עיבוד שפה טבעית - פרויקט מסכם

תיאור המשימה

בפרויקט תממשו מודל לתרגום פסקאות מגרמנית לאנגלית.

לצורך בחינת הביצועים, נשתמש במדד BLUE, שהוא מדד סטנדרטי למדידת ביצועים במשימות תרגום. המדד יחושב עבור כל אחד מהפסקאות המתורגמות, וציון הדיוק יהיה ממוצע ציוני BLEU שהתקבלו עבור תרגומי כל הפסקאות.

לצורך חישוב מדד הBLEU ניתן לכם קוד לדוגמא בו נשתמש לחישוב ביצועי המודלים שלכם. עליכם לוודא שהקוד מסוגל לרוץ על הקבצים אותם אתם מגישים.

הסבר על מבנה הציון בתרגיל:

- **40%** - מימוש מלא של מודל תרגום מגרמנית לאנגלית. על המודל לקבל ציון BLEU ממוצע של לפחות 30% בתיוג הקובץ val.labeled כאשר הוא מתאמן על הקובץ train.labeled בלבד.
- **10%** - עמידה בפורמט: מימוש tagger, המקבל קובץ בפורמט comp.unlabeld, מתייג אותו ומוציא קובץ בשם comp_id1_id2.labeld בפורמט המדויק.
- **30%** - תחרות מבוססת BLEU בתיוג קובץ התחרות comp.unlabeld.
- **20%** - כתיבת דו"ח מפורט, תמציתי, ועניני (בן עד 3 עמודים) אשר יציג את העבודה. הדו"ח יכלול את כל הסעיפים הנדרשים ויעמוד בתנאי ההגשה.

נתונים:

הסבר על הקבצים המצורפים –

1. train.labeled – קובץ המכיל 10000 זוגות של פסקאות בגרמנית ובאנגלית. כל זוג משפטים מופרד בשורה ריקה, ולפני כל משפט מופיעה שורה המכילה את השפה בה הוא נכתב.
   עליכם להשתמש בקובץ זה בשלב האימון.
2. val.labeled – קובץ המכיל 1000 זוגות של פסקאות בגרמנית ובאנגלית, בפורמט זהה לפורמט של הקובץ הקודם.
   עליכם להשתמש בקובץ זה להערכת הביצועים של המודל שלכם.
3. comp.unlabeled – קובץ המכיל 2000 פסקאות בגרמנית בלבד. בנוסף, עבור כל פסקה מופיעים השורשים של כל משפט בה, וכן שניים מה-Modifiers של השורש, אם קיימים כאלו.
4. val.unlabeled – קובץ המכיל את המשפטים מהקובץ val, בפורמט הזהה לקובץ comp.

אימון (**Train**):

תוכלו לאמן כל מודל בסיסי (מגודל base או large), אין להשתמש במודלים שאומנו למשימות אחרות, בפרט לא למשימות תרגום.

אין להשתמש בקובץ התחרות לשום מטרה בשלב האימון או הולדיציה פרט להרצת המודל עליו ושמירת התוצאות לקובץ.

עליכם לאמן מודל בהתבסס על הקובץ המתויג train.labeled ולשמור אותו לזיכרון.

הסקה ותיוג (**Inference**):

עליכם לבנות מתייג אשר מקבל קובץ לא מתויג ומתייג אותו באמצעות המודל שיטען מהזיכרון.

עליכם לתייג את הקובץ val.unlabeled בהתבסס על מודל שאומן על הקובץ train.labeled בלבד.

יש להגיש קובץ val_id1_id2.labeled בפורמט הזהה לחלוטין לקובץ train.labeled.

יש לדווח בדו"ח את ממוצע ציוני ה-BLEU שקיבלתם על הקובץ val_id1_id2.labeled, בהשוואה לקובץ המתויג val.labeled.

שימו לב כי יש לדווח את אחוזי הדיוק בהשוואה לתיוג שבוצע בפועל.

מבחן (**Test**):

עליכם לתייג את הקובץ comp.unlabeled.

יש להגיש קובץ comp_id1_id2.labeled בפורמט הזהה לחלוטין לקובץ train.labeled.

על בסיס התיוג הזה יקבע הציון התחרותי שלכם.

סביבת עבודה:

על הפרויקט לרוץ על המכונה שניתנה לכם בסביבת azureml_py38 ללא התקנת ספריות נוספות.

אין להשתמש בספריות נוספות ללא אישור מסגל הקורס בפורום המיועד לפרויקט.

אין להשתמש בקבצי נתונים שלא סופקו במהלך ההגשות בקורס.

הגשה:

קובץ zip בלבד, בשם Project_123456789_987654321.zip(עבור שני סטודנטים שמספרי הזהות שלהם 123456789 ו 987654321). הקובץ הנ"ל יכלול:

1. דו"ח מפורט, תמציתי, ועניני (בן עד 3 עמודים, ובשם report_123456789_987654321.pdf) אשר יציג את העבודה. הדו"ח יכיל הסברים תמציתיים, דיווח וניתוח תוצאות, ויכלול לפחות:

   1. מספרי תעודות זהות של המחברים, ללא שמות.
   2. תיאור ניסויים אותם ביצעתם במהלך העבודה על הפרויקט.
   3. תיאור האלגוריתם בו השתמשתם לאימון המודל ולתיוג הקובץ val.
   4. אחוז הדיוק שהתקבל על הקובץ val.
   5. תיאור ניסויים אשר ביצעתם לשיפור התוצאות בתחרות.
   6. תיאור האלגוריתם בו השתמשתם לאימון המודל התחרותי ולתיוג הקובץ comp.
   7. אחוז הדיוק הצפוי לקובץ comp.
   8. הסבר על חלוקת העבודה בין שני חברי הקבוצה.

2. קבצים מתויגים בשמות val_id1_id2.labeld ו- comp_id1_id2.labeld.

3. קבצי הקוד של הפרויקט על הקוד להיות מתועד וקריא.

בנוסף, הקוד צריך להיות מסוגל לרוץ על המכונה הוירטואלית שסופקה לצורך הפרויקט. אנא כתבו ממשקי הרצה פשוטים לאימון, מבחן וייצור קבצי התחרות המתויגים.

4. ממשק לתיוג קבצי התחרות בשם generate_comp_tagged המקבל קובץ בפורמט unlabeled, מתייג אותו ומוציא קובץ בשם comp_id1_id2.labeld בפורמט הזהה לפורמט הקובץ train.labeled. בקובץ צריכה להיות פונקציה שמתייגת את קובץ הval ואחרת שמתייגת את קובץ הcomp, ומוציאה כל אחד מהם לקובץ בשם המתאים.

5. בעקבות המשקל הגדול של המודלים המאומנים אין לצרף אותם להגשה, אלא יש להעלות אותם לdrive (google או Microsoft) ולשים קישור שמאפשר הורדה של המודל בדוח.

העתקות:

בשל אופי המשימה והמורכבות שלה, קל לבדוק העתקות של קטעי קוד \ קבצים מלאים. למען הסר הספק אנו מדגישים כי אין להעביר קוד בין סטודנטים, בין אם להגשה ובין אם לא. אין להעתיק קטעי קוד מוכנים מהאינטרנט, ובכלל אין להסתמך על שום מקור אחר לקוד מלבד פרי יצירתכם והחבילות החיצוניות אשר צוינו בסעיף הרלוונטי.