שאלה 1:

סעיף a:

1. משפטים עם מילים בעלות דוד משמעות שמית:

Ford makes the best cars.

המילה Ford במשפט מכוונת ליצרנית המכוניות, אף אותה מילה יכולה לשמש גם כשם של בן אדם.

I can get to JFK with a taxi.

המילה JFK מתייחסת לשדה התעופה בניו יורק, אך היא גם יכולה לשמש ככינוי שך נשיא ארה"ב לשעבר.

ii על ידי שימוש בפיצ'רים למעט המילה עצמה ניתן לתת לה הקשר בתוך המשפט, שיעזור לפתור את בעיית הדו-משמעות. בנוסף, מילים שהן named entities עלולות להיות נדירות באימון ולכן הפיצ'רים עוזרים בהכללה.

iii דוגמאות לפיצ'רים: (נניח כי המילה היא wi, במיקום i במשפט)

* Wi-1 is part of a named entity **and** wi starts with a capital letter.
* Wi-1 is a prepesition that refers to a location (e.g. at, inside, out of…)

סעיף b:

מימדים:

, ,

סיבוכיות חיזוי עבור משפט באורך T: (נניח גודל חלון w)

חישוב : מכפילים פעמים את וקטור המילה (one hot) עם המטריצה E המכילה וקטורי קידוד באורך D – סה"כ .

חישוב : .

חישוב **:**

סה"כ עבור מילה בודדת במשפט:

ולכן עבור משפט באורך T:

סעיף d:

התוצאות הטובות ביותר עם דיוק של 83% F1:

2019-06-08 11:25:13,782:DEBUG: Token-level confusion matrix:

go\gu PER ORG LOC MISC O

PER 2967 49 55 12 66

ORG 143 1638 120 53 138

LOC 60 94 1873 20 47

MISC 40 54 52 1009 113

O 58 54 15 26 42606

2019-06-08 11:25:13,782:DEBUG: Token-level scores:

label acc prec rec f1

PER 0.99 0.91 0.94 0.92

ORG 0.99 0.87 0.78 0.82

LOC 0.99 0.89 0.89 0.89

MISC 0.99 0.90 0.80 0.85

O 0.99 0.99 1.00 0.99

micro 0.99 0.98 0.98 0.98

macro 0.99 0.91 0.88 0.90

not-O 0.99 0.89 0.87 0.88

2019-06-08 11:25:13,782:INFO: Entity level P/R/F1: 0.82/0.85/0.83

ניתן להסיק מה-confusion matrix שהטעויות המשמעותיות הן:

True class: ORG, predicted class: PER,LOC,O

True class: MISC predicted class: O

באופן כללי רואים שהמודל הכי מתקשה עם זיהוי ORG לפי מספר השגיאות בשורה זאת בטבלה.

מגבלות המודל:

1. המודל לא משתמש בהמשכיות של תיוגים שביצע על מילים קודמות, כלומר הוא לא מתחשב בעובדה שרצף של מילים יכול להוות שם אחד של אותה יישות ולכן יש לתייג את כולן בצורה זהה. לדוגמה:

x : Jordan won the Samsung Tel Aviv marathon

y': LOC O O ORG ORG LOC O

האירוע "מרתון סמסונג תל אביב" מהווה ישות אחד, אך המודל תייג חלק מהמילים בצורה שונה.

1. במשפטים ארוכים המודל אינו מסוגל להתחשב בהקשר של מילה כדי לפתור דו משמעות. לדוגמה:

x : in his interview, Jordan said that he was very excited from his victory

y': O O O LOC O O O O O O O O O

ברור שהמשפט עוסק בראיון הניצחון של Jordan במרתון, אך המודל לא מסוגל להבחין בכך ולכן מתייג את השם ב LOC.