⚡️🚀המאמר היומי של מייק 22.06.24:⚡️🚀

GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer

המאמר הזה הוא שפצור קל של המאמר שסקרנו אתמול 21.06.24. המאמר מציע גישה לאימון והיסק של מודל לזיהוי NER המורכב משלבים הבאים:

מעברים כל קטגוריה שברצוננו לזהות דרך טוקנייזר - הקטגוריות מופרדות על ידי טוקן מיוחד הנקרא "ENT"

מעבירים דרך הטוקנייזר את כל הטוקנים של הטקסט. ד״א הטוקנים של הקטגוריות מופרדות מהטוקנים של טקסט על ידי טוקן מיוחד "SEP"

מכניסים את הטוקנים מהשלבים הקודמים לטרנספומר דו-כיווני (encoder) כמו BERT או ROBERTA

מעבירים את הייצוגים תלויי הקשר של הקטגוריות דרך FFN דו שכבתי (יש כזה בטרנספורמר) כדי לקבל ייצוג של כל קטגוריה.

מפעילים את מה שנקרא במאמר הקודם: Structured Span Prediction כלומר כדי לזהות את הקטגוריה של הטוקנים i עד i+n: לוקחים את הייצוג של טוקן ה-i ואת זה של טוקן i+n ומעבירים את השרשור שלהם דרך FFN דו שכבתי (מבנה דומה לסעיף הקודם) וכך מפיקים ייצוגו של ה-span הזה

כדי לשערך הסתברות ש- span (תת-סדרה של טוקנים רצופים) שייך לקטגוריה j מחשבים סיגמואיד של המכפלה פנימית של ייצוג הקטגוריה j מסעיף 4 עם ייצוג ה-span מהסעיף הקודם.

מפעילים אלגוריתמיםן גרידים כדי לזהות spans השייכים לכל קטגוריה (המאמר לא מרחיב על כך, צריך להביט בקוד)

https://arxiv.org/abs/2311.08526