בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Part 3: BiLSTM Tagger- Assigment 3:

בחלק זה של המשימה, נדרשנו לממש 2-layer biLSTM כך שהקלט לBILSTM מיוצג כל סעיף בצורה אחרת.

בחלק זה השתמשנו בדוגמאות מהתרגיל הקודם (POS,NER). בחלק של טעינת הדאטה השתמשנו בחלקים של קוד מהתרגיל הקודם. כדי לממש את זה בצורה הטובה ביותר פעלנו כך:

1. בזמן טעינת הדאטה בנינו אוצר מילים מכל המילים שנצפו בtrain, נקרא vocab.
2. בנוסף ל vocab שיצרנו בנינו מילון שכל מילה מיוצגת באינדקס שנקרא word_to_idx ולlaybel בנינו מילון שנקרא laybel_to_idx .
3. במידה ואנחנו משתמשים בייצוג הסעיף השלישי הוספנו מילון לכל הprefix והsufix של המילה שנקרא idx_pre_suf וכמובן שהוספנו לvocab את התחיליות והסייפות.
4. כמו כן, כאשר נצפית מילה בסט הוולידציה, שלא היתה באימון, הגרלנו אינדקס של מילה כלשהי מהאוצר מילים שכבר יש.
5. במידה ואנחנו משתמשים בייצוג החלק השני אזי עשינו מילון נוסף של idx_word כי אנחנו ממירים כל מילה שקיבלנו בsentence לייצוג באותיות, ולכן נצטרך לקבל בחזרה את המילה עצמה כדי לקבל את האותיות המרכיבות אותה, וכל אות במילה קיבלה אינדקס שהוא מס' הASCII שלה.

נראה את הייצוגים השונים של כל קלט לBILSTM ולאחר מכן את הFLOW של המודל.

- בסעיף א כל מילה בsentence קיבלה ייצוג בטבלת הEMBEDDING ושירשרנו אתכל הייצוגים יחד והכנסנו לקלט לרשת.
- בסעיף ב לקחנו כל מילה בsentence ופירקנו אותה לאותיות המרכיבות אותה כך שכל אות במילה קיבלה אינדקס(מספר ASCII), ובנינו טבלת אמבדינג שאוצר המילים שלה הם כל אותיות הASCII. שירשרנו את ווקטורי האותיות והכנסנו כקלט לLSTM רגיל. לאחר מכן לקחנו מהפלט רק את הlast_state של כל מילה והכנסנו לBILSTM.
- בסעיף ג כל מילה בsentence קיבלה ייצוג בטבלת הEMBEDDING ובנוסף כל רישא וסיפא של המילה קיבלה ייצוג בטבלת EMBEDDING ולאחר מכן סכמנו לכל ווקטור של מילה את הווקטורים שמייצגים את הprefix והsuffix שלה בטבלת הEMBEDDING והכנסנו כקלט לרשת.
- בסעיף ד השתמשנו גם בייצוג של סעיף א' וגם בייצוג של סעיף ב'. לקחנו את הייצוג של המילה מסעיף א' ושירשרנו אליה את הייצוג מסעיף ב' והכנסנו כקלט לרשת.

לאחר שקיבלנו את הייצוג המתאים, הכנסנו אותו כקלט לBILSTM עם 2 שכבות ואת הפלט שקיבלנו הכנסנו לשכבה לינארית וקיבלנו את הפרדיקציה.

בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

היפר פרמטרים:

Repr A(POS) :

EMBEDDING_DIM = 100,**HIDDEN_LAYER_LSTM** = 110, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 500, **DEV_BATCH_SIZE** = 50.

Repr A(NER) :

EMBEDDING_DIM = 100,**HIDDEN_LAYER_LSTM** = 110, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 500**, DEV_BATCH_SIZE** = 50.

שימוש ב**DROPOUT**: no/no
שימוש ב**איתחול יוניפורמי לEMBEDDING**:yes/yes

Repr B (POS):

EMBEDDING_DIM = 50,**HIDDEN_LAYER_LSTM** = 100**, EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 50, **DEV_BATCH_SIZE** = 500, **CHAR_DIM_EMBED** = 100, **LSTM_CHAR_OUTPUT_DIM** = 100.

Repr B(NER) :

EMBEDDING_DIM = 50,**HIDDEN_LAYER_LSTM** = 100, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 50, **DEV_BATCH_SIZE** = 500**, CHAR_DIM_EMBED** = 100, **LSTM_CHAR_OUTPUT_DIM** = 100.

שימוש ב**DROPOUT**:no/no
שימוש ב**איתחול יוניפורמי לEMBEDDING**:no/yes

Represent C(POS) :

EMBEDDING_DIM = 30**,HIDDEN_LAYER_LSTM** = 100, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 500**, DEV_BATCH_SIZE** = 50.

Represent C(NER) :

EMBEDDING_DIM = 100,**HIDDEN_LAYER_LSTM** = 100, **EPOCHS** = 5**, LR** = 0.01 **BATCH_SIZE** = 50, **DEV_BATCH_SIZE** = 500.

שימוש ב**DROPOUT**: no/no
שימוש ב**איתחול יוניפורמי לEMBEDDING**: no/yes

בס"ד
שירה יאיר 315389759
אסנת אקרמן 315747204

Represent D(POS) :

EMBEDDING_DIM = 50,**HIDDEN_LAYER_LSTM** = 100, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 50, **DEV_BATCH_SIZE** = 500, **CHAR_DIM_EMBED** = 100, **LSTM_CHAR_OUTPUT_DIM** = 100.

Represent D(NER) :

EMBEDDING_DIM = 50,**HIDDEN_LAYER_LSTM** = 100, **EPOCHS** = 5, **LR** = 0.01 **BATCH_SIZE** = 50, **DEV_BATCH_SIZE** = 500, **CHAR_DIM_EMBED** = 100, **LSTM_CHAR_OUTPUT_DIM** = 100.

שימוש בDROPOUT:no/no
שימוש ב**איתחול יוניפורמי לEMBEDDING**:no/yes

- הערה נוספת- מכיוון שהדאטה של NER לא יציב- תגית 'O' משויכת לרוב המילים, הוספנו אלמנט שיעזור להתמודד נכון יותר עם דאטה שכזה.
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.
- הערה – התמודדנו עם בעיית אורכי הרצפים ע"י הספרייה המובנית של Pytorch .
כאשר חישבנו את הloss התעלמנו מהריפוד של y ע"י הדגל ignoe_index כדי שהloss יתעלם מערכים לא נכונים.

לסיכום המודל שהניב את התוצאות הטובות ביותר:

אצלנו דווקא מודל C עם ההתייחסות לרישות וסיפות של המילה נתן לנו את האחוזים הגבוהים ביותר עבור 2 סוגי הדאטה. בPOS הגענו ל 94% וב-NER הגענו ל 93%.

בס"ד

שירה יאיר 315389759
אסנת אקרמן 315747204

גרפים-