Review 138: G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment, 05.09.2023

https://github.com/nlpyang/geval/tree/main, https://arxiv.org/abs/2303.16634.pdf  
  
סקירה זו נכתבה על ידי עדן יבין

**Paper: https://arxiv.org/abs/2303.16634v3**

תהליך אבלואציה של מודל עוסק בהערכת הביצועים שלו. אבל איך נוכל להעריך טקסט שנוצר על ידי מודל גנרטיבי? מדדים שונים כגון BLEU, מאפשרים לבצע הערכה אך צורכים טקסט נוסף שנכתב על ידי אדם כרפרנס. בעיה, כי השגת טקסט שכזה הינו יקר, מה הפתרון אז?

המאמר נכתב על ידי חוקרים ממיקרוסופט ומציג שיטה הנקראת G-EVAL. השיטה בבסיסה משתמשת ב-GPT כדי להעריך את ביצועי מודל השפה שיצר טקסט כלשהוא בהינתן איזו שהיא משימה. היא מדמה תהליך של מילוי שאלון הערכה בכך שמאפשרת להגדיר קריטריונים שונים שחשובים לנו, החוקרים, להערכת איכות הטקסט. איך השיטה עובדת:

המשתמש יוצר prompt בו הוא מתאר מה המשימה שמודל אחר ביצע + הסבר על הקריטריונים להערכת הטקסט.

שימוש במודל שפה נוסף כדי לפרק את שיטת האבלואציה לכמה צעדים, למשל: ״תקרא את כל הטקסט ושים לב לכמות מילות הקישור״, ״תקרא כל משפט ושים לב לכמות המילים במשפט״ וכדומה. שלב זה מבוצע על ידי שימוש ב-Chain of Thought (CoT) כדי לייצר את השלבים השונים האלו המדמים טופס הערכה עם תתי סעיפים.

נותנים למודל השפה לתת את הציונים השונים לכל קריטריון שהגדרנו בשלב אחד לטקסט שנרצה להעריך. הציון צריך להיות מספר שלם משום שניסויים שנערכו מראים שהם מתקשים עם float.

מגדירים פונקציית ניקוד אשר משקללת גם את הניקוד יחד עם התפלגות הניקוד מהתפלגות הטוקנים של המודל. הסיבה היא למקרה ויהיה שוויון בין הניקוד של טקסטים שונים נוכל להתחשב בניקוד יחד עם שקלול ההסתברות לקבל את אותו ניקוד.

כדי להשיג את ההתפלגות של הטוקנים מ-GPT4 הם דגמו כל קריטריון 20 פעמים כדי לקבל את התפלגות הטוקנים והשתמשו בזה בפונקציית הניקוד (יכול להיות ש-20 לא מספיק לעניות דעתי). עם GPT3.5 ניתן ישירות להשיג את ההסתברות לכל טוקן אז אין צורך לבצע זאת עם מודל זה. את התוצאות הם השוו מול מדדים אחרים בתחום. התוצאות כולן הן מול הערכה שביצע אדם על הטקסט לפי הקריטריונים שהוגדרו מראש (אותם קריטריונים שהשתמש גם במודל). ניתן לראות שהשיטה של החוקרים הביאה לתוצאות טובות באופן ניכר ויותר קרובות להערכה שבוצעה על ידי אדם.