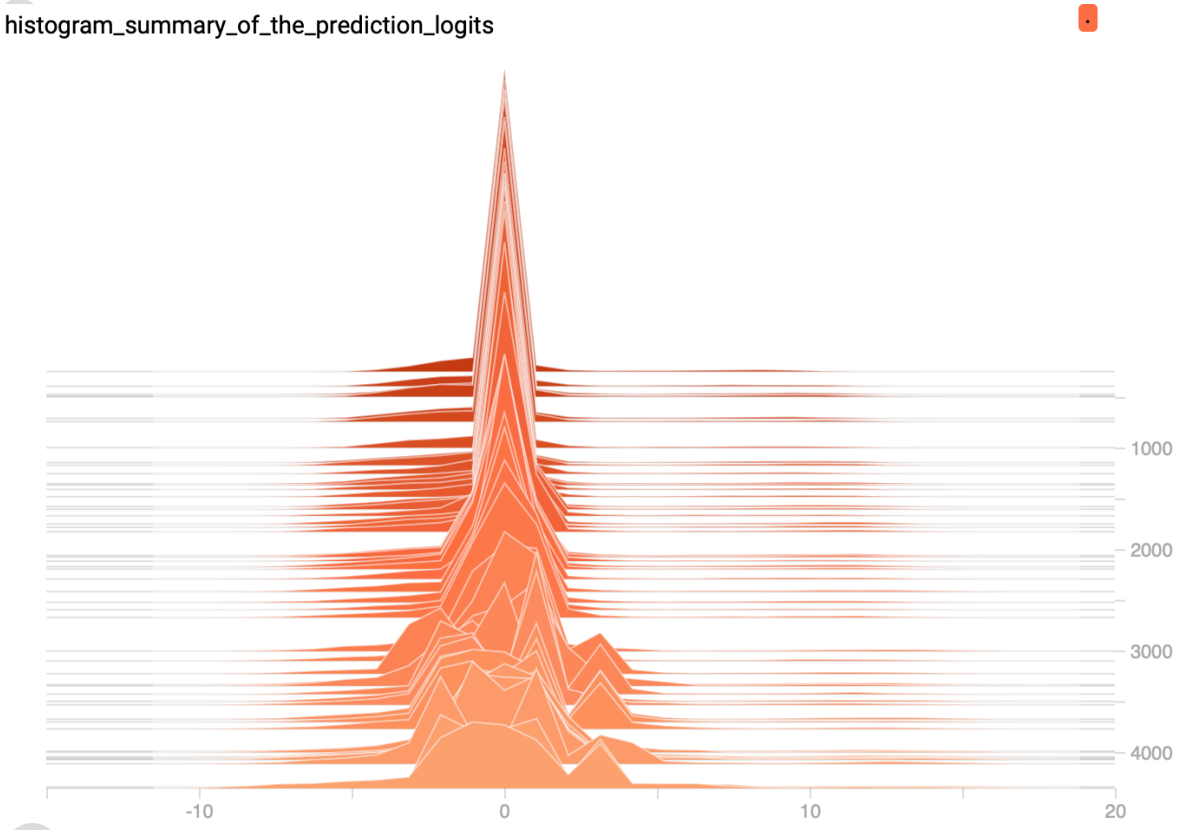
דרך לפתרון – שימוש ב Attention שייתן משקל גבוה לכל מילים הרצופות הקודמות שהן חלק משם של יישות.

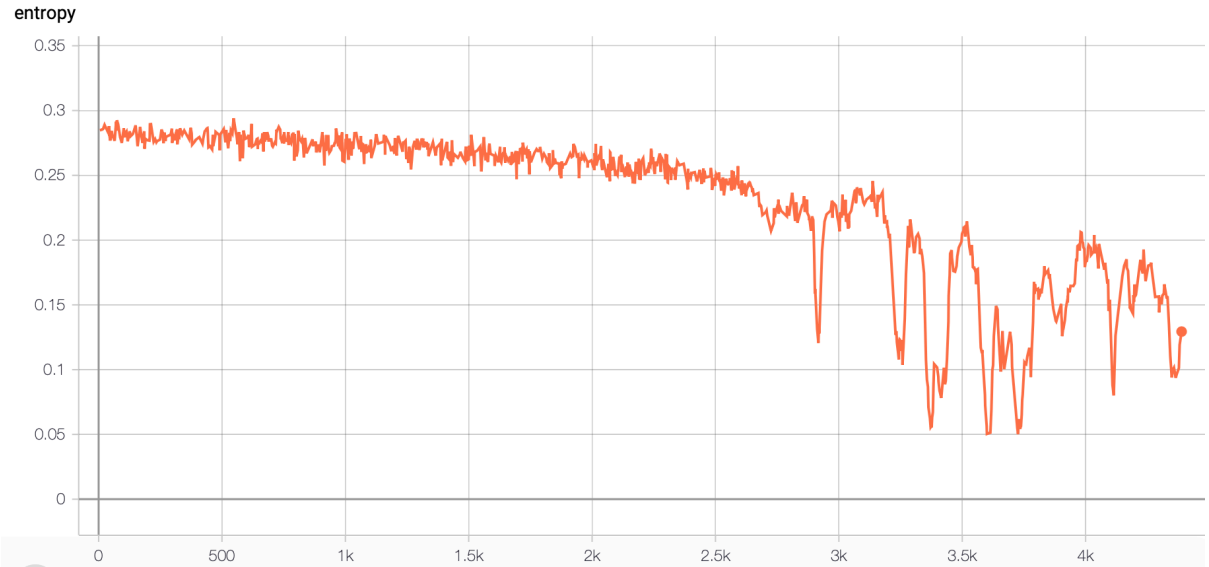
מגבלה שניה היא חוסר יכולת להסתכל על המילים הבאות במשפט. לדוגמה במשפט שהצגנו למעלה אם המודל היה יודע שהמילה Park עומדת להגיע בסוף השם, סביר יותר שהיה מתייג את המילה Ariel כ LOC.
דרך לפתרון – שימוש ב bi-directional RNN.

שאלה 3:

הגרפים:







<u>סעיף c:</u>

i. האנטרופיה המקסימלית האפשרית מתקבלת כאשר ההתפלגות המתקבלת אחידה – והערך הוא: $\log(5)$ (יש 5 תיוגים אפשריים)

ii. ניתן לראות מהגרף שהאנטרופיה הולכת ויורדת ככל שהמודל מאומן יותר. זה מרמז לנו שהמודל הולך וצובר יותר ביטחון בהחלטות שלו. תופעה זאת משתקפת בהיסטוגרמה המצורפת – ניתן לראות שבשלבים מוקדמים של האימון כל מסת ההסתברות מרוכזת סביב מרכז ציר ה X, כלומר המודל לא יודע להבדיל בין הישויות השונות ומקצה להן הסתברויות דומות. כאשר מסתכלים על שלבים מאוחרים יותר רואים שהמודל מצליח יותר להבדיל בין הישויות ונותן לכל אחת הסתברות שונה.

<u>סעיף d:</u>

ניתן לראות מטבלת ה Token-level scores שהמודל מתקשה יותר עם ישויות מסוג ORG בעוד שמצטיין בזיהוי של PER:

label	acc	prec	rec	f1
PER	0.99	0.94	0.94	0.94
ORG	0.99	0.91	0.78	0.84
LOC	0.99	0.86	0.94	0.90
MISC	0.99	0.87	0.81	0.84
O	0.99	0.99	1.00	0.99

לכן נצפה שירבה לטעות בזיהוי של ארגונים בעלי שמות של אנשים
דוגמאות מהמודל:

```
x : I    like Tommy Hilfiger 's   shirts
y': O    O    PER   PER      O    O
p : 1.00 1.00 0.97  0.98     1.00 1.00
```

ברור שכאן הכוונה לחברת הביגוד ולא לבן אדם ששמו Tommy Hilfiger.

```
x : I    made it   to   Guinness 's   book
y': O    O      O    O    ORG        O    O
p : 1.00 1.00 1.00 1.00 0.63     1.00 1.00
```

בדוגמה זאת המודל צודק, אך ניתן לראות שהוא מאוד לא בטוח בתשובה – המילה Guinness קיבלה הסתברות נמוכה של 0.63.