בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Assignment 2 – Window-based Tagging – part 1

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% הצלחה על סט הבדיקה.

NER:

HIDDEN_LAYER = 110, 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'.

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 74% הצלחה על סט הבדיקה.

עיבוד הדאטה

כדי לשפר את הביצועים, בזמן טעינת הדאטה ויצירת אוצר המילים בשביל טבלת האמבדינג- הפכנו את כל המילים ל-LOWER CASE כך שה-"רעש" של מילים זהות אך בכתיב שונה לא יפריע ללמידה נכונה יותר.

דבר נוסף שעשינו הוא להפוך כל ספרה למילה 'DG'. לדוגמה המספר 12 מתורגם למילה DGDG. זאת על מנת לטשטש דיוק יתר על מספרים מסוימים וניסיון להקל על המודל לקבץ את כל המספרים לאותו תיוג.

לאחר הפעלת אפשרויות אלו קיבלנו שיפור משמעותי באחוזי הדיוק.

בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204

Considerations

1. עבור מילים שנמצאות בVALID אך לא בTRAIN יצרנו מלכתחילה מילה ריקה שאותה הוספנו לאוצר המילים של טבלת האמבדינג, כך שבכל פעם שנתקלנו במילה כזו, שלפנו מהטבלה את הווקטור שמשויך למילה הריקה. מילה זו אומנה יחד עם כל שאר הווקטורים בטבלה.

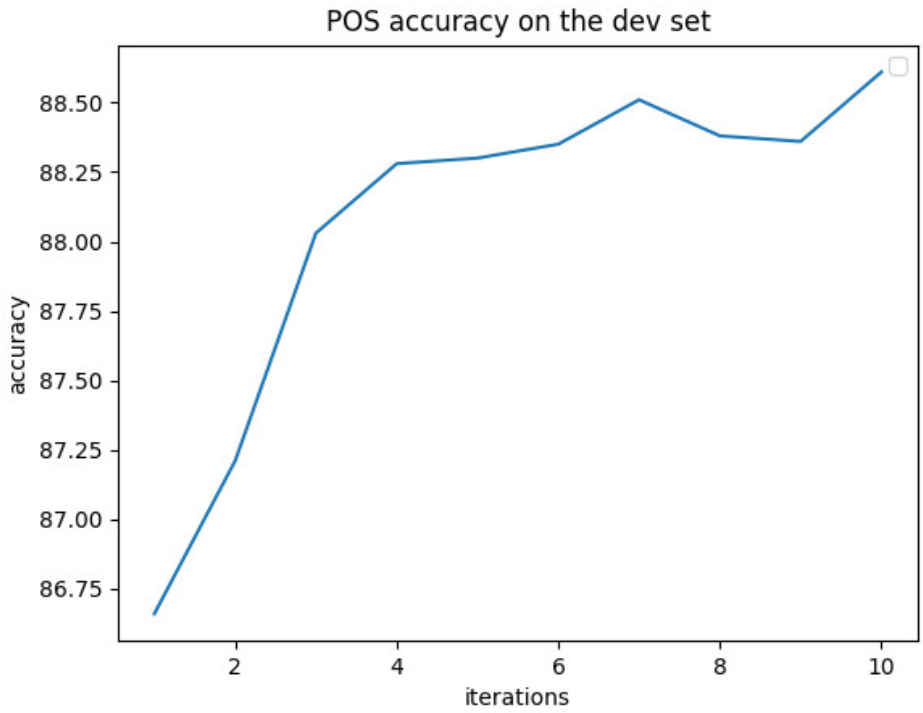
2. כדי ליצור חלון בגודל המתאים עבור המילים הראשונות והאחרונות במשפט הוספנו לכל משפט בתחילתו את המילים <START>,<START> ובסוף המשפט הוספנו <END>,<END>. כך שבתוספת המילים האלו ניתן ליצור חלון בגודל הנדרש לכל מילה במשפט. לדוגמה:
עבור המשפט Hello World נייצג את המילים במשפט הזה באופן הבא:
<START>, <START>, **Hello**, World, <END>
<START>, Hello, **World**, <END>, <END>
המילים הנוספות <START>, <END> נכנסו לאוצר המילים שבנינו עבור המודל כך שגם הן קיבלו אינדקס וייצוג בטבלת האמבדינג.

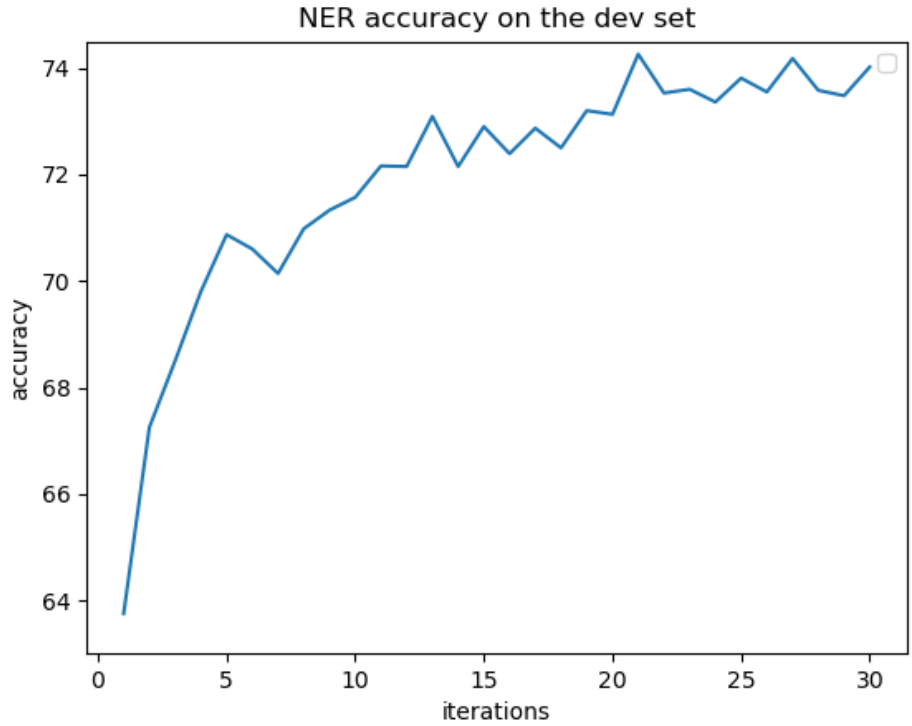
בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204

גרפים

Accuracy:





בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204

:loss