בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Assignment 2 – Window-based Tagging – part 3

בחלק זה של המשימה, נדרשנו לטעון ווקטורים מאומנים מראש עבור טבלת האמבדינג במקום לאתחל אותה אקראית.

כדי לממש את זה בצורה הטובה ביוצר פעלנו כך:

1. טענו את המילים בtrain וחילקנו לכל אחד אינדקס שנשמר במילון word_to_idx.
2. טענו את המילים של יואב מ-vocab.txt ושמרנו אותו בset שנקרא yoav_vocab.
3. עברנו על כל מילה בyoav_vocab כך ש:
   a. אם היא נמצאת כבר ב-word_to_idx, המשך הלאה.
   b. אם היא לא נמצאת, הוסף אותה למילון עם האינדקס הבא בתור.
4. כך הרחבנו את המילון שלנו שיכלול גם את כל המילים בעלות ווקטור ממושקל מראש.
5. כעת בנינו מטריצה עם len(word_to_idx) שורות ו-50 עמודות. נקראת weights_matrix.
6. עבור כל 'מילה' ו'אינדקס' ב-word_to_idx :
   a. אם קיים ווקטור ממושקל עבור ה'מילה', הצב אותו במקום 'אינדקס' במטריצת המשקולות
   b. אחרת, הצב ווקטור מאותחל רנדומית בהתפלגות אחידה במקום 'אינדקס' במטריצת המשקולות
7. את מטריצת המשקולות שלחנו למודל שבזמן בניית שכבת האמבדינג שלך את הטבלה כפרמטר weights.

אופן הפעולה הזה יצר מצב שבזמן הtrain עבור מילה שקיבלנו מראש יש ווקטור מאומן מטבלת האמבדינג, ועבור מילה חדשה יש ווקטור מאותחל אקראית.

כאשר נצפית מילה בסט הוולידציה, פעלנו באותו אופן כמו בחלק הראשון- שייכנו אותה לווקטור של המילה הריקה.

בזמן טעינת המילים מהtrain ויצירת word_to_idx תרגמנו את כל המילים להיות lower, זאת על מנת לשפר ביצועים.

היפר- פרמטרים

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

POS:

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

קיבלנו 88.6% הצלחה על סט הבדיקה.

NER:

HIDDEN_LAYER = 50, EPOCHS = 30, LR = 0.01 ,BATCH_SIZE = 100

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'.

בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

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 כל נוירון בהסתברות של חצי.

בסופו של דבר קיבלנו 75% הצלחה על סט הבדיקה.

המסקנה היא ששימוש בטבלת pre-embedding שיפר מעט אך לא משמעותית את אחוזי הדיוק על סט הוולידציה.

**הערה – ביום האחרון להגשה עלה לנו רעיון בנוגע להתייחסות למילה בvalid שלא נמצאת בtrain .
עד כה השתמשנו בווקטור המילה הריקה כדי לייצג את כל המילים הללו, אך לדעתנו דרך נכונה יותר לבצע זאת היא לשלוף מתוך טבלת האמבדינג את הווקטור הקיים הקרוב ביותר אל המילה המבוקשת. בעזרת הפונקציה שהתבקשנו לממש בחלק 2 ניתן לחשב ולשלוף את המילה הקרובה ביותר וכך הסיכוי לתייג נכון עולה.
במקום לאחד את כל המילים הלא ידועות בלי קשר לוגי אמיתי ביניהן, נתייג אותן למילה הקרובה ביותר מבחינה לוגית.
לצערנו נשארה שעה להגשה ואנחנו לא מספיקות לממש את הרעיון ולבדוק את התוצאות.

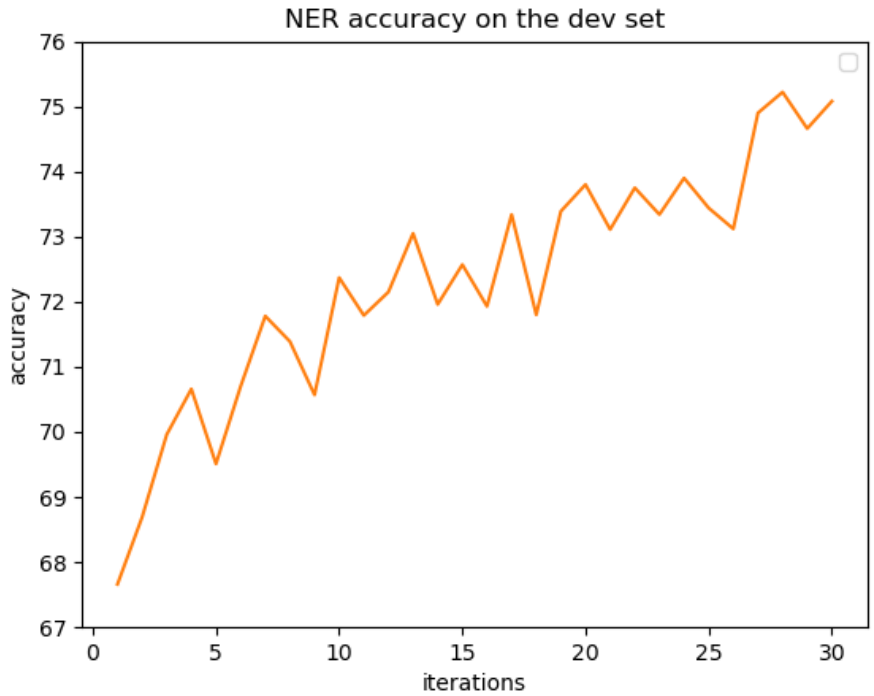
ייתכן ששיפור זה היה מקפיץ את אחוזי ההצלחה הרבה מעל מה שצפינו.

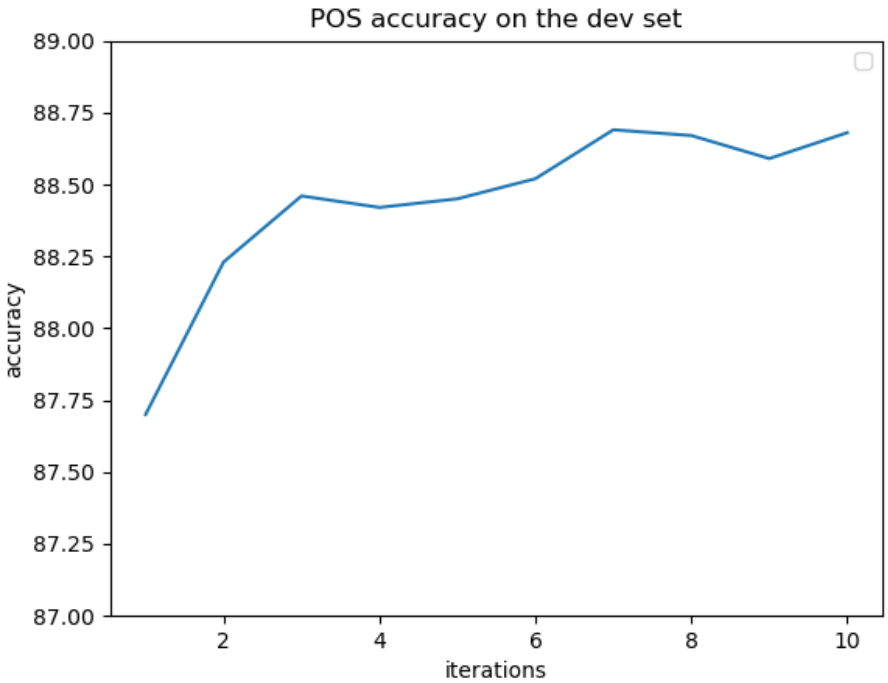
בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204

גרפים

Accuracy





שירה יאיר 315389759
אסנת אקרמן 315747204

בס"ד

Loss: