עיבוד שפות טבעיות - תרגיל 4

סמסטר א' תשפ"ה

מגישים

שם	ת.ז
אור ישלח	314861584
עודד דואק	039945522

<u>חלק א':</u> יצירה של מודל Vec2Word ואימונו על קורפוס הכנסת

סעיף ב'

בחירת ערכים

- המשמעות של גודל הוקטור (**vector_size**) היא מספר המימדים בווקטור ש-Word2Vec מייצר עבור כל מילה.

- הקורפוס שלנו קטן יחסית ולכן לא נרצה וקטור גדול מדי שעלול להכניס רעשים ולגרום ל-Overfitting.
חיסרון נוסף בווקטור כזה הוא צריכת זיכרון גדולה יותר וזמן אימון ארוך יותר.
- לא נרצה גם וקטור קטן מדי כי הוא עלול לגרום לייצוג פשוט מדי של המילים,
דבר שיגרום להבנה מועטה על המשמעות וההקשר שלהן.

ניסינו וקטור בכמה גדלים (50,100,150,200) **ובחרנו לבסוף בוקטור באורך 100** מפני שקיבלנו בעזרתו את התוצאות הכי טובות.

- המשמעות של **window** היא המרחק המקסימלי בין מילת המטרה לבין המילים הסובבות אותה שבעזרתן בונים את וקטור הייצוג.

<u>לדוגמא:</u> אם למשל window=4 מסתכלים על 4 המילים מכל צד של מילת המטרה.

- חלון קטן יותר גורם ללמידה מקומית יותר סביב מילת המטרה (לעיתים מידע תחבירי או פונקציונלי).
- חלון גדול יותר גורם להבנה של קונטקסט רחב יותר הקשור לנושא של המילה.

לאחר ניסינות (5-8) החלטנו **לבחור בגודל חלון של 5** (ערך ברירת המחדל הוא 5).

- המשמעות של **min_count** היא התעלמות מכל המילים המופיעות בקורפוס מתחת לערך זה.

<u>לדוגמא:</u> אם min_count=2 אז יוצרים את וקטורי המילים רק בעזרת מילים המופיעות לפחות פעמיים בקורפוס.

- ערך נמוך עלול לגרום להתחשבות במילים המופיעות מעט.
מילים אלו יכולות לכלול טעויות כתיב, מורפולוגיה לא שכיחה או סתם להיות מילים שהשימוש בהן מועט.
דבר זה עלול לגרום להכנסת רעש, לצריכת זיכרון גדולה יותר ולזמן אימון ארוך יותר.
- ערך גבוה מדי עלול להסיר מילים שעוזרות בהבנת משמעות והקשר.

הקורפוס שלנו קטן יחסית ולכן יש הרבה מילים שמופיעות מעט שלא נרצה להתעלם מהן.
לאחר ניסיונות (2-8) **החלטנו לבחור בערך 5** .

שאלות

1. הסבירו מה המשמעות, ומה היתרונות והחסרונות של הגדלת והקטנת גודל הוקטור- vector_size.

המשמעות של גודל הוקטור (**vector_size**) היא מספר המימדים בווקטור ש-Word2Vec מייצר עבור כל מילה.

- כאשר גודל הוקטור **קטן** אז הייצוג של כל מילה קצר יחסית ולכן אינו כולל הרבה מידע.
- כאשר גודל הוקטור **גדול** יחסית אז המידע על משמעות כל מילה וההקשר שלה יכול להיות עשיר יותר.

	יתרונות	חסרונות
הגדלה של גודל הווקטור	• מאפשר לקודד יותר מידע, מה שמסייע לייצוג עשיר ומדויק יותר של משמעות המילה ושל ההקשר שבו היא מופיעה.	• שימוש בווקטור מאוד גדול יכול לגרום ל-Overfitting ולהכנסת רעש ל-embedding (בעיקר כאשר הקורפוס קטן יחסית) • צריכת זיכרון גדולה יותר וזמן אימון ארוך יותר
הקטנה של גודל הווקטור	• צריכת זיכרון קטנה יותר • זמן אימון קצר יותר • פחות סיכוי ל-Overfitting - מה שמאפשר הכללה טובה יותר	• ייצוג פשוט של כל מילה, שמוביל להבנה מוגבלת של המשמעות שלה ושל ההקשר שבו היא מופיעה. • עלול לגרום לחוסר הבחנה בין מילים דומות אך בכל זאת שונות.

2. הסבירו מה הבעיות שיכולות לעלות משימוש במודל הנ"ל , שאומן על הקורפוס שלנו. התייחסו בתשובתכם לאופן שבו יצרנו את הקורפוס, לגודל שלו ולשימושים פוטנציאלים של המודל.

- בעיות שיכולות לעלות משימוש במודל Word2Vec שאומן על הקורפוס שלנו:

1. הקורפוס שלנו קטן יחסית מה שעלול לגרום ל-Overfitting ובעיות בהכללה.

2. ייתכנו מקרים רבים שבהן מילים מופיעות מעט בקורפוס ולכן הווקטור המייצג אותן נבנה ממעט דוגמאות ולכן לא ייצג טוב את משמעותן.

3. הקורפוס שלנו נבנה מפרוטוקולים של וועדות ומליאות של הכנסת ולכן עוסק בנושאים ספיציפיים. המודל ילמד הקשרים שרלוונטיים בעיקר לנושאים אלו ולא ילמד ייצוגים עשירים ומגוונים. במילים אחרות, המודל לא ייצג מספיק טוב את השפה ובפרט, המודל יתקשה להכליל לתחומים אחרים שאינן שיח פוליטי.

- השימוש הפוטנציאלי של המודל מתאים בעיקר לטקסטים הקשורים לכנסת/חוקים/דיונים, פחות מומלץ לשפה עברית כללית.

חלק ב': דמיון בין מילים

סעיף ד'

הרצנו את הפונקציה most_similar עם הפרמטרים positive,negative ו-topn=3. לכן קיבלנו 3 אפשרויות החלפה לכל מילה.
בהתחלה ניסינו פרמטרים שונים ל-Word2Vec כך שיתקבלו תוצאות אופטימליות לדמיון מילים ולהחלפת המילים האדומות מבלי ששמנו ערכים ל-positive ול-negative (מלבד מילת המטרה האדומה). על מנת שנוכל לבצע השוואות רצינו לקבל תוצאות שאין בהן אקראיות. לשם כך, טענו את המודל במידה והוא קיים במיקום שקיבלנו בתור פרמטר.
לאחר ניסיונות רבים עם פרמטרים שונים ל-Word2Vec בחרנו בערכים vector_size=100,window=5, min_count=5.

שמנו לב שעדיין יש לא מעט מילים אדומות שלא הוחלפו במילים טובות מספיק. ניסינו להוסיף ל-positive ול-negative מילים שישפרו את ההחלפה. חשבנו על מילים כאלו ובדקנו את כמות המופעים שלהן בקורפוס. אם מילה הופיעה הרבה אז יתכן ו-Word2Vec למד טוב את המשמעות שלה ויצליח למצוא מילה דומה.
ניסינו להוסיף ל-positive מילים נרדפות או דומות למשל הוספה של המילה מרוצים ל-positive של המילה שמחים והוספה של המילה המליאה ל-positive של המילה הוועדה (כך קיבלנו את המילה קואליציה).
בנוסף, ניסינו גם לתקן את 3 הפלטים לכל מילה על ידי הוספת מילים ל-negative. למשל, אם ראינו שמתקבלות תוצאות ביחיד במקום ברבים אז שמנו בנוסף למילה ברבים ב-positive גם את המילה ביחיד ב-negative כדי לעזור למודל להבין שהוא צריך למצוא מילים ברבים.

כדי למצוא מילה טובה שתחליף את המילה בוקר ניסינו להוסיף ל-positive מילים כמו ערב או לילה. כאשר הוספנו את המילה ערב קיבלנו את המילה לילה בתור אפשרות להחלפה.
ניסיון נוסף שעשינו הוא לקיחת תוצאה סבירה מתוך 3 האפשרויות בפלט והוספה שלה לחלק של ה-positive.
כך למשל במהלך הניסיונות למציאת מילת החלפה למילה דקות ניסינו לשים ב-positive מילים כמו שניות, שעות, רגעים. ראינו בפלט את המילה ספורות ושמנו אותה ב-positive של המילה דקות וקיבלנו את המילה שניות בתור מילת החלפה. במהלך דומה מצאנו את המילה מפעל בתור תחליף למילה קידום והוספנו אותה ל-positive.

ניסינו הוספה של מילים בודדות בחלק של ה-positive וניסינו גם צירופים שלהן. כך למשל השתמשנו במילים מרוצים ורגועים ביחד ב-positive של המילה שמחים (וכך קיבלנו את המילה מצפים). דוגמא נוספת היא הוספת המילים מפעל ותפקיד ל-positive של המילה קידום (וכך קיבלנו את המילה תקן).
היה חשוב לנו לקבל משפט בעל משמעות הגיונית ולכן גם אם קיבלנו אפשרויות מתאימות מבחינה תחבירית המשכנו לחפש תוצאות נוספות. למשל, היתה לנו אפשרות להחליף את המילה היקר במילה המתועב אך משום שזה לא התאים להגיון אז המשכנו לנסות עד שמצאנו את המילה המחליפה שהתאימה למשפט.

לבסוף הצלחנו למצוא לכל מילה לפחות תחליף טוב אחד מתוך שלושת האפשרויות ואז בחרנו לכל מילה באיזה תחליף נשתמש במשפט.

נסביר על הפונקציה שלנו `replace_words_with_similar` שמבצעת את ההחלפה למילים האדומות:

קלט:

- משפט מקור שעליו נבצע את ההחלפה.
- רשימה של טאפלים: כל טאפל מכיל:
  - מילה אדומה

- רשימת positive: מילים שרוצים לקרב אליהן.
- רשימה negative: מילים שרוצים להתרחק מהן.

- מודל Word2Vec
- topn: מספר המילים הדומות שיישקלו.

אתחול משתנים:

- אתחל רשימה ריקה שתכיל מידע על ההחלפות שבוצעו (מילה מוחלפת ומה הוחלף במקומה).

לולאה:

- עבור כל מילה מסומנת ברשימה:
  1. נבדוק אם המילה נמצאת במודל:
     - אם המילה נמצאת, נריץ את `most_similar` עם:
       - רשימת positive: המילה עצמה ורשימת ה-positive שהוגדרה.
       - רשימת negative: רשימת ה-negative שהוגדרה.
       - מספר התוצאות (topn).
     - אם המילה לא נמצאת, ננסה להריץ את `most_similar` רק עם רשימת ה-positive ו-negative (ללא המילה המסומנת באדום).

החלפה:

- מתוך רשימת המילים שחזרה מ- `most_similar`, נבחר את אחד המועמדים מתוך השלושה ונחליף את המילה המסומנת במילה המחליפה במשפט.
- נשמור ברשימה את המילה שהוחלפה ואת המילה המחליפה.
- נעבור למילה הבאה ברשימה.

פלט:

- נחזיר את המשפט החדש לאחר ההחלפות ואת רשימת ההחלפות שבוצעו.

דוגמת הרצה על המשפט:

משפט מקורי: אין מניעה להמשיך לעסוק בנושא.

המילה מניעה מוגדרת להחלפה:

- רשימה חיובית (positive): המילה "בעיה"
- רשימה שלילית (negative): ריקה (אין מילים להתרחק מהן)
- topn = 3

תהליך ההרצה:

- המילה מניעה נמצאת במודל:
  נשתמש בה כבסיס ב-`positive` (יחד עם המילים ברשימת ה-positive)
- הפרמטרים שנשלחו ל-`most_similar`:
  - `positive`: ["מניעה","בעיה"]
  - `negative`: []
  - `topn=3` (שלושת המילים הכי דומות)

פלט `most_similar`:

- כוונה
- התנגדות
- הגדרה

- נבחר את המילה השנייה ברשימת התוצאות - התנגדות, ונחליף במשפט עם המילה המקורית
- נוסיף אותה לרשימת ההחלפות:
  - (מניעה, התנגדות)

פלט:

- משפט חדש: אין התנגדות להמשיך לעסוק בנושא.
- דאטה של ההחלפות:
  - (מניעה, התנגדות)

שאלות

1. האם המילים הכי קרובות שקיבלתם בסעיף א' תואמות את הציפיות שלכם? הסבירו.
   גם אם תאמו לציפיות וגם אם לא, נסו להסביר מדוע זה עבד או לא עבד טוב.

המילים הכי קרובות שקיבלנו בסעיף א' תואמות את הציפיות שלנו.
בחלק מהמקרים נראה שקיבלנו מילים בעלות משמעות דומה, בחלק קיבלנו מילים שיכולות להחליף את המילה מבלי לפגוע בתחביר, בחלק קיבלנו אותן מילים אך עם מורפולוגיה שונה ובחלק קיבלנו מילים שאינן מתאימות.
צפינו זאת משום שידענו שאחרי אימון ובחירת פרמטרים למודל נקבל תוצאות טובות אך משום שהקורפוס קטן ועוסק בשיח פוליטי אז נקבל גם תוצאות פחות טובות.

נסתכל על כמה דוגמאות מהפלט שקיבלנו:

אבחנה	Most Similar Words	Original Word
קשרים פוליטיים והיסטוריים, משקף הקשרים תרבותיים ומדיניים בקורפוס.	(אבותינו, 0.6668) (ודבש, 0.6462) (זבת, 0.6031)	ישראל

	(פלסטין, 0.5667) (היהודים, 0.5666)	
זיהוי הקשר בין "גברת" לשמות משפחה/פרטיים, תוצאה מצופה מהשפה הפורמלית בקורפוס.	(גמליאל, 0.9469) (מאור, 0.9427) (טלי, 0.9392) (ענת, 0.9382) (נעמי, 0.9359)	גברת
הקשרים בין "מים" ו"חשמל" או "מבנים" מראים השפעה של הקורפוס (דיונים כלכליים), במקום המשמעות הישירה של "מים" כמשאב טבע.	(מבנים, 0.9296) (חשמל, 0.9256) (חיילים, 0.9247) (הלימוד, 0.9221) (רווח, 0.9215)	מים
"מותר" הוא אנטונימי ל"אסור", מה שמשקף קשר ניגודי. שאר המילים משקפות הקשרים נפוצים בקורפוס.	(שאסור, 0.9023) (מותר, 0.8686) (ואסור, 0.8135) (כדאי, 0.7986) (תנו, 0.7844)	אסור

המודל Word2Vec לומד על סמך הקשרים בקורפוס ולא על סמך הבנה סמנטית. המודל מושפע ממילים שמופיעות לצדן בקורפוס, ולא ממשמעותן הישירה.

2. אם ניקח שתי מילים שנחשבות להפכים (antonyms), למשל "אהבה" ו"שנאה", או "קל" ו"כבד".
האם היינו מצפים שהמרחק בין שני וקטורי המילים שלהן יהיה קצר או ארוך? הסבירו.

המרחק הקוסינוסי (Cosine Similarity) בין שני וקטורים מחושב כך:

$$Cosine\ Similarity = \frac{\overline{A}\cdot\overline{B}}{|\overline{A}|\cdot|\overline{B}|}$$

אם הזווית בין $\overline{A}$ ל- $\overline{B}$ קרובה ל-°180 (מנוגדים), המרחק הקוסינוסי יהיה שלילי.

באופן תיאורטי, מאחר והמשמעות של מילים שהן הפכים היא מנוגדת, סביר שהן יופיעו בהקשרים שונים או מנוגדים.
לכן, היינו מצפים שהמרחק בין וקטורי מילים הנחשבות ל-antonyms יהיה <u>ארוך</u>.
במילים אחרות, ה-Cosine Similarity ביניהן יהיה קרוב ל-0 או שלילי.

במודל Word2Vec, משמעות של מילה נקבעת לפי ההקשר שלה בטקסט והפכים יכולים להופיע באותו הקשר (למשל, "אהבה" ו-"שנאה" במשפטים המתארים רגשות, למשל: "היא אהבה/שנאה אותו מאוד").
הקרבה מבחינת הקונטקסט גורמת למודל לתת ייצוג של וקטורים קרובים.
לכן, פעמים רבות המרחק בין וקטורי מילים של מילים שנחשבות להפכים דווקא יהיה <u>קצר</u>.

3. מצאו שלושה זוגות של מילים שנחשבות להפכים (antonyms) הקיימות בקורפוס שלנו ובדקו את המרחק ביניהן.
האם הציפייה שלכם מסעיף 2 מתקיימת עבורן עם המודל שבניתם?

מילה 1	מילה 2	cosine similarity
קל	כבד	0.8
פותח	סוגר	0.9
בעד	נגד	0.87

הציפייה שהמרחקים של הוקטורים יהיו גדולים לא מתקיימת מהסיבה שהוסברה בסעיף הקודם (מילים שיכולות להופיע באותו קונטקסט). עם זאת, לעיתים מילים שנחשבות להפכים יקבלו וקטורים פחות קרובים, והסיבה יכולה להיות שהן מופיעות הרבה בקורפוס והמודל למד שיש הבדל משמעותי ביניהן או שיש להן משמעות נוספת שהקורפוס עושה בה שימוש.

4. האם המשפטים הכי קרובים בסעיף ג' תאמו לציפיות שלכם? הסבירו.
גם אם תאמו לציפיות וגם אם לא, נסו להסביר מדוע זה עבד או לא עבד טוב.

המשפטים הכי קרובים מסעיף ג תאמו לציפיות שלנו.
בחרנו משפטים שרובם קצרים המופיעים הרבה בקורפוס. קל למודל למצוא משפט ממש דומה, המכיל כמה מילים זהות והשינוי המרכזי הוא שם חבר הכנסת או שם אחר כלשהו.
שינוי זה נחשב קטן גם לפי חישוב המרחק הוקטורי וגם לפי המשמעות הסמנטית (מילים זהות שעיקר השינוי הוא שם).

נסתכל על כמה דוגמאות מהפלט שקיבלנו:

אבחנה	Most Similar Sentence	Original Sentence
התאמה גבוהה, והמודל עבד כמצופה.	אני מאוד מודה לך .	אני מודה לך מאוד .
התאמה גבוהה, והמודל עבד כמצופה.	אני מזמין את חבר הכנסת רחבעם זאבי .	אני מזמין את חבר הכנסת יורם לס .
המודל מזהה את המשמעות הזהה למרות היפוך סדר המילים.	אדוני היושב - ראש , אני רוצה לסכם .	אני רוצה לסכם , אדוני היושב - ראש .

משפטים אינם תואמים לחלוטין, אך הדמיון המבני ניכר.	רבותי חברי הכנסת , אני מתכבד לפתוח את ישיבת הכנסת , היום יום שלישי , י"ב בטבת התש"ס , 21 בדצמבר 1999 .	חברי הכנסת , אני פותח את ישיבת הכנסת , יום שני , כ"ו באייר התשנ"ג , 17 במאי 1993 .

חלק ג': סיווג

```
================ BINARY CLASSIFICATION EVALUATION =============
=== Sentence Embeddings Features ===
KNN Results:
              precision    recall  f1-score   support

       first      0.767     0.901     0.829      2282
      second      0.880     0.726     0.796      2282

    accuracy                          0.814      4564
   macro avg      0.824     0.814     0.812      4564
weighted avg      0.824     0.814     0.812      4564
```

עבור אותם פרמטרים ותנאים שהשתמשנו בהם בתרגיל 3 קיבלנו תוצאות קצת פחות טובות עבור וקטור המאפיינים שנוצר באמצעות Word2Vec (עם TF-IDF קיבלנו accuracy=0.849 ועם Word2Vec קיבלנו accuracy=0.814).

לדעתנו דבר זה קורה ממספר סיבות:

- **גודל הקורפוס**

הקורפוס שלנו קטן יחסית עבור אימון מודל Word2Vec מה שעלול לגרום ל-Overfitting ובעיות בהכללה. המודל לא לומד מספיק טוב את משמעות המילים כך שקשה לבצע סיווג טוב בעזרת וקטורי המילים.

- **בחירת ערכים אופטימליים עבור KNN**

דבר זה נעשה עבור שימוש ב-TF-IDF. ייתכן שתתקבל תוצאת סיווג טובה יותר ל- Word2Vec עבור בחירת ערכים שונים ל-KNN.

- **דמיון בוקטורי המילים של הדוברים**

מילים שונות המופיעות אצל דוברים שונים יכולות להתאים בקונטקסט ולכן לקבל וקטורים קרובים ולא לעזור בסיווג בעזרת Word2Vec. לעומת זאת, אם המשקלים נקבעים על סמך תדירות הופעתן אצל דובר (כמו ב-TF-IDF) הן יוכלו לעזור לסיווג.

- **Word2Vec בונה ייצוג על סמך משמעות והקשר ולא נועד למשימות סיווג**

דוברים שונים יכולים להשתמש במילים שונות לגמרי ולכן יתקבל סיווג טוב בעזרת TF-IDF אך מילים אלו יכולות להיות בעלות משמעות דומה מבחינת Word2Vec ולכן לקבל וקטורים קרובים שלא עוזרים בסיווג.

חלק ד': שימוש במודלי שפה גדולים

שאלות

1. האם קיבלתם משפטים הגיוניים? מבחינת התוכן, קוהרנטיות ומבחינה תחבירית.

ברוב המקרים קיבלנו משפטים הגיוניים.
לפעמים קיבלנו השלמות זהות למילים המקוריות ולפעמים גם כאשר ההשלמות לא היו זהות (מבחינת תוכן וקוהרנטיות) קיבלנו בד"כ מילים שמתאימות לקונטקסט.
מבחינה תחבירית, ברוב המקרים קיבלנו תוצאות טובות שלא פוגעות בתחביר המשפט.

דוגמא למשפט שהושלם בצורה מעולה:

- *מקורי:* אני רוצה לתת תמונה קצת יותר רחבה , [MASK] אני אענה גם [MASK] השאלה הזאת .
  פלט: אני רוצה לתת תמונה קצת יותר רחבה , ואז אני אענה גם על השאלה הזאת.
  ההשלמות הגיוניות מבחינת רצף המשפט וזהות למילים המקוריות.

2. האם קיבלתם השלמות קרובות למילים החסרות האמיתיות? פרטו.

ברוב המקרים החיזויים שהתקבלו בעזרת DictaBERT היו זהים או דומים סמנטית למילים המקוריות.

דוגמא להשלמה מדויקת:

- *מקור:* מיכאל שפיר , הנהלת ענף ציוד רפואי מלשכות המסחר , בבקשה .
  פלט: מיכאל שפיר , הנהלת ענף ציוד רפואי מלשכות המסחר , בבקשה .

דוגמא להשלמה חלקית:

- *מקור:* אתם עשיתם את זה בענק , ואנחנו כננס על גבי ענק מוסיפים עוד משהו .
  פלט: אתם עשיתם את זה בענק ,אתם כננס על גבי ענק מוסיפים עוד משהו .
  המילה אתם היא השלמה לא נכונה, אך היא כינוי גוף בדומה למילה המקורית (ואנחנו).
  המילה עוד היא השלמה מדוייקת.

היו מקרים מסוימים שבהם ההשלמות לא היו קרובות.
דוגמא להשלמה לא נכונה:

- *מקור:* אדוני הרפרנט , אולי תגלה לי מי חתום מתחת לאמיר לוי ?
  פלט: אדוני הרפרנט , אולי תגלה לי מי חתום על לאמיר לוי ?

3. השוו את התוצאות שקיבלתם עכשיו לאלו שקיבלתם בתרגיל בית 2. האם יש שיפור בתוצאות לדעתכם?

בתרגיל בית 2 השתמשנו בשיטות המבוססות על התדירות של המילה במשפט ולא התעמקנו בהקשר שלה. לכן, רוב ההשלמות שקיבלנו היו טוקנים נפוצים בקורפוס (כמו פסיקים). המודל לא ניבא בצורה נכונה את המילה עם ההקשר הנכון.

ב־DictaBERT ניכר שיפור משמעותי באיכות ההשלמה. ברוב המשפטים שקיבלנו ניתן לראות שהמודל משלים באופן הגיוני את המשפט, במקום מילה נפוצה מאוד וחסרת הקשר.

4. האם יש משפטים שעבורם המודל עבד פחות טוב? אם כן ואם לא, הסבירו מה לדעתכם הסיבה לכך.

עבור המשפטים שבדקנו לא ראינו מקרים שבהם DictaBERT היה פחות טוב.
המודל DictaBERT משתמש בלמידה עמוקה ובהבנה דינאמית של קונטקסט (לעומת סטטית כמו של Word2Vec) על ידי שימוש בשכבות attention ולכן מבין מילים בהקשרים שונים.
DictaBERT אומן על טקסט ענק בעברית וזה מאפשר לו להכיר מגוון רחב מאוד של מילים ומורפולוגיות ולהכליל טוב.

5. באילו מקרים לדעתכם יש למודל סיכוי גבוה יותר להכשל או להחזיר תוצאה פחות טובה? מדוע?

למודל יש סיכוי גבוה יותר להכשל במקרים הבאים:

- אם המילה שצריך להשלים היא מילה נדירה או ספציפית לתחום מסויים, המודל יציע מילה נפוצה יותר.
- אם המשפט קצר אז למודל יהיה קשה יותר להבין הקשר ולמצוא את המילים החסרות.
- אם הקונטקסט הוא דו משמעי, המודל יכול לבחור השלמות המתאימות מבחינה סמנטית אך הן עדיין לא יהיו נכונות.
- שימוש במורפולוגיה והתאמה של מין, יחיד/רבים במשפט יכול לגרום למודל לחזות מילה עם מורפולוגיה שגוייה.