How Often Are Errors in Natural Language Reasoning Due to Paraphrastic Variability?

הרעיון המרכזי

המאמר מנסה להבין - כאשר מודל NLI טועה, האם זו באמת בעיית reasoning , או פשוט רגישות לניסוח אחר של אותו משפט?
כלומר, אולי המודל "מבין" את הבעיה, אבל נופל רק בגלל שינוי wording קטן.

מדובר ב – continuation של ה – Anchor Paper :

- **anchor paper** - בדק את שבריריות המודל ע"י paraphrases.
- **המאמר הזה** - מנסה למדוד בצורה מתמטית, כמה מהטעויות של מודלי NLI נגרמות בגלל paraphrastic variability (שינויי ניסוח בעלי אותה משמעות).

התרומה המרכזית

החוקרים מציעים metric חדש בשם - **PC** - Paraphrastic Consistency
מדד חדש הבודק מה ההסתברות שמודל יישאר עקבי כאשר מציגים לו שתי גרסאות שונות של אותו רעיון.

הסבר פשוט ל – PC

אם המודל:

- מצליח בשני הניסוחים
  או
- נכשל בשני הניסוחים

אז המודל נחשב כעקבי, ואין מדובר בטעות של הנובעת מ – paraphrastic variability
אבל אם - ניסוח אחד מצליח וניסוח אחר נכשל – אזי יש paraphrastic inconsistency .

ההבדל בין PC לבין Relaxed Fooling Rate (וגם לבין Strict)

- **במאמר המקורי (ה – Anchor)** – השתמשו ב relaxed וב – Strict על paraphrases **שהציגו תוצאות <u>נכונות</u> עוד לפני ה – Paraphrase.**
  - משמע, המדדים שבהם השתמשו בחנו עד כמה מודלי ה NLI שבריריים כשמבצעים paraphrase
- **במאמר הנוכחי**, הציעו והשתמשו ב PC, הבודק **האם מודל ה – NLI נשאר עקבי להחלטה שלו,** גם לאחר ביצוע Paraphrase
  - משמע, המדד הזה בודק את אמינות המודלים (consistent). עד כמה אפשר לסמוך על המודל שישאר יציב תחת paraphrases. (אם המודל לא אמין והוא משתנה את

הדאטהסט שהם בנו

הם בנו dataset חדש בשם – **PARANLU**, הכולל:

- 7782 paraphrases שנכתבו ע"י בני אדם
- ועוד paraphrases אוטומטיים

המטרה - לבדוק robustness של מודלי reasoning .

מתודולוגיה

החוקרים מפרידים בין 2 סוגי כשלים :

- **Language Understanding Failure**
  - שגיאה בה המודל לא הבין את הניסוח.
- **Reasoning Failure**
  - שגיאה בה המודל הבין את המשפט, אבל reasoning עצמו שגוי.

וזו בעצם אחת התרומות הכי חשובות של המאמר.

תהליך יצירת ה - paraphrases

החוקרים הם השתמשו ב - Label-Preserving Paraphrases
כלומר – paraphrases שמשנים בהם את המילים, אבל שומרים על אותה המשמעות הלוגית (במטרה להגיע לאותו ה label של ההקשר הסמנטי).

ממצא מרכזי

כל המודלים שנבדקו עדיין הראו PC נמוך, כזה שניתן עוד לשפר, גם כאשר ה - accuracy שלהם היה גבוה מאוד.

כלומר - מודלים יכולים לקבל ציוני accuracy גבוהים, אבל עדיין להתנהג בצורה לא עקבית כאשר אותו reasoning problem מנוסח אחרת.

ממצא חשוב נוסף – Pretraining שיפר משמעותית את ה – consistency, אבל בדיקות מעמיקות הבחנו fine tuning נוספים לא תמיד שיפרו את ה consistency, ולעיתים אף פגעו בו.
מסקנה – במצבים כאלו המודל נהיה "מותאם יתר" למשימה, ופחות יציב סמנטית

ממצא מעניין מאוד - המודלים היו יותר robust מול paraphrases אוטומטיים,
ופחות robust מול paraphrases שבני אדם כתבו.
ככל הנראה כי paraphrases אנושיים:

- יותר מגוונים
- יותר טבעיים
- פחות צפויים