בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Assignment 2 – Window-based Tagging – part 4

בחלק זה של המשימה, נדרשנו להוסיף את ה3- suffixו וה3- prefix של המילים שנמצאות בדאטה ולבצע פעולת חיבור ב forword בין החמישיה של המילים האמיתיות מהדאטה לבין התחילית והסיפא של כל אחת מהמילים,
ובנוסף, נתבקשנו לאמן אותם בשני המודלים, במודל של החלק הראשון, או במודל של החלק השלישי שבו קיבלנו וקטורים מאומנים מראש ולראות היכן האחוזים גבוהים יותר- ז"א איזה מודל יותר טוב.

כדי לממש את זה בצורה הטובה ביוצר פעלנו כך:

1. בזמן טעינת הדאטה בנינו אוצר מילים מכל המילים שנצפו בtrain. (כל המילים באוצר המילים שלנו הפכו ל-lower בכ"מ כדי לשפר ביצועים). נקרא vocab.
2. בנוסף ל vocab שיצרנו ולword_to_idx שיצרנו ב2 המשימות הקודמות הוספנו מילון:idx_to_prre_suf, זה מילון שהkey בו הוא האינדקס של המילה האמיתית, והvalue הוא tuple: (idx_prefix,idx_sufix) .
3. ובמידה ואנחנו משתמשים במודל בחלק השלישי שבו קיבלנו משקולות מאומנות, טענו גם את אוצר המילים שקיבלנו מראש (גם הן lower), נקרא yoav_vocab.
4. איחדנו את 2 הקבוצות כך שלא יהיו כפילויות בין המילים vocab_union, וגם משם לקחנו את הsuffix והprefix של המילים.
5. דרסנו את פונקציית הforword שיצרנו קודם, ומימשנו אותה שונה בצורה הבאה:
   א.יצרנו שני tensores אחד של suffix ואחד של prefix .
   ב. סכמנו כל מילה בtensor עם התחילית והסיפא שלה וככה חיברנו ויצרנו tensorחדש ואותו אנחנו מכניסים לרשת.

6. כאשר מסתכלים על המודל של החלק השלישי, אנחנו עושים את אותו התהליך אך, גודל מטריצת המשקלים גדלה מכיוון שהוספנו את התחילית והסיפא של כל מילה.
בנוסף מכיוון שגילינו שיכול להיות שבזמן הvalidation יש מילה שכביכול נמצאת בword_to_idx , אך בפועל במילון הidx_pre_suf לא נמצאת.וזאת מכיוון שמילה זו היא תחילית או סיפא של מילה שהייתה ב itrain ולכן הוספנו גם את התחיליות והסיפא למילון הidx_pre_suf. כמו כן, כאשר נצפית מילה בסט הוולידציה, פעלנו באותו אופן כמו בחלק הראשון והשלישי- שייכנו אותה לווקטור של המילה הריקה.

היפר- פרמטרים

לאחר מספר נסיונות ובדיקת פרמטרים שונים הגענו למסקנה שהפרמטרים הבאים הם האידיאלים בשביל לקבל את אחוזי הדיוק הגבוהים ביותר.

גילינו שהמודל שהביא את התוצאות הטובות ביותר גם לpos וגם לner (או עם המקולות המאומנות או בלי)

POS:

HIDDEN_LAYER = 110, EPOCHS = 10, LR = 0.01 ,BATCH_SIZE = 100

הוספנו dropout עם הסתברות של 0.5. הפעלנו אותו אחרי חישוב השכבה הלינארית הראשונה.

ההרצה ללא שימוש בטבלת האמבדינג נתנה אחוזים טובים יותר ולכן בחרנו בה כדי לייצר את קובץ הtest.

סה"כ קיבלנו 89.5% הצלחה על סט הבדיקה.

<u>NER:</u>

HIDDEN_LAYER = 110, EPOCHS = 30, LR = 0.01 ,BATCH_SIZE = 30

בנוסף, מכיוון שהדאטה של NER לא יציב- תגית 'O' משויכת לרוב המילים, הוספנו אלמנטים שיעזרו להתמודד נכון יותר עם דאטה שכזה.

דבר ראשון שעשינו הוא (כפי שנתבקשנו בתרגיל) לחשב את אחוזי הדיוק על המודל בזמן הוולידציה ללא התחשבות בהצלחות על תגית 'O'.

בנוסף, כשהגדרנו את הלוס להיות CrossEntropy, הוספנו באתחול משקולות עבור כל תגית, כך שכל תגית קיבלה את המשקולת 1.0 ותגית 'O' קיבלה את המשקולת 0.1. כך למעשה הורדנו משקל משמעותי מהתגית הדומיננטית בדאטה ונתנו הזדמנות לתגיות האחרות להילמד טוב יותר.

הוספת המשקלים שיפרה משמעותית את אחוזי ההצלחה.

דבר אחרון שהוספנו והוביל לשיפור הוא dropout בהסתברות 0.5. לאחר חישוב השכבה הראשונה ברשת ביצענו את הdropout והעלמנו 50 אחוז מהנוירונים.

בסופו של דבר קיבלנו 76% הצלחה על סט הבדיקה.

לאחר שאימנו את המודל, ובחרנו את הפרמטרים שהביאו לתוצאות הטובות ביותר לנו, ראינו שהגענו לתוצאות יותר טובות מבלי להוסיף את הוקטורים המאומנים מראש.

<u>ניתוח כל המצבים:</u>

במהלך כתיבת הקוד חשבנו שאכן ה substring יעזור לנו וגילינו שהיה שינוי מזערי בpos ושינוי יותר משמעותי בner . עבור הner היה שיפור משמעותי יותר מכיוון שהדאטה פחות יציב והוספת התייחסות לsubstring נתנה יותר מידע לתיוג נכון ולעומת זאת ב pos כבר היה תיוג יחסית נכון ותוספת המידע לא שיפרה יותר מדי.

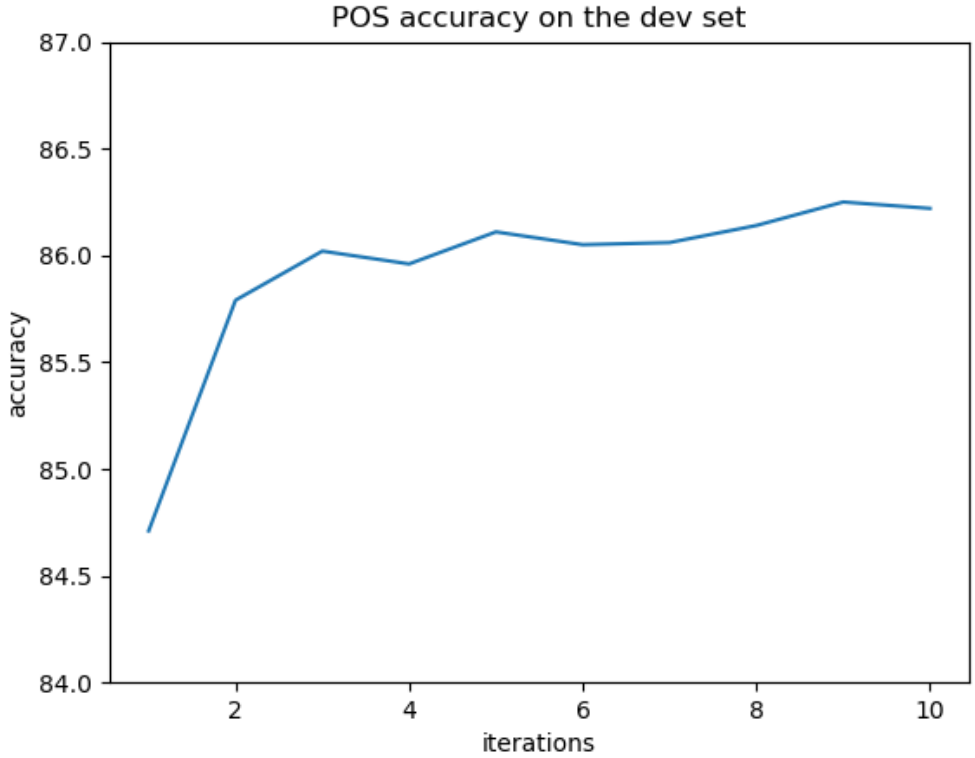
*הערה כפי שכתבנו בחלק 3 אם היינו מתייחסים למילה שהופיעה ב valid ולא בtrain כפי שציינו שם,היינו מקבלים תוצאות טובות יותר.

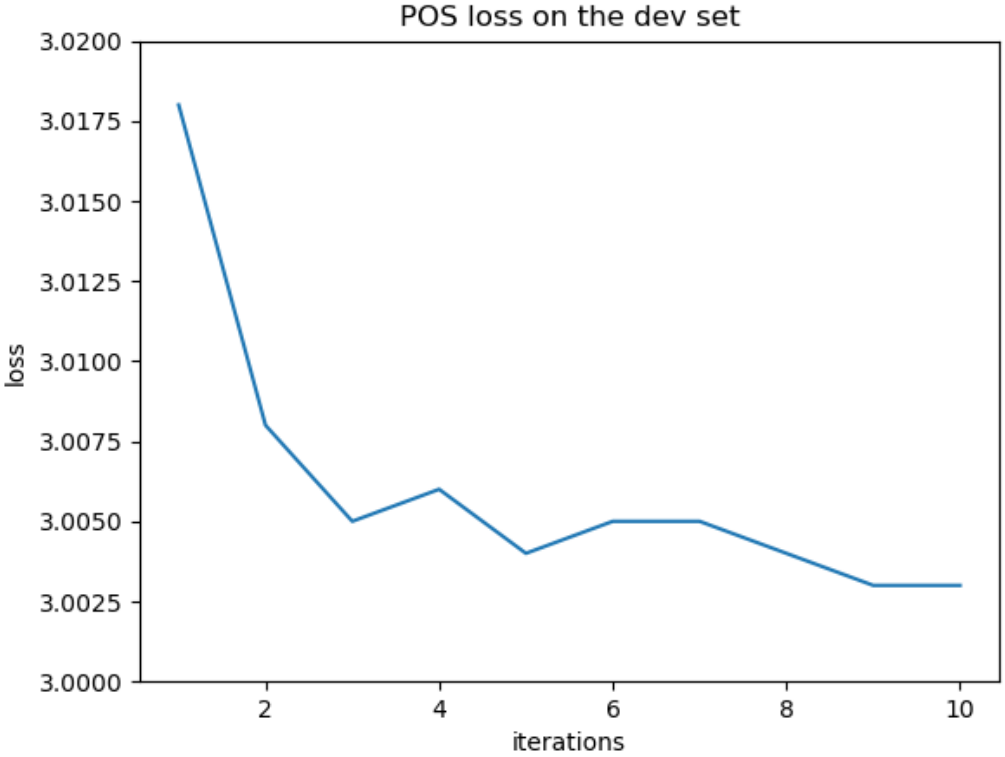
שירה יאיר 315389759
אסנת אקרמן 315747204

בס"ד

גרפים

POS- with pre-trained embedding:

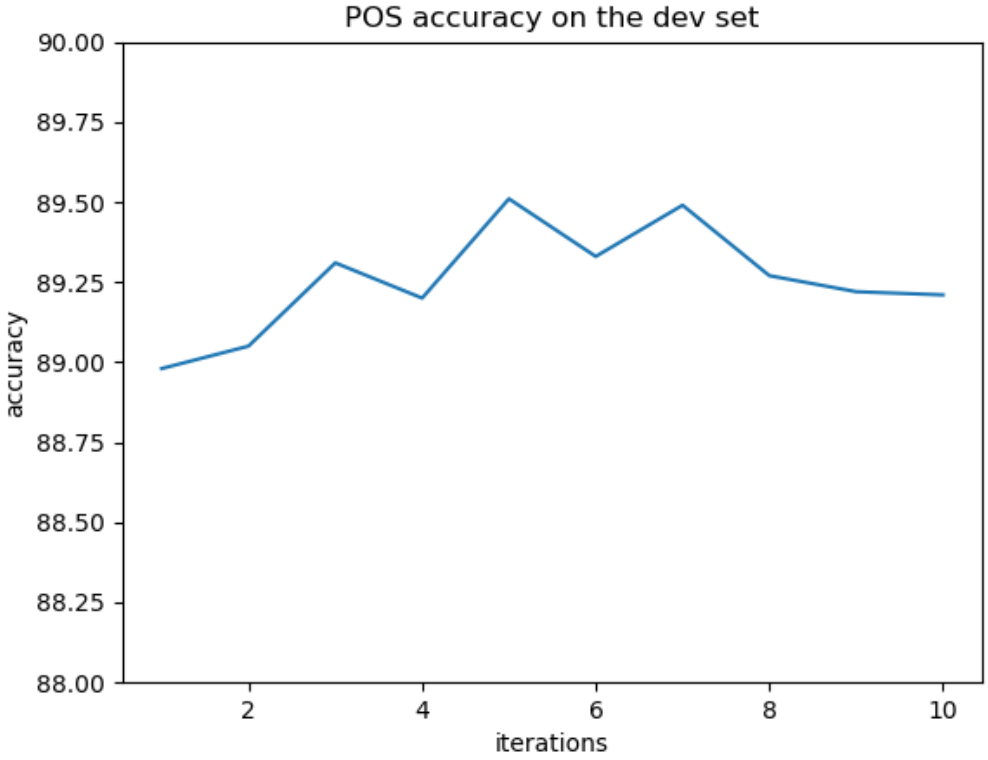


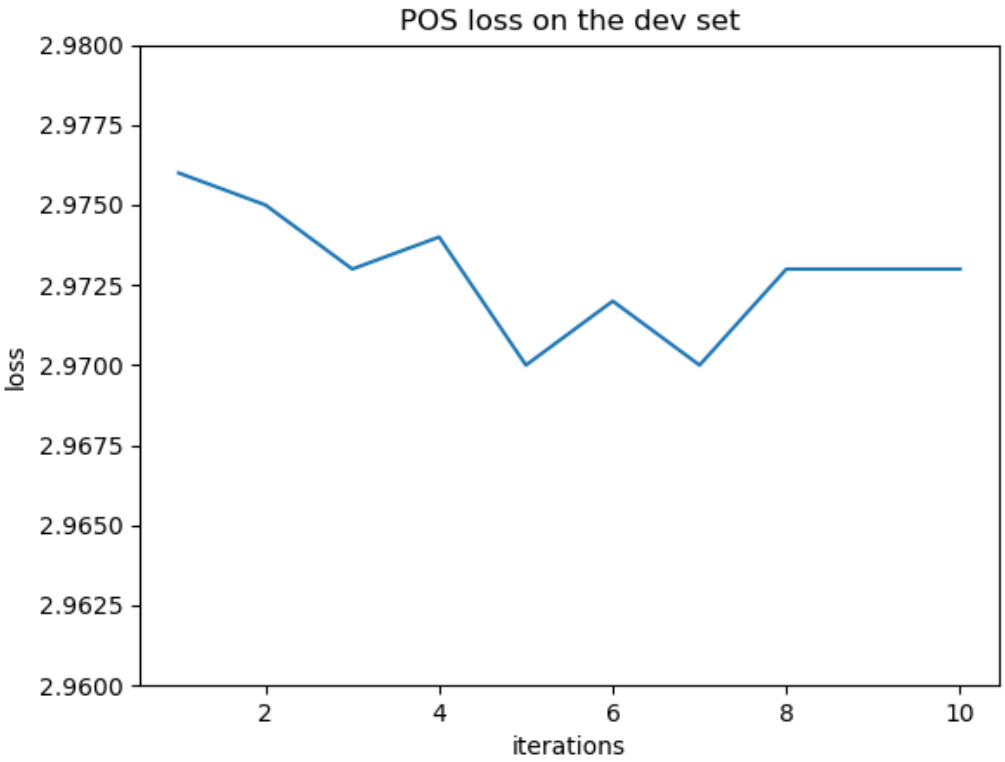


שירה יאיר 315389759
אסנת אקרמן 315747204

בס"ד

POS- no pre-trained embedding:

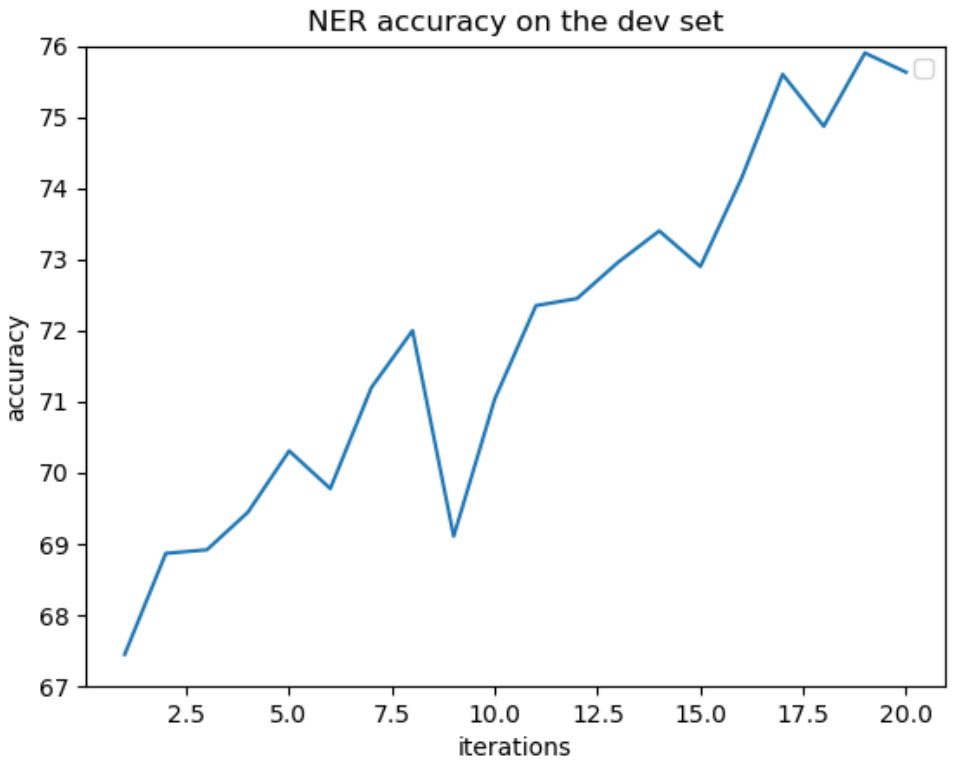


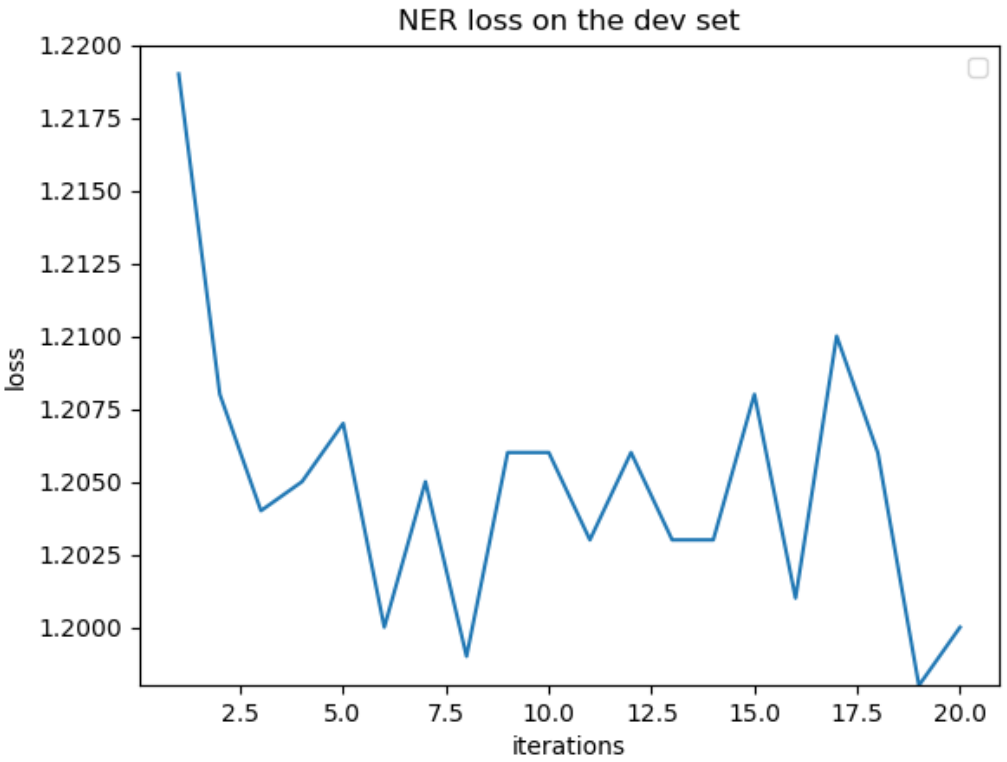


שירה יאיר 315389759
אסנת אקרמן 315747204

בס"ד

NER- with pre-trained embedding:

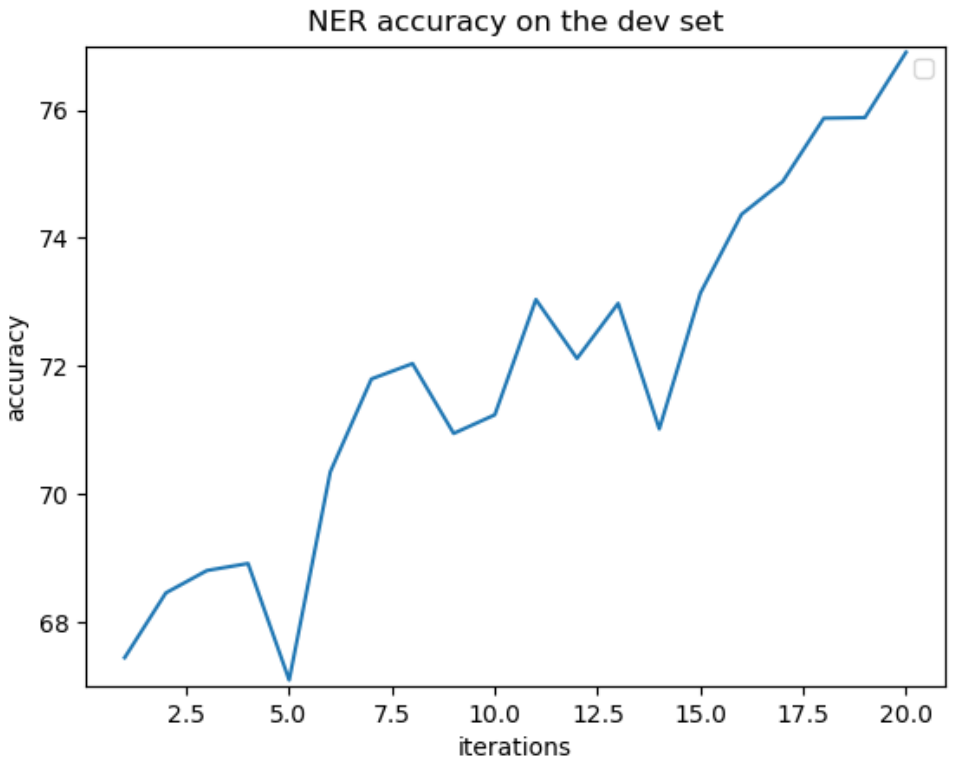


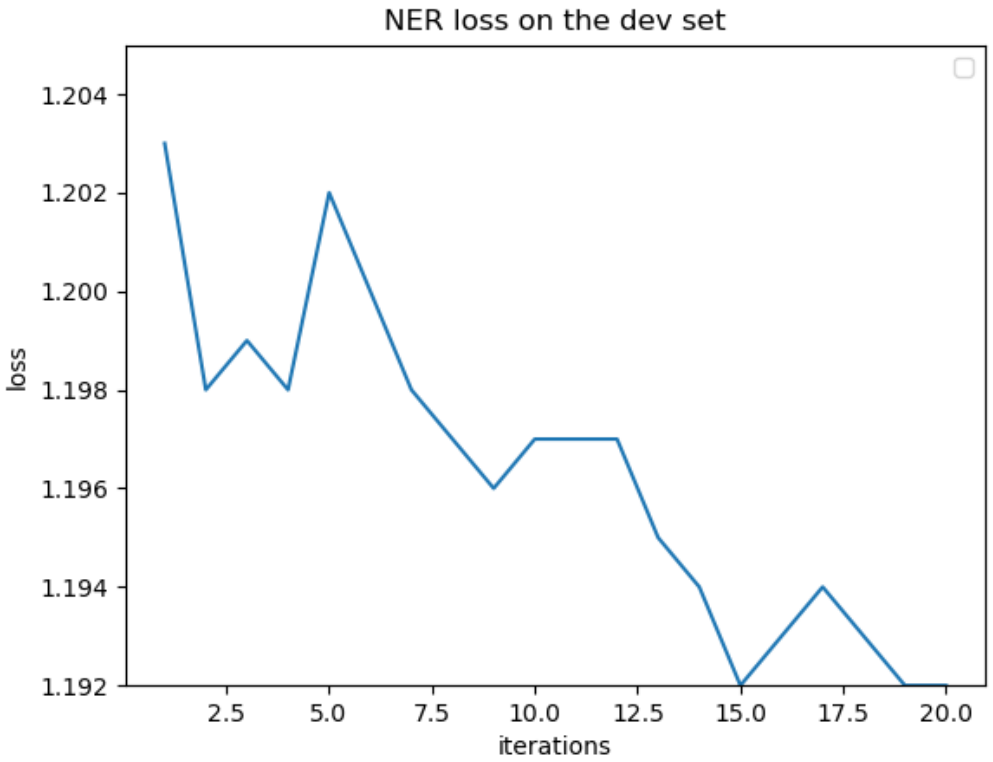


שירה יאיר 315389759
אסנת אקרמן 315747204

בס"ד

NER-no pre-trained embedding:





בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204