תרגיל 4 NLP

אור מלכאי – 203569264 – ormalkai@gmail.com

איציק מור – 204240246

סטפן גולדברג – 316802222

שאלה 1

סעיף a

i. נציג מספר דוגמאות לambiguity במשפט הבא(https://www.washingtonpost.com/business/economy/mary-schapiro-to-leave-sec-next-month/2012/11/26/d61d4c74-37df-11e2-8a97-363b0f9a0ab3_story.html?noredirect=on&utm_term=.39a79ff1710c):

S.E.C. chief Mary Shapiro to leave Washington in December.

התיוג המקורי הוא:

(ORG S.E.C.) chief (PER Mary Shapiro) to leave (LOC Washington) in December.

בתיוג הזה S.E.C זה למעשה acronym ל- Security and Exchange Commission וזה למעשה ארגון.

בנוסף Mary Shapiro תויג כבנאדם ו-Washington תויג כמקום.

במשפט הנ"ל ניתן לראות שתי דוגמאות לambiguity הראשונה, הקלה יותר, היא שWashington במקרה הזה זה שם של עיר, אך מאחר והעיר נקראה על שם של בנאדם, בקלות ניתן למצוא מופעים של Washington שיתויגו כבנאדם. בנוסף נשים לב שהמילה chief לא נכללה בתיוג "בנאדם" של Mary Shapiro, בקלות היה ניתן לתייג את S.E.C chief Mary Shapiro בתיוג אחד של בנאדם, או אולי אפילו תיוג ותיוגים פנימיים.

ii. הצורך בפיצ'רים נוספים מלבד המילה עצמה נועד כדי לפתור קונפליקטים שנובעים מambiguity. למשל, אם היינו פותרים קונפליקטים ע"י אומד MLE של תיוג למילה, על פי התיוג הנפוץ ביותר בtraining היינו בטוח טועים על התיוג שראינו פחות. כדי לפתור זאת עלינו להבין מה הcontext של המילה ולהסתכל על פיצ'רים נוספים. יתר על כן, הצורך בפיצ'רים נוספים נובע גם מהצורך לטפל ב"rare words" שלרוב הן named entities.

iii. ישנם מספר פיצ'רים שניתן לחשוב שיעזרו בתיוג named entities למשל, המילה הקודמת/הבאה, או באופן כללי חלון של מילים, או "תקציר" של המילים בחלון מסוים. פיצ'ר נוסף יכול להיות הPart Of Speech של המילה. יתר על כן ניתן לחשוב גם על פיצ'רים שקשורים בסינטקס של המילה, למשל prefix או suffix, לדוגמא מילים שמסתיימות בly רוב הסיכויים לא יהיו Person או Organization.

סעיף b

i. $e = (1,(2w+1) * D)$

$W = ((2w+1) * D, H)$

$U = (H, C)$

כאשר D הוא גודל הembedding, H הוא גודל השכבה החבויה ו-C הוא מספר ה-classים.

ii. לפי המימדים שתיארנו בסעיף הקודם, עבור חלון יחיד בגודל w, הסיבוכיות היא:

$O((2w+1)D)$ עבור חישוב $\mathbf{e}^{(t)}$

$O((2w+1)DH+H)$ עבור חישוב $\mathbf{h}^{(t)}$

$O(HC+C)$ עבור חישוב $\mathbf{y}^{(t)}$

לכן, סה"כ עבור משפט באורך T נקבל $O((2w+1)DHT+HC)$

סעיף d

i. התוצאה הטובה ביותר שקיבלנו עבור ה-development set היא:

$$F1 = 0.83, Precision = 0.82, Recall = 0.85$$

Confusion Matrix

Gold\Guess	PER	ORG	LOC	MISC	O
PER	2981	18	65	14	71
ORG	156	1592	133	75	136
LOC	47	68	1910	22	47
MISC	45	46	58	1007	112
O	47	41	16	29	42626

לפי הConfusion Matrix ניתן לראות כי באופן יחסי המודל מנבא באופן לא רע, כלומר רוב האנשים מזוהים כאנשים, רוב המקומות מזוהים כמקומות, רוב ה"שונות" מזוהים כ"שונות" ורוב הO מזוהים כO. לעומת זאת, הרבה ארגונים מזוהים כאנשים, ניתן לראות זאת גם על בסיס הRecall הנמוך, 0.76, ברמת הToken.

ii. הבעיה עיקרית של windowed model, היא שהוא לא נעזר בפרדיקציות שבחלון בשביל לנבא פרדיקציות נוספות, בדומה למשל לlog linear models שראינו בתרגיל הקודם עבור POS. בגלל המגבלה הזו ניתן לראות "רצפים" של entities שלא מזוהים כמו שצריך, למשל עבור המשפט:
Starting(go:O, gu:O) on(go:O, gu:O) May(go:O, gu:O) 13(go:O, gu:O) next(go:O, gu:O) year(go:O, gu:O), (go:O, gu:O) the(go:O, gu:O) Test(go:ORG, gu:MISC) and(go:ORG, gu:O) Country(go:ORG, gu: ORG) Cricket(go: ORG, gu: ORG) Board(go: ORG, gu: ORG)

במקרה הנ"ל קל לראות ש"Test and Country Cricket Board" זה בעצם ארגון, ואם המודל היה משתמש בתיוגים סביב כדי לנבא את Test ואת and הוא כנראה לא היה טועה. באופן דומה ניתן גם לומר שwindowed model לא נעזר במידע משאר המשפט, כלומר חלקי המשפט שלא נמצאים בחלון. עבור משפטים שמכילים "לוואי" ארוך ניתן לחלץ מידע רלוונטי דווקא מחלקי משפט "רחוקים".

שאלה 2

סעיף a

i. ישנם שני הבדלים בין מודל הRNN לבין מודל הwindow.
ראשית המימד של W_e הוא (D, H) במודל RNN לעומת ((2w+1) * D, H) עבור W במודל הwindow.
שנית, W_h לא קיימת במודל הwindow והמימד שלה הוא (H, H).

ii. עבור cell יחיד בRNN הסיבוכיות היא:
$O(D)$ עבור חישוב $\mathbf{e}^{(t)}$
$O(H*H+DH+H)=O\big(H(H+D+1)\big)=O(H(H+D))$ עבור חישוב $\mathbf{h}^{(t)}$
$O(HC+C)$ עבור חישוב $\mathbf{y}^{(t)}$

לכן, סה"כ עבור משפט שלם באורך T נקבל $O((D+H)HT)$ בהנחה ש $D,H \gg C$

סעיף b

i. ראשית עלינו להבין את החישוב של F1 ובפרט את הPrecision והRecall ברמת ה-entity.
<u>Precision</u> הוא החלק היחסי של ה-named entities שהמודל חזה נכונה שמיושרים עם -named entities ב-development set, מתוך כל הnamed entities שהמודל חזה. $Precision=\frac{tp}{tp+fp}$
<u>Recall</u> הוא החלק היחסי של ה-named entities בdevelopment set שהמודל חזה נכונה מתוך כל הnamed entities בdevelopment set. $Recall=\frac{tp}{tp+fn}$
<u>F1</u> הוא הממוצע ההרמוני שלהם. $F1=\frac{2*Precision*Recall}{Precision+Recall}$

נתבונן במשפט הבא:
תיוגים נכונים:
$"Harry(PER)\ Potter(PER)\ and(O)\ the(LOC)\ chamber(LOC)\ of(LOC)\ secretes(LOC)"$
תיוגים אפשריים:
1. $"Harry(PER)\ Potter(PER)\ and(O)\ the(LOC)\ chamber(LOC)\ of(LOC)\ secretes(O)"$
2. $"Harry(PER)\ Potter(PER)\ and(O)\ the(O)\ \ chamber(O)\ \ of(O)\ \ secretes(O)"$

נניח ובחרנו בתיוג הראשון על פני התיוג השני, ברור כי ה-cross entropy תקטן מכיוון ש ברמת הtokens אנחנו צודקים יותר. לעומת זאת ברמת ה-entity הF1 גם כן יקטן:

$$Precision1=\frac{1}{1+1}=0.5, Recall1=\frac{1}{1+1}=0.5, F1=\frac{2*0.5*0.5}{0.5+0.5}=0.5$$
$$Precision2=\frac{1}{1+0}=1, Recall2=\frac{1}{1+1}=0.5, F1=\frac{2*1*0.5}{0.5+1}=\frac{1}{3}$$

ii. ישנן לפחות שתי סיבות לכך שקשה לבצע אופטימיזציה ישירות על הF1. הסיבה המתמטית היא שהפונקציה לא דיפרנציאבילית. הסיבה הפרקטית היא שבשביל לחשב F1 דרוש חישוב על פני כל הdata וזה יקר.

סעיף d

i. במידה ולא היינו מבצעים masking ה loss היה כולל גם את השגיאה שלנו על הpadding שהוספנו.

לכן הגרדיאנטים של הpadding שהוספנו היו משפיעים על הפרמטרים של הרשת במהלך הbackpropagation. באמצעות השימוש בmasking למעשה הloss על הpadding הוא 0, לכן גם הגרדיאנטים של הpadding יתאפסו ולא נשפיע על הפרמטרים של הרשת.

סעיף g

נתאר שתי מגבלות של מודל RNN:

1. מגבלה ראשונה במימוש הRNN שלנו היא שהRNN עובדת משמאל לימין, כלומר מתחילת המשפט לסופו. ברור שאם היינו יודעים מידע על המשך המשפט היינו יכולים לתייג יותר טוב. פתרון פשוט, בדומה לbiLSTM ניתן לבצע מעבר גם מהסוף להתחלה. דוגמא מהdata:

   May(go:O, gu:O) 15(go:O, gu:O) v(go:O, gu:O) Duke(go:ORG, gu:ORG) of(go:ORG, gu:O) Norfolk(go:ORG, gu:ORG) ‘s(go:ORG, gu:O) XI(go:ORG, gu:ORG) ((go:O, gu:O) at(go:O, gu:O) Arundel(go:LOC, gu:LOC)) (go:O, gu:O)

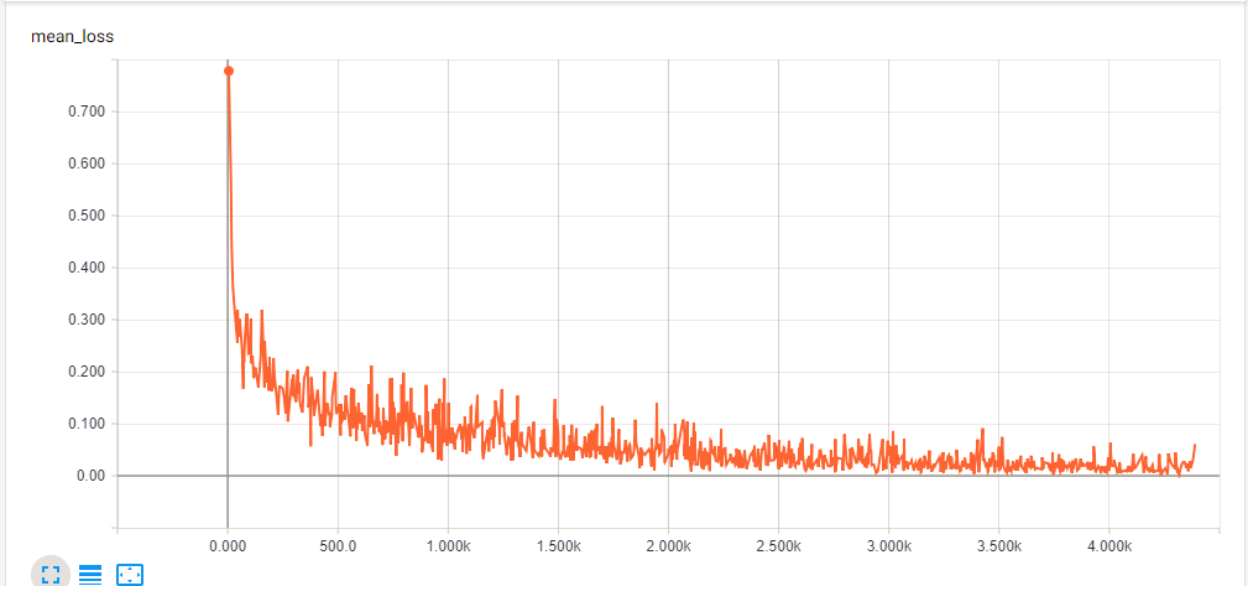
   בדוגמא הנ"ל טעינו על ‘s ועל of, כנראה שלא היינו טועים אם היה לנו מידע על המשך המשפט.

2. המגבלה השנייה במימוש שלנו היא שהיינו רוצים ל"תגמל" תיוגים רציפים של אותם entities. הדוגמא לעיל עדיין טובה, שכן היינו רוצים להאמין שבתוך היער של הORG נמשיך לתייג ORG ולא O. יכולנו לפתור זאת ע"י רגולריזציה בloss שנותנת עדיפות לתיוגים שכנים זהים. לחילופין, יכולנו להכניס קלט נוסף לכל cell שהוא התיוג הקודם.

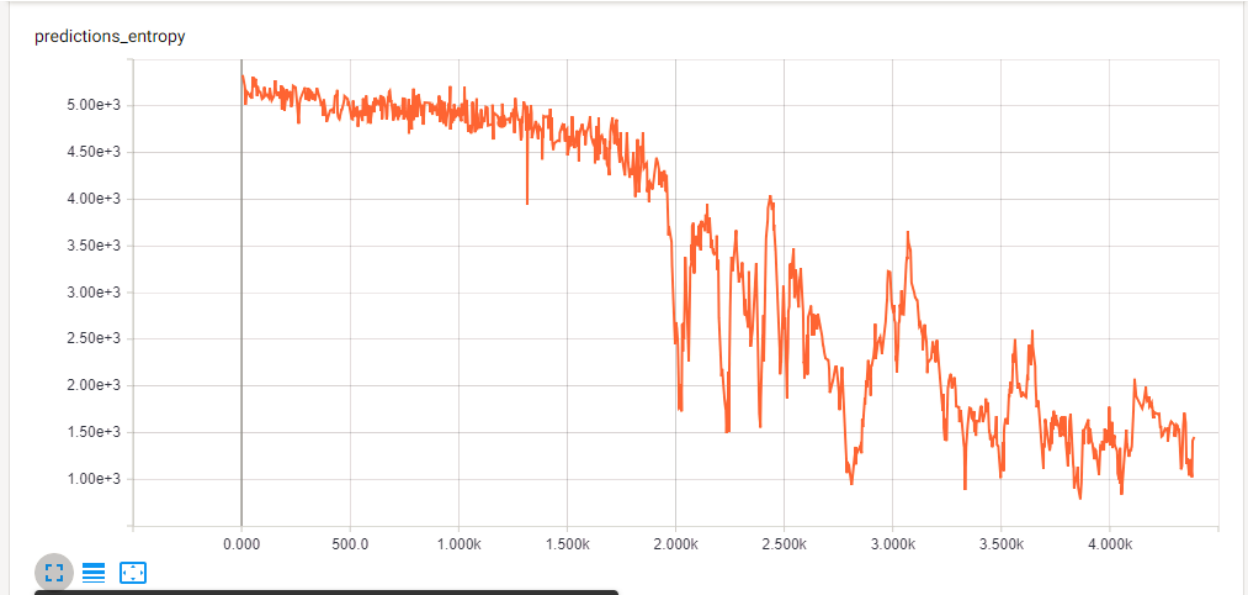
שאלה 3

סעיף c

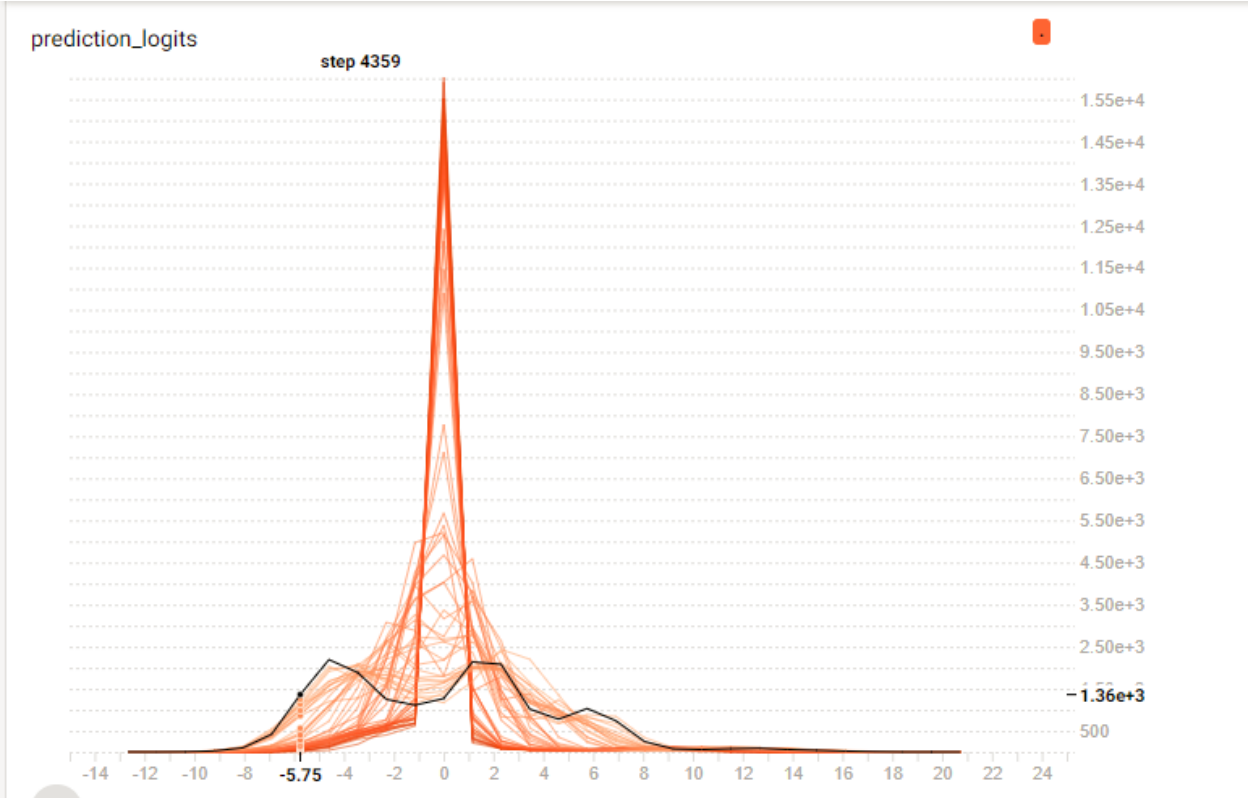
i. עבור timestep יחיד האנטרופיה המקסימלית היא אינסוף, במקרה שקיבלנו $\hat{y} = 0$

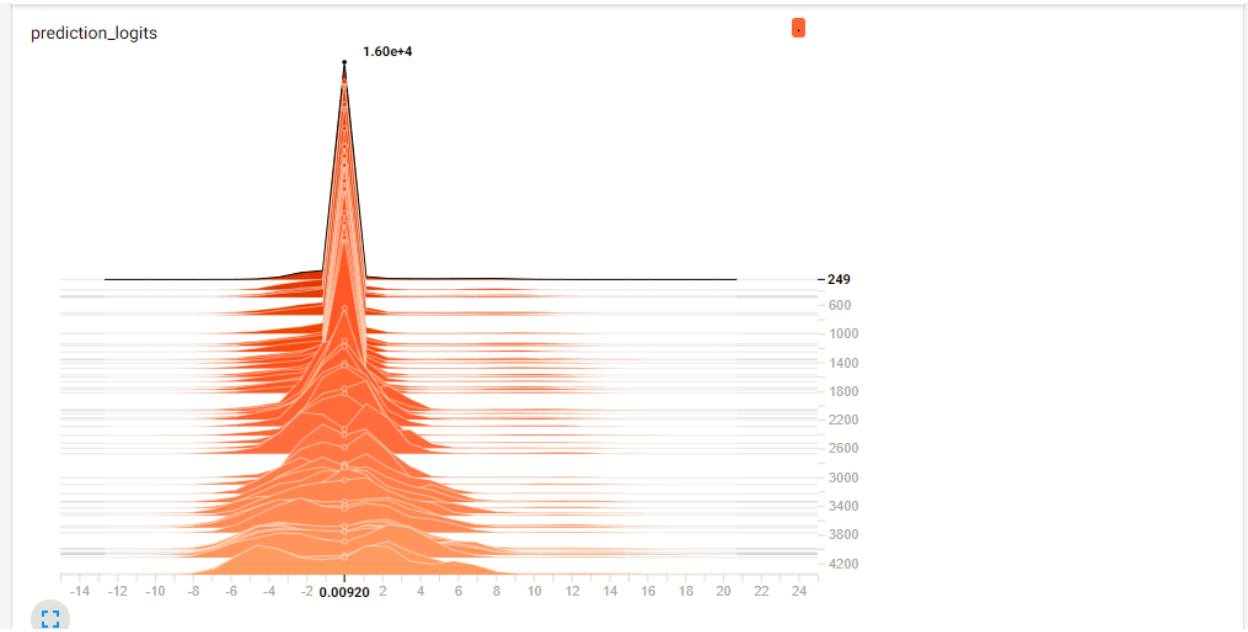


ii. בגרף לעיל קל לראות שהloss יורד ככל שהstepים עולים, בדיוק כפי שהיינו מצפים.



בגרף לעיל ניתן לראות שהאנטרופיה יורדת ככל שהstepים עולים.





בשני הגרפים לעיל ניתן לראות שההיסטוגרמה של הפרדיקציות הופכת לשטוחה ורחבה ככל שהstepים עולים.

<u>ניתוח הגרפים:</u>

האנטרופיה למעשה מכמתת את אי הוודאות שיש לנו בdata. מתורת האינפורמציה האנטרופיה למעשה מייצגת את מספר הביטים הדרוש כדי לתאר את y. לעומת זאת הcross entropy למעשה מכמתת את מספר הביטים הדרושים כדי לעשות encoding לy על ידי אומד $\hat{y}$ (מילים אחרות ללומר שזו דרך למדידת מרחק בין y ל $\hat{y}$).

הcross entropy תמיד גדולה מהאנטרופיה, ולכן מיזעור שלה למעשה מאפשר לנו להשתמש במספר ביטים גדול יותר כדי לתאר את y. ניתן לראות זאת בצורה יפה בהיסטוגרמה, שכן ככל שהצעדים עולים כך השונות בהיסטוגרמה שלנו גדולה יותר והיא רחבה יותר, כלומר אנחנו מתארים יותר ערכים ומתקרבים לתיאור הyים האמיתיים.

סעיף d

Confusion Matrix

Gold\Guess	PER	ORG	LOC	MISC	O
PER	2954	73	36	16	70
ORG	106	1754	51	93	88
LOC	34	162	1845	22	31
MISC	34	60	39	1056	79
O	35	73	22	36	42593

לפי הConfusion Matrix ניתן לראות כי באופן יחסי המודל מנבא באופן לא רע, כלומר רוב האנשים מזוהים כאנשים, רוב המקומות מזוהים כמקומות, רוב ה"שונות" מזוהים כ"שונות" ורוב הO מזוהים כO. לעומת זאת, הרבה ארגונים מזוהים כאנשים, וכן הרבה "מקומות" מזוהים כ"ארגונים".

ניתן לראות תימוכין בקלות ע"י אותו משפט שבדקנו במודל הקודם. קודם לכן, נשים לב כי המגבלות שתיארנו עבור מודל הRNN עדיין קיימות גם עבור GRU.

X	May	15	V	Duke	of	Norfolk	‘s	XI	(	at	Arundel	)
Y*	O	O	O	ORG	ORG	ORG	ORG	ORG	O	O	LOC	O
Y’	O	O	O	ORG	ORG	ORG	O	ORG	O	O	ORG	O
P	1.00	1.00	1.00	1.00	0.99	1.00	0.80	0.98	1.00	1.00	0.62	0.99

במשפט לעיל ניתן לראות שטעינו עבור LOC שחזינו כORG, ניתן לראות שגם ההסתברות נמוכה 0.62, להבדיל מהמקרים בהם צדקנו בהם ההסתברות מובהקת, לפחות 0.98. בנוסף ניתן לראות שהמגבלות של המודל עדיין קיימות, שכן היינו רוצים לעודד תיוגים זהים למילים שכנות וכאן שוב טעינו בO בתוך יער של ORG.

נראה דוגמא נוספת שמדגימה את אותם העקרונות:

X	Shearer	‘s	Euro	96	striking	partner	Teddy	Sheringham	withdrew	from	the	squad	with	an	injury	on	Friday	.
Y*	PER	O	MISC	MISC	O	O	PER	PER	O	O	O	O	O	O	O	O	O	O
Y’	ORG	O	MISC	MISC	O	O	O	ORG	O	O	O	O	O	O	O	O	O	O
P	0.41	0.91	0.86	0.89	1.00	1.00	0.42	0.74	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00