דו״ח סיכום פרויקט

Grammatical Error Correction – Mistake Types Evaluation

הערכת חומרת סוגי שגיאות התחביר השונות באנגלית

אופיר שיפמן, אוגוסט 2019

מנחה : ד״ר עמרי אבנד

תודה מיוחדת וענקית ללשם חשן על הסיוע במהלך הפרויקט

תקציר :

עבודה זו באה על מנת לייצר מדד לחומרתן של טעויות תחביריות שונות בשפה האנגלית בראיית דוברי אנגלית. במסגרתה השתמשנו ב-MTurk על מנת לאסוף נתונים על אודות חומרת הטעויות בראי הקוראים וביצענו רגרסיה לינארית החוזה את ציון המשפט בהתאם לטעויות המופיעות בו ; חילצנו את המשקלות שקיבלה כל סוג טעות ברגרסיה, ואלו מעידות להבנתנו על חומרת הטעות בראי הקוראים. המשפטים נדגמו ממאגר NUCLE וחילוץ המשקלות המעידות על חומרתן בוצע הן בהתבסס הן על תיוגי הטעויות של NUCLE והן על הטעויות המחולצות באמצעות ERRANT.

התוצאות רועשות ולכן יש לקחת אותן בעירבון מוגבל, אך ניתן להצביע על מספר ממצאים ראשוניים בהם העובדה כי מיידעים (determiners), המהווים את הטעות הנפוצה ביותר במאגר, אינם בעיה המפריעה מאוד לקוראים ; טעויות הקשורות בפעלים קיבלו ציון גבוה ונראה שהן חשובות ; ולא קיימת קורלציה בין מידת הנפוצות של טעות לבין מידת החומרה שלה.

במסגרת העבודה יצרנו מספר מאגרי מידע שניתן לעשות בהם שימוש באופנים נוספים, ובראשם מאגר המכיל את משפטי NUCLE כווקטורים של טעויות בעלי ציון מתוקנן המשקף את חומרת הטעויות בראי הקוראים.

צעדי המשך אפשריים לעבודה כוללים שימוש בתוצאות אלו באלגוריתמי GEC, ובמערכות הערכה (evaluation) לאלגוריתמים אלו. זאת, לצד איסוף מידע נוסף לטובת הגדלת מהימנות התוצאות, או הרחבתן למאגרים נוספים ולהקשרים קונקרטיים (כגון הפרדה בין שגיאות המפריעות במאמר אקדמי לעומת שגיאות המפריעות במייל).

רקע:

אלגוריתמים רבים מנסים לבצע תיקון שגיאות דקדוק באנגלית (GEC – Grammatic Error Correction), ואחרים לבצע הערכה (evaluation) לטיב התיקונים. אלו משתמשים בקורפוסים שונים ותיוגים שונים לסוגי הטעויות, ובמדדים שונים על אודות טעויות אלו, בדגש על כמות הטעויות.

נושא שלא נחקר מספיק בתחום זו הוא חומרת הטעויות השונות בראי הקוראים. האם יתכן שישנן טעויות נפוצות שקהילת הGEC עומלת למצוא אלגוריתמים טובים לפתרונן אך אינן באמת מפריעות לקורא הממוצע? האם יתכן שישנן טעויות קטנות ולא נפוצות שהופכות משפט ללא מובן בעליל, או למותיר רושם רע על הכותב? עבודה זו מנסה לשפוך אור על שאלות מסוג זה ולספק מידע על אודות חומרת סוגי הטעויות השונות בראי קרואים דוברי אנגלית. בחרנו להתמקד בעבודה זו בקורפוס NUCLE כקורפוס דגל בתחום; ניתוח הטעויות נעשה על סמך התיוג המקורי של מאגר זה, וכן אוסף הטעויות המחולץ מהמאגר באמצעות מערכת ERRANT, מכיוון שזוהי מערכת אוטומטית ושימושית לחילוץ שגיאות תחביר באופן חוצה קורפוסים. בעתיד, יש מקום לבצע עבודה דומה על קורפוסים ושיטות תיוג נוספים על מנת לאתגר או לאשרר את הממצאים.

שיטה:

את חומרת הטעויות בחרנו להעריך באמצעות העלאת משפטים מ-NUCLE ל-direct assement ב-MTurk. שיטת הערכת משפט בודד התבססה כמה שניתן על עבודתם של Grahm ונוספים שחקרו את ביצוע ההערכה בכלים אלו לטובת משימות תרגום (machine translation). במסגרת זאת משפטים בעלי טעויות מNUCLE עלו להערכה בMTurk תחת השאלה[1]:

The English mistakes in the following text bother me (1 = it doesn't bother me at all, 100 = it really bothers me)":

כל HIT הכיל 100 משפטים לדירוג וכלל התוצאות נורמלו לכדי z-score, כך שתשובותיו של כל אדם (MTurk worker) היו בעלות ממוצע 0 וסטיית תקן 1 (הן אם הוא ענה על HIT בודד של 100 משפטים, והן אם על יותר מאחד). נבחרו רק עובדים מארה"ב, בעלי HIT Approval Rate > 95%, והם קיבלו שכר של 0.5$[2].

כל HIT הכיל 100 משפטים, שנבחרו וסודרו אקראית, והכילו (שוב, בהשראת העבודה של GRAHAM):

1. 70 משפטים ייחודיים בעלי טעויות - כלומר משפטים שהופיעו בHIT זה בלבד
2. 15 משפטי בקרה – מתוך מאגר של 200 משפטים, כך שניתן היה להשוות בין תשובות המשיבים השונים.
3. 15 משפטים "מושלמים" - משפטים שהופיעו במאגר NUCLE אך לא תויגו כבעלי טעויות (אלו לא חזרו על עצמם).

---

[1] בחרנו להיצמד לניסוח בו עשו שימוש בעבודה קודמת בתחום, אינני בטוח שזהו ה-ניסוח האידיאלי לשאלה, ואני חושב שניתן לחשוב על ניסוחים שונים כתלות במהות הפרויקט. האם נדרשת דיפרנציאציה בין עד כמה טעויות מסוימות מפריעות לאנשים כאשר הם קוראים מייל מקולגה בעבודה לעומת חיבור בבית הספר או מאמר אקדמי? יתכן שאלו בעיות שונות הדורשות כלים שונים, ואז נדרשים פה ניסוחים שונים של השאלה עליה משיבים אנשים.

[2] חלקם התלוננו על כך שהעבודה רבה ולא שווה את השכר, אך הנתונים נאספו בכל זאת בתוך פחות מ48 שעות

מתוך המאגר נבחרו רק משפטים שמכילים לפחות 7 מילים[3] ולא מכילים את התווים או המחרוזות הבאות: "; ", "*", "[", "]", "&", "http". זאת משום שמשפטים שלא ענו לדרישות אלו הכילו רעש רב, בעיקר במשפטים ה"מושלמים" ומתוך רצון שלא לסנן באופן שונה את המשפטים ללא הטעויות המשמשים לבקרה ואת אלו המשמשים להערכה.

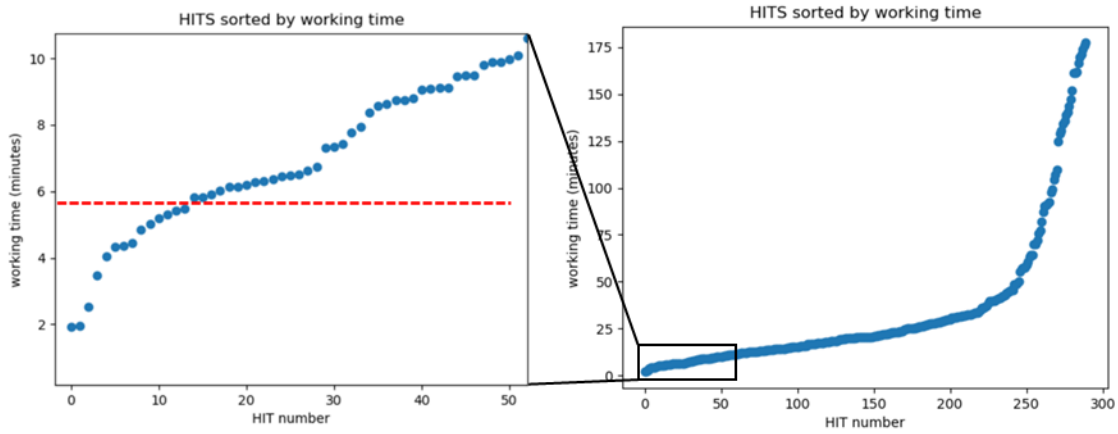
כמו כן, מכיוון שהתוכן בNUCLE מכיל לעיתים רווחים מיותרים בין סימני פיסוק למילה, המשפטים עברו את העריכה הבאה בהשוואה לקובץ המקור:

1. מחיקת התו רווח אם הוא הגיע לפי אחד מהתווים הבאים: $ % ! (/ . ,
2. מחיקת התו רווח אם הוא הגיע אחרי התו:)

<u>משתתפים וסינון:</u>

לאחר פיילוט קטן העלנו 290 HITים במבנה המפורט לעיל לNUCLE, אותם מילאו 245 אנשים שונים. אנשים אלו סוננו בעזרת 3 מדדים שונים:

1. משך הזמן למענה על כל 100 המשפטים:



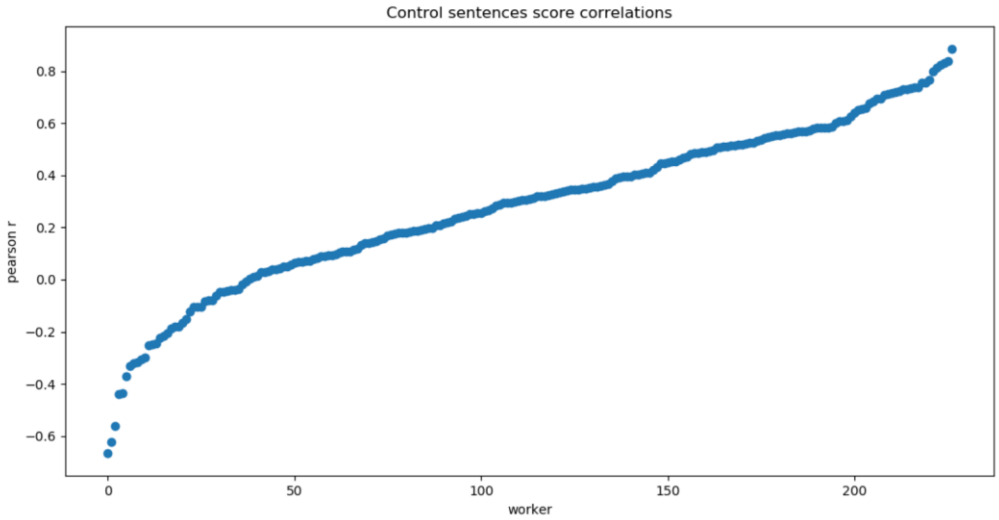
גרף 1: משך הזמן למענה על כל 100 המשפטים של כל אחד מ-245 המשיבים על הסקר. זמן מילוי ממוצע: 33.5 דקות, זמן מילוי חציוני: 20.4 דקות. בשמאל: הגדלה של הגרף עבור המשיבים שענו מהר במיוחד; אנשים שזמן המענה המינימאלי שלהם היה מתחת לקו האדום סוננו החוצה מן הנתונים.

בחרנו לסנן החוצה את כל מי שזמן המילוי המינימאלי שלו לHIT היה נמוך מ350 שניות (5.83 דקות) מתוך הבנה שאלו ככל הנראה רמאים שענו על הסקר במהירות ללא התייחסות לשאלה באופן אמיתי. 12 מתוך 245 אנשים סוננו החוצה באופן זה.

2. לאחר סינון זה התבצע סינון נוסף על סמך הציון שנתנו האנשים למשפטים ה"מושלמים". נערך t-test שבדק האם הציון שקיבלו המשפטים אשר מתויגים ללא טעויות, גבוה באופן מובהק ($\alpha = 0.05$) מאלו שמתויגים כבעלי טעויות. 6 אנשים נוספים סוננו באופן זה.

[3] כאשר ההגדרה של מילה פה היא כל דבר שהופיע בקורפוס בין שני רווחים, לרוב היו אלו ציטוטים קצרים או קטעי טקסט לא רלוונטים.

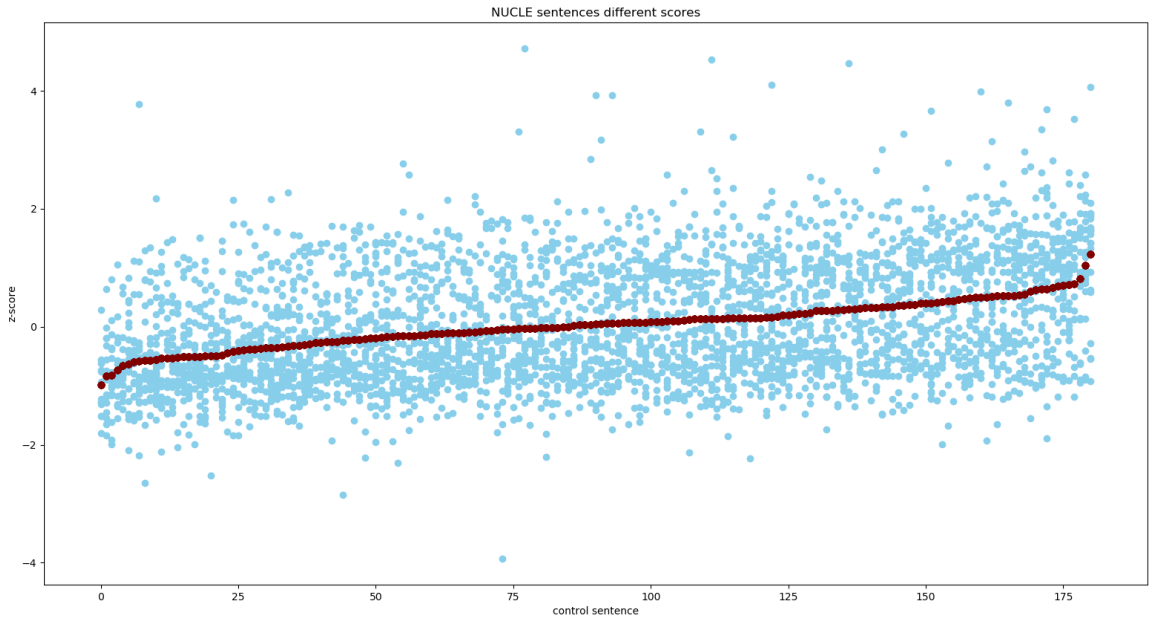
3. לאחר מכן, בדקנו את ציוני האנשים על משפטי הבקרה שהם ענו עליהם, בהשוואה לתשובות שאר המשיבים על אותם המשפטים. יודגש שמרבית האנשים שסוננו בשלב 1 ו-2 נמצאו בקורלציה שלילית עם שאר המשיבים, אך העלאת הרף (למשל סינון כל מי שענה בפחות מ-10 דקות) לא תרמה באופן מובהק לסינון אנשים שנמצאו בקורלציה שלילית, עובדה שנלקחה בחשבון בבחירת רף הסינון. לחישוב הקורלציה השתמשנו במדד פירסון r בין וקטור הציונים של אותו אדם על משפטי הבקרה, לממוצע של כל שאר המשיבים על אותם המשפטים, בניקוי האדם שבדקנו. בחישוב זה נלקחו בחשבון רק משפטי בקרה עליהם השיבו יותר מ-15 אנשים[4].



גרף 2 : קורלציה בין ציוני משיב ספציפי על משפטי הבקרה, לאלו של כלל המשיבים על אותם המשפטים ; תשובותיהם של אנשים בעלי קורולציה נמוכה מ 0.4- סוננו.

אנשים בעלי קורלציה שלילית קיצונית (מתחת ל 0.4-) סוננו גם הם ממאגר הנתונים, וכך נופו 5 משיבים נוספים. רף זה נבחר בהתאם להחלטה מראש לסנן 10% מהמשיבים.

[4] כי נראה שהחל מרמה של 15 משיבים שונים ניתן לדבר על הסכמה כלשהי בציון של משפט (graham et al, 2015)



גרף 3 : ציונים שונים על משפטי הבקרה- באדום הציון הממוצע של כל משפט, בכחול כלל הציונים. ניתן לראות כי ישנה שונות לא מבוטלת ברמת המשפט. ממוצע של כלל התשובות הינו 0.029.

בסה״כ, מתוך 245 משיבים על 290 HITים, נופו 23 משיבים שענו על 31 HITים. לאחר תהליך זה נותרו 25,900 משפטים עם ציון (מתוכם 200 משפטי בקרה שחזרו פעמים רבות ו15% משפטי בקרה ללא טעויות). כלל האנשים שנופו קיבלו את התשלום בכל זאת.

לאחר קבלת התוצאות נעשו מספר ניסיונות לשנות את רף הסינון במדדים השונים ולראות כיצד אלו משפיעים על התוצאות. בדיקות אלו הובילו לתוצאות דומות מאוד לתוצאות המוצגות (מלבד הגדלת הרווח בר הסמך, לאור הקטנת המדגם). כלומר, נראה שהנתונים מורעשים מאוד, אך לא נראה שהחמרת הרף במדדים הקיימים משנה את התוצאות באופן מהותי.

<u>תוצאות</u> :

לאחר סינון נתונים לא מהימנים, נותרנו עם מאגר של סה״כ 22,015 משפטים המתויגים כבעלי טעויות (רובם שונים זה מזה, אך ישנן גם חזרות על משפטי הבקרה בתוך ספירה זו). כל משפט כזה תורגם לווקטור של טעויות (z-score משפטים), ולכל וקטור טעויות מוצמד ציון ERRANT או NUCLEבהתאם לשיטת התיוג הרצויה . שחזרו על עצמם באיסוף המידע חזרו גם במאגר זה ומופיעים כווקטור טעויות זהה, בעלי ציונים שונים.

מאגרים אלו נראים כך, ונשמרו כחלק מהפרויקט (2 קבצים, אחד לכל מערכת תיוג שגיאות) :

Nucle_ID	Vt	Vm	V0	Vform	SVA	ArtOrDet	Nn	Npos	Pform	Pref	Prep	Wci	Wa	Wform	Wtone	Srun	Smod	Spar	Sfrag	Ssub	WOinc	WOadv	Trans	Mec	Rloc-	Cit	Others	Um	TotalMistakes	z-score
156790	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0.480722443
76902	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.537444856
42856	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	2	0.480722443
37424	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0.849418123
60918	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.480722443
123098	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1.135866304
23352	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1.135866304
144463	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	-1.135866304
100181	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.537444856
61478	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.565806062

על מאגר זה ביצענו רגרסיה לינארית שמטרתה לחזות את הציון שינתן למשפט, בהינתן וקטור טעויות מסוים. המטרה האמיתית של הרגרסיה הייתה חילוץ המשקלות השונות שינתנו לכל סוג טעות, כאשר משקולת גבוהה מעידה כי טעות זו מנבאת שהמשפט יהיה בעל ציון גבוה, קרי הטעויות בו מפריעות לקורא. הבחירה ברגרסיה לינארית לא נבעה מכך שאנחנו חושבים שיש קשר לינארי פשוט בין הגורמים אלא בעיקר מכיוון שזהו מודל פשוט המאפשר לחלץ את המשקלות בקלות. מאגרי המידע קיימים וניתן לבחון גם דרכי אחרות לחלץ מהם משקלות אלו או אחרות.

ערכנו את הרגרסיה גם בהוספת משתנה של סך הטעויות, על מנת לבחון האם נדרשת שליטה מסוימת על משתנה זה, שעשוי להשפיע על המשקולות השונות. התוצאות אינן שונות במקרה זה (פירוט בטבלה).

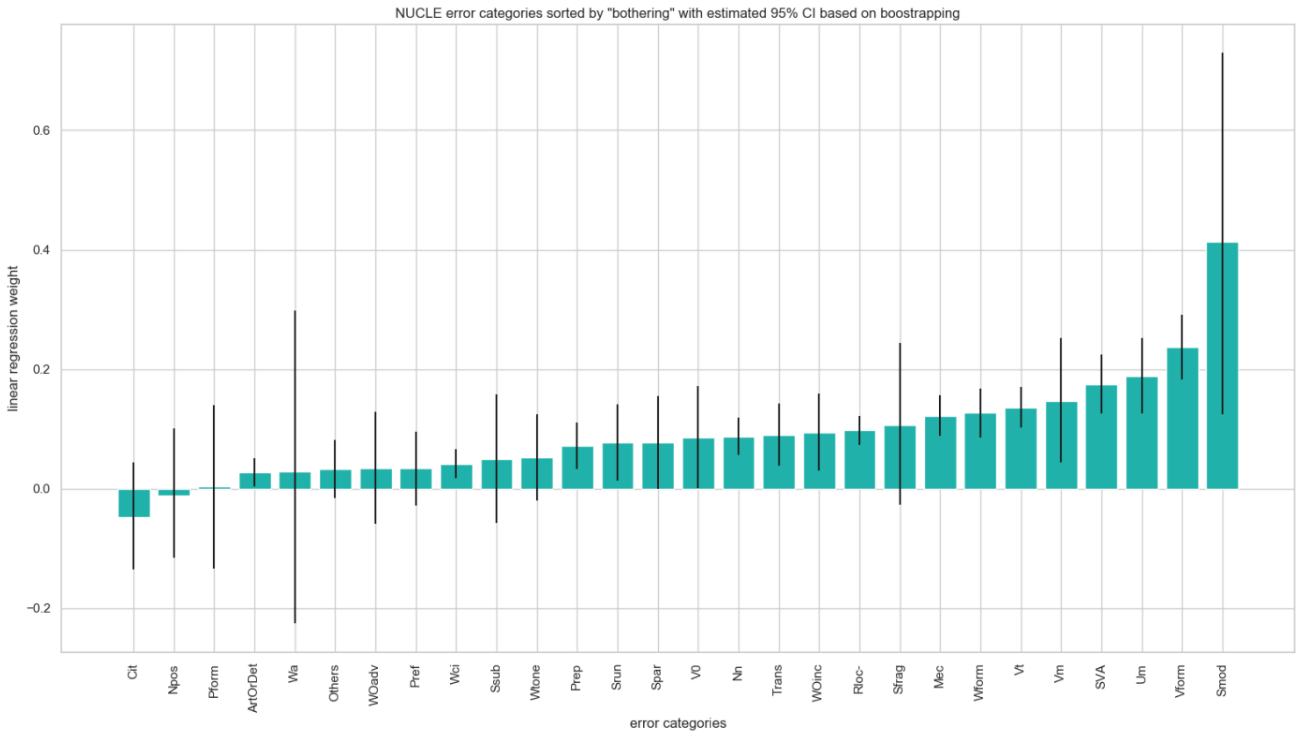
בנוסף לרגרסיה על המאגר כולו ביצענו Bootstrapping, כלומר דגימה מחדש עם חזרות של המשפטים השונים 10,000 פעמים ובחנו את הציון שכל סוג טעות מקבלת וכן את הדירוג (rank) של סוג הטעות ביחס לטעויות האחרות. הממצאים (הן על שיטת התיוג של NUCLE[5] והן על ERRANT[6]) מתוארים בטבלאות והגרפים הבאים (מאגר המידע המלא קיים ושמור במסגרת הפרויקט, כך שניתן לחלץ נתונים נוספים).

<u>טבלה 1 - ציוני הטעויות לפי NUCLE וכמות ההופעות שלהן במדגם:</u>

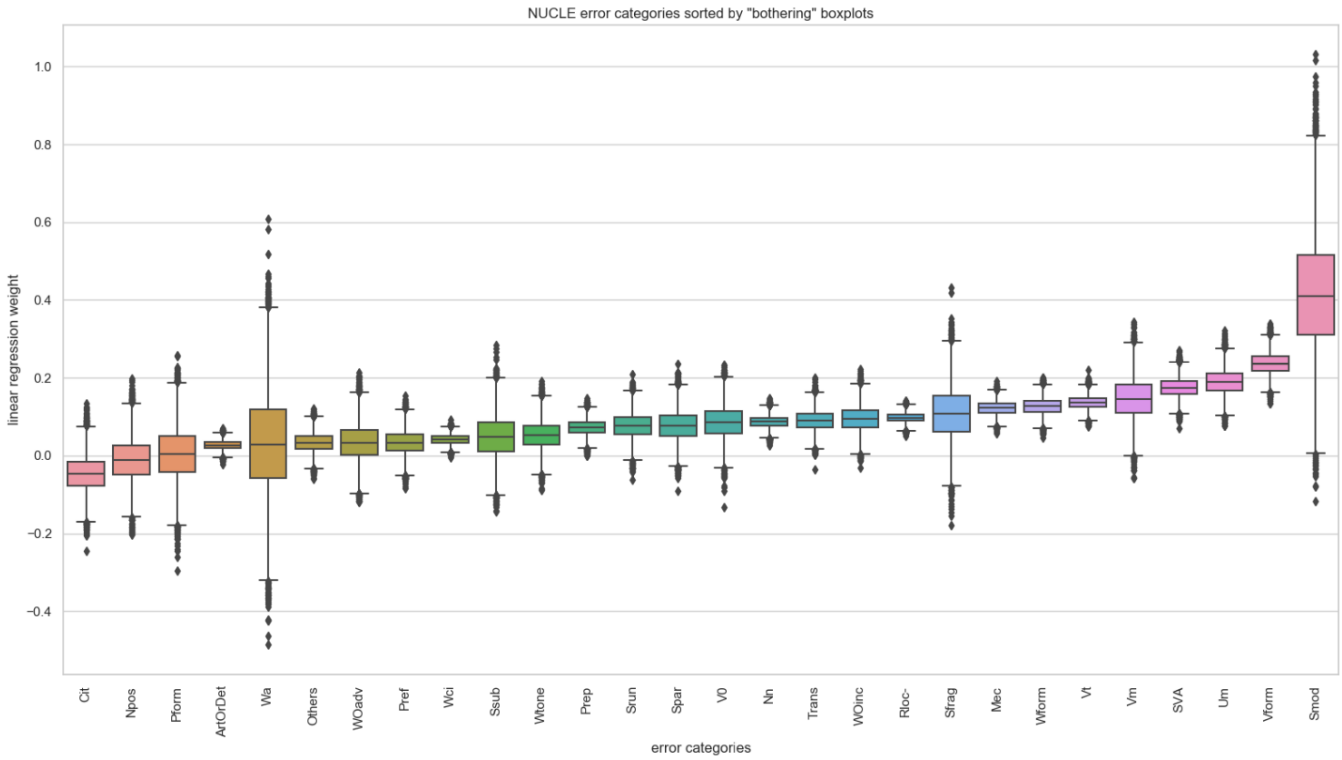
	Total appearances	% of appearances	weight	Regularized weight
Smod	58	0.13%	0.414	0.599
Vform	1423	3.19%	0.237	0.417
Um	1062	2.38%	0.189	0.373
SVA	1467	3.29%	0.174	0.353
Vm	413	0.93%	0.146	0.338
Vt	3258	7.31%	0.136	0.312
Wform	2262	5.08%	0.127	0.313
Mec	2899	6.51%	0.122	0.306
Sfrag	161	0.36%	0.107	0.294
Rloc-	5022	11.27%	0.098	0.282
WOinc	696	1.56%	0.094	0.279
Trans	1270	2.85%	0.090	0.276
Nn	3754	8.43%	0.087	0.275
V0	449	1.01%	0.086	0.277
Spar	499	1.12%	0.078	0.261

[5] https://www.aclweb.org/anthology/N15-1124,
[6] https://www.aclweb.org/anthology/C16-1079, https://github.com/chrisjbryant/errant

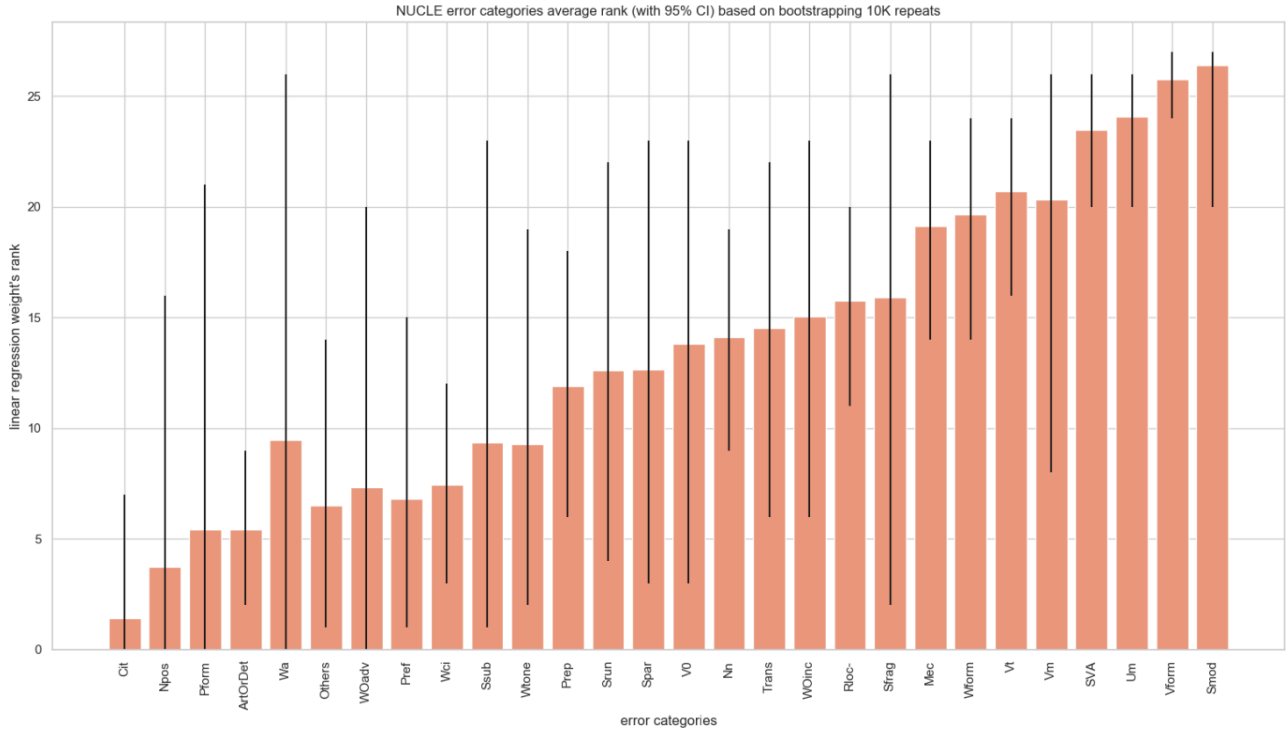
Srun	828	1.86%	0.077	0.264
Prep	2473	5.55%	0.072	0.260
Wtone	599	1.34%	0.052	0.237
Ssub	328	0.74%	0.049	0.239
Wci	5482	12.31%	0.042	0.227
Pref	929	2.09%	0.034	0.224
WOadv	316	0.71%	0.034	0.218
Others	1463	3.28%	0.034	0.219
Wa	58	0.13%	0.029	0.215
ArtOrDet	6563	14.73%	0.027	0.211
Pform	171	0.38%	0.004	0.186
Npos	248	0.56%	-0.011	0.174
Cit	394	0.88%	-0.048	0.139



גרף 4 – ציוני הרגרסיה, ורווח בר הסמך של הציונים בהתבסס על הסימולציות החוזרות. ציון המשקולת הממוצע: 0.0092.



גרף 5 – boxplots של הציונים והחזרות עליהם. כל תיבה מכילה 50% מהתוצאות.

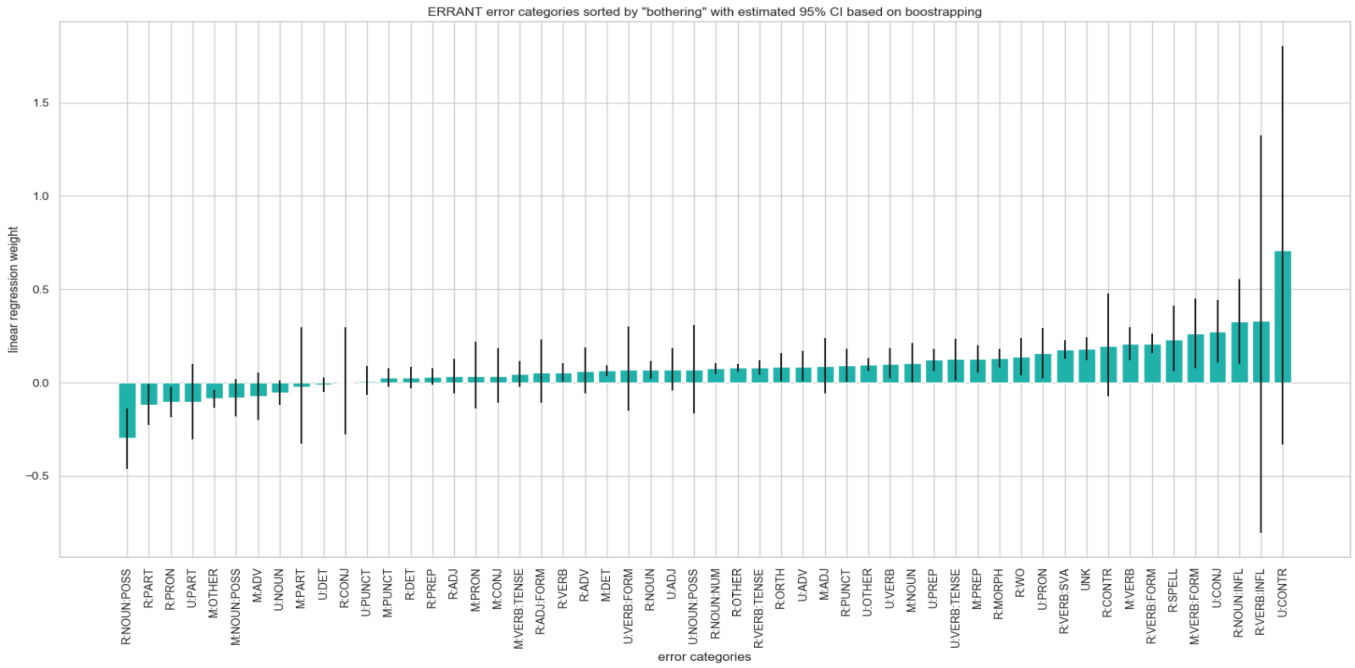


גרף 6 – הדירוג הממוצע של הטעויות השונות לאורך 10,000 דגימות מהמאגר. בכל איטרציה המשקולות דורגו לפי סדר הציונים בין 0 ל-27, ציון גבוה משמעותו שהטעות קיבלה ציון משקולת גבוה. בשחור מסומן הרווח בר הסמך של הדירוג (95% מהדירוגים במהלך ה-bootsrap נפלו בטווח זה) ; ניתן לראות שהשונות גדולה.

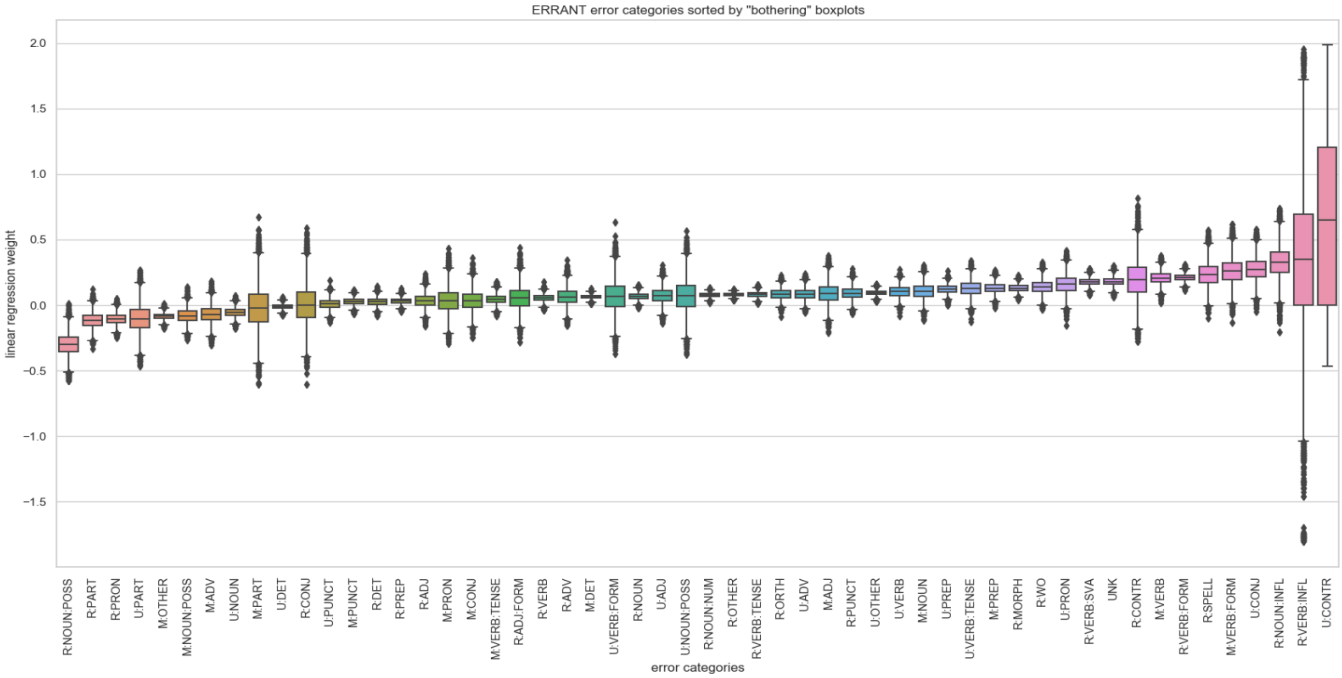
טבלה 2 - ציוני הטעויות לפי ERRANT וכמות ההופעות שלהן במדגם :

	total_appearance	perc_of_appearance	weights	regularized_weights
U:CONTR	2	0.00%	0.708	0.840
R:VERB:INFL	6	0.01%	0.330	0.469
R:NOUN:INFL	67	0.14%	0.327	0.464
U:CONJ	154	0.31%	0.274	0.413
M:VERB:FORM	107	0.22%	0.262	0.405
R:SPELL	148	0.30%	0.232	0.348
R:VERB:FORM	1712	3.49%	0.209	0.336
M:VERB	499	1.02%	0.207	0.347
R:CONTR	56	0.11%	0.195	0.336
UNK	1062	2.16%	0.181	0.316
R:VERB:SVA	1447	2.95%	0.179	0.305
U:PRON	211	0.43%	0.159	0.299
R:WO	373	0.76%	0.139	0.281
R:MORPH	1677	3.41%	0.130	0.267
M:PREP	773	1.57%	0.129	0.260
U:VERB:TENSE	316	0.64%	0.126	0.260
U:PREP	1020	2.08%	0.122	0.261
M:NOUN	377	0.77%	0.105	0.221
U:VERB	542	1.10%	0.102	0.232
U:OTHER	2964	6.04%	0.095	0.215
R:PUNCT	442	0.90%	0.092	0.217
M:ADJ	179	0.36%	0.091	0.224
U:ADV	604	1.23%	0.085	0.243
R:ORTH	585	1.19%	0.084	0.252
R:VERB:TENSE	2676	5.45%	0.082	0.229
R:OTHER	6308	12.84%	0.079	0.210
R:NOUN:NUM	4088	8.32%	0.076	0.203
U:NOUN:POSS	67	0.14%	0.071	0.193
U:ADJ	260	0.53%	0.071	0.208
R:NOUN	1396	2.84%	0.069	0.209
U:VERB:FORM	51	0.10%	0.068	0.195

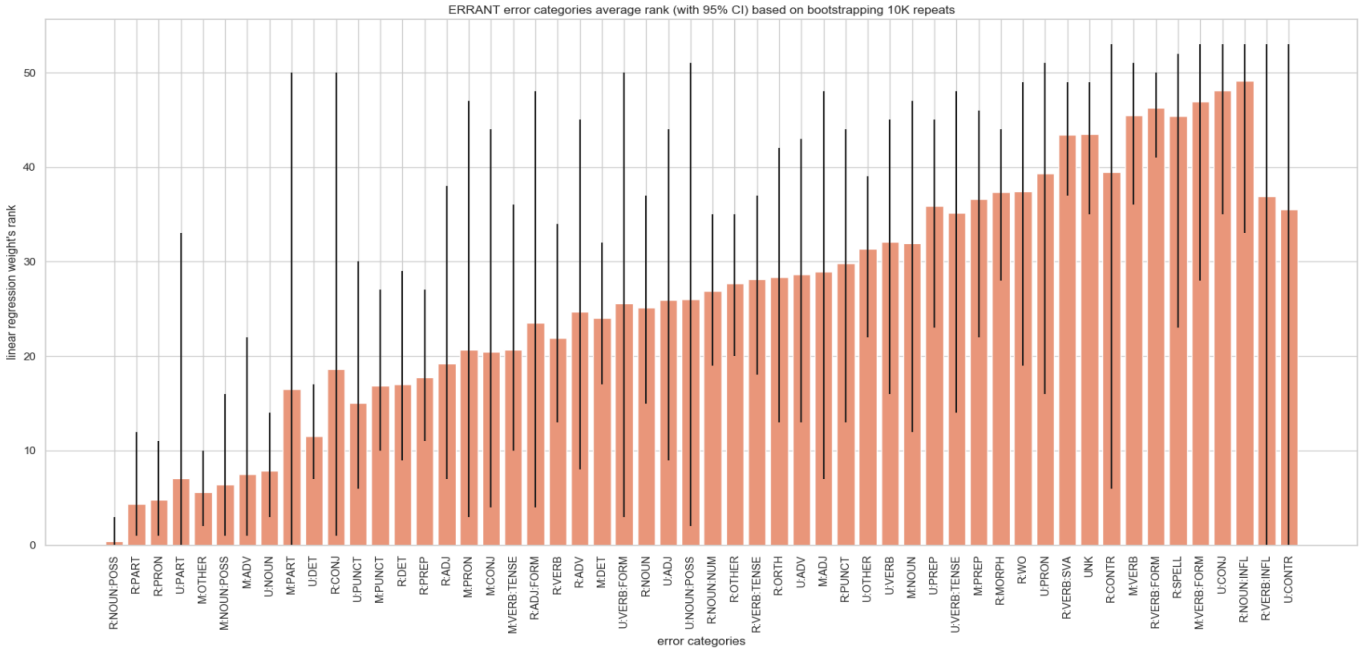
M:DET	4212	8.58%	0.065	0.224
R:ADV	282	0.57%	0.063	0.218
R:VERB	1515	3.08%	0.054	0.190
R:ADJ:FORM	153	0.31%	0.054	0.193
M:VERB:TENSE	598	1.22%	0.046	0.163
M:CONJ	166	0.34%	0.036	0.172
M:PRON	153	0.31%	0.034	0.163
R:ADJ	491	1.00%	0.034	0.168
R:PREP	2045	4.16%	0.032	0.153
R:DET	1120	2.28%	0.027	0.162
M:PUNCT	1363	2.78%	0.026	0.160
U:PUNCT	635	1.29%	0.009	0.143
R:CONJ	46	0.09%	0.002	0.137
U:DET	2570	5.23%	-0.011	0.114
M:PART	37	0.08%	-0.023	0.110
U:NOUN	727	1.48%	-0.054	0.090
M:ADV	206	0.42%	-0.072	0.062
M:NOUN:POSS	216	0.44%	-0.080	0.065
M:OTHER	1462	2.98%	-0.086	0.049
U:PART	87	0.18%	-0.102	0.035
R:PRON	486	0.99%	-0.103	0.026
R:PART	274	0.56%	-0.117	0.018
R:NOUN:POSS	100	0.20%	-0.296	-0.153



גרף 7 – ציוני הרגרסיה, ורווח בר הסמך של הציונים בהתבסס על הסימולציות החוזרות על תיוגי ERRANT. ציון המשקולת הממוצע הוא : 0.0068.



גרף 8– boxplots של הציונים והחזרות עליהם לפי ERRANT. כל תיבה מכילה 50% מהתוצאות.



גרף 9 – הדירוג הממוצע של הטעויות השונות לאורך 10,000 דגימות מהמאגר על תיוגי ERRANT. בכל איטרציה המשקולות דורגו לפי סדר הציונים בין 0 ל-27, ציון גבוה משמעותו שהטעות קיבלה ציון משקולת גבוה. בשחור מסומן הרווח בר הסמך של הדירוג (95% מהדירוגים במהלך ה-bootsrap נפלו בטווח זה); ניתן לראות שהשונות גדולה מאוד. לאור מספרם הרב של סוגי הטעויות, אין מספיק נתונים על חלקן.

דיון ומסקנות:

מתוצאות המחקר על שתי מערכות התיוג ניתן להצביע על מספר מסקנות:

1. מיידע, קרי determiner לא כל כך מפריע לאנשים - על אף שזו הטעות הכי נפוצה, היא מדשדשת בסוף הדירוג עם משקולת מאוד נמוכה בשתי שיטות התיוג.
2. באופן כללי נראה כי אין קשר בין כמה שהטעות הטעות נפוצה למידה בה היא מפריעה לאנשים (קורלציה[7] של 0.06- בין אחוזי ההופעות של הטעויות במאגר לציון המשקולות שלהן)
3. ככלל טעויות שקשורות בפעלים מפריעות לאנשים בשתי מערכות התיוג והן מקבלות ציונים גבוהים.
4. בתיוג האוטומטי של ERRANT, למעלה מ-12% מהטעויות סווגו כ"אחר" (other). טעויות אלו עדיין מקבלות ציון גבוה ומדורגות בתחילת השליש השני של הטעויות בדירוג לפי משקלות הרגרסיה. לכן יתכן ושווה להשקיע מאמץ נוסף בפירושן וביכולתן לתקנן.
5. עוד עולה מERRANT כי איות (spelling) הינו בעל חשיבות וטעויות מסוג זה מפריעות לאנשים.

[7] פירסון r

6. בקריאת הנתונים מומלץ לא לייחס חשיבות רבה במיוחד לטעיות נדירות במידה ניכרת שכן נראה שהן פחות מהימנות (למשל Smod ו-Wa לפי תיוג NUCLE, או U:CONTR ו- R:VERB:INFL לפי תיוג ERRANT).

<u>צעדי המשך אפשריים:</u>

1. הגדלת המהימנות:
   a. על מנת להגדיל את מהימנות התוצאות ניתן לאסוף עוד דאתא, כלומר מספר אנשים שמתייגים כל משפט. רמת ההסכמה של מתייגים שונים על משפט בודד מאוד נמוכה, בעבודה זו יצאנו מנקודת הנחה שברמה הכללית, כאשר מסתכלים על משפטים רבים ותיוגים רבים השונות הזו לא תפריע מכיוון שיש הרבה מידע, אך ניתן לאתגר אמירה זו; במיוחד עבור הטעויות שלא חזרו פעמים רבים בקורפוס כולו.
   b. אפשרות נוספת ופשוטה יותר על מנת להגדיל את המהימנות היא לאחד סוגי טעויות לקטגוריות רחבות יותר המכילות מספר טעויות שונות כל אחד.
2. הרחבת המחקר:
   a. כיוון המשך מעניין נוסף יכלול ניסוחים שונים לשאלה אשר נשאלים האנשים המבצעים את ההערכה בMTurk, כתלות במהות הפרויקט. במסגרת זאת, מעניין יהיה לבחון האם נדרשת דיפרנציאציה בין הציון אשר מקבלות טעויות מסוימות בהקשר של מייל מקולגה בעבודה לעומת חיבור בבית הספר או מאמר אקדמי. יתכן שאלו בעיות GEC שונות הדורשות כלים שונים, ואז נדרשים פה ניסוחים שונים של השאלה עליה משיבים אנשים (ויתכן שההבדל לא גדול).
   b. ניתן לבצע את הניתוח על קורפוסים אחרים (ושיטות תיוג טעויות שונות).
3. שימוש בתוצאות המחקר בGEC:
   a. ניתן להשתמש במשקולות אלו בעת לימוד מערכת תיקון שגיאות, למשל על ידי מתן משקל לטעויות באלגוריתם הלומד או בעזרת דגימה חוזרת של משפטים בעלי בעיות המפריעות יותר לאנשים.
   b. כמו כן, ניתן להשתמש במשקולות לטובת ביצוע הערכה (evaluation) של מדדים אוטומטים ומערכות תיקון שגיאות אוטומטיות, לדוגמא על ידי מתן משקלות לשיטות נוכחיות שמתבססות על עריכות של טעויות לפי סוג הטעות. בתוך כך ניתן גם להרחיב את MAEGE כדי שמערכת זו תבחן את מדדי ההערכה ולא תניח שכל הטעויות הן באותה רמה.

אנצל את המקום הנותר על דף זה להודות ללשם חשן שסייע לי רבות לאורך כל הפרויקט; ולעמרי על ההזדמנות, האכוונה, האתגר והסיוע. תודה 😊

GIT: https://github.cs.huji.ac.il/ofirshifman/GEC_ME_PROJECT

Project files:

DA\results:

1. Batch_3727145_batch_results.csv – original Mturk output csv.
2. filtered_results_with_zscores.csv – the filtered results with z-scores.
3. controls_df.csv – control sentences data only.
4. sentences_mistakes_scores.csv – sentences as a vector of NUCLE mistakes with z-scores.
5. sentences_mistakes_scores_errant.csv – sentences as a vector of ERRANT mistakes with z-scores.
6. mistakes_weights.csv – more statistic information about NUCLE weights.
7. mistakes_weights_errant.csv - more statistic information about ERRANT weights.
8. bootstrap.csv – 10,000 iterations bootstrap results
9. bootstrap_errant.csv - 10,000 iterations bootstrap results on ERRANT
10. ranks.csv - 10,000 iterations bootstrap mistakes ranking
11. ranks_errant.csv - 10,000 iterations bootstrap ERRANT mistakes ranking
12. graphs – graphs folder.

NUCLE\my_NUCLE_parser:

13. my_parser.py – this file parse NUCLE corpus into several databases (regular, perfect and control sentences), according to different filters that serves to create the MTurk csv file.
14. batchCreator.py – python script that write hard-coded JS script for MTurk
15. results_processing.py – main results processing file, including data filtering and re-formatting.
16. results_analysis.py – main results analysis file, create different data sets, and plot the results (imported to results_processing.py and being used by it).

NUCLE\to_Mturk:

17. c_sentences.csv, c_sentences.txt – project control sentences.
18. m_sentences.csv, m_sentences.txt – project mistake sentences – sentence that has been evaluated by one worker only.
19. p_sentences.csv, p_sentences.txt - project perfect sentences – sentences without mistakes.
20. mTurk_csv.csv – final csv to be uploaded to MTurk.

מקורות:

1. GRAHAM, Y., BALDWIN, T., MOFFAT, A., & ZOBEL, J. (2017). Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering, 23(1), 3-30. doi: 10.1017/S1351324915000339
2. Graham, Y., Baldwin, T., & Mathur, N. (2015). Accurate Evaluation of Segment-level Machine Translation Metrics. HLT-NAACL.
3. Dahlmeier, Daniel et al. "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English." *BEA@NAACL-HLT* (2013).
4. Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of Error Types for Grammatical Error Correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada.
5. Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in esl sentences using linguistically enhanced alignments. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan.