# עיבוד שפה טבעית מתקדם – תרגיל 1

ניב אקהאוס

## שאלות פתוחות

1. להלן שלושה QA datasets שמשתמשים ב-QA לאנוטציה של קונספטים אינטרינזיים:
   a. הדאטאסט Quoref; הוא מודד את התכונה האינטריזית של Coreference בכך שהוא מאפשר לשאול שאלות שהתשובה עליהן מאלצת את המודל לקשר ישויות ואיזכורים לישויות לאותו הרפרנס. מקור: הרצאה 1.
   b. הדאטאסט Chinese Electronic Health Records (CEHR); הוא מודד את התכונה האינטריזית של Named Entity Recognition בכך שהוא מאפשר לחלץ Named Entities באמצעות שאלת שאלות ישירה על הדאטא הרפואי. מקור: https://www.semanticscholar.org/paper/Nested-Named-Entity-Recognition-for-Chinese-Health-Chiang-Lin/9b36588b26d0cec2590a27bd7a71a0c1008e080c
   c. הדאטאסט SQuAD; הוא מאפשר למדוד את התכונה האינטריזית של Named Entity Recognition ע"י פרומפטינג של שאלות שמצפות לקבל זיהויי NER כתשובה, כמו שנעשה במחקר: https://arxiv.org/abs/2203.01543
2. עבור שני הסוגים של המשימה:
   a. עבור Interactive summarization:
      i. הגדרת המשימה: יצירת סיכום לטקסט באינטראקציה ושיתוף פעולה עם המשתמש האנושי, כלומר באמצעות משוב לתוצרים שמודל השפה מייצר, עד לתוצאת הסיכום הרצויה.
      ii. בנצ'מארק ראוי לציון: The Evaluation Framework for IntSumm כפי שהוצג במאמר: https://www.semanticscholar.org/paper/Extending-Multi-Document-Summarization-Evaluation-Shapira-Pasunuru/a7dab9725ca411b49ebcc09544119cd463a9b94c לא מדובר בדאטאסט אלא בבנצ'מארק לאבליואציה, שכן בשביל להעריך את הביצועים של Interactive Summarization אי אפשר פשוט להשתמש בדאטאסט, אלא בפידבק אנושי בזמן אמת.
      iii. אתגרים מובנים: האתגר המרכזי הוא ההתאמה של מודל המסכם לטווח רחב של אנשים פוטנציאליים שיכולים לתת לו משוב על הסיכומים שהוא פולט. אנשים שונים עשויים לתת משוב באופן שונה, בניסוחים שונים, או עם מידה שונה של שביעות רצון שמתבטא באותו משוב טקסטואלי. כמו כן אנשים שונים עשויים להיות מרוצים במידה שונה מאותו הסיכום או לרצות פלט סופי בסגנון שונה, מה שיקשה על מודל ללמוד להכליל.
   b. עבור Multi-document summarization:
      i. הגדרת המשימה: יצירת תמצית טקסטואלית אחת של המידע החשוב המופיע במספר רב של מסמכים, כך שהיא תסכם אותם יחד.
      ii. בנצ'מארק ראוי לציון: Multi-News כפי שהוצג במאמר: https://aclanthology.org/P19-1102.pdf וזמין ב-HuggingFace תחת השם multi_news. יש בו כ-56K דגימות, כל אחת מורכבת מסיכום שנכתב ע"י עורכי חדשות מקצועיים למספר כתבות חדשותיות באותו נושא ממקורות חדשותיים שונים.
      iii. אתגרים מובנים: כאשר המשימה היא לסכם מספר רב של מסמכים לידי תמצית אחת, יש קושי גדול יותר להוציא את המשותף בין המסמכים, במקום להתייחס אליהם כמו מסמך אחד

ארוך. לכן האתגרים המרכזיים הם למנוע חזרתיות ויתירות של הפרטים שמוזכרים בסיכום והאבליואציה – אם נשתמש באבליואציה בעזרת מטריקה כמו ROUGE למשל (המטריקה הנפוצה למטלות summarization) אז היא עשויה לחטוא עוד יותר מבסיכום מסמך בודד, מאחר שעלייה במספר המסמכים מעלה גם את מספר הדרכים לתמצת באופן שונה מרפרנס קיים. כמו כן כמות מסמכים גדולה יותר גורמת לקלט למודל להיות גדול יותר, מה שעשוי להקשות יותר מבחינת כמות הטוקנים האפשריים כקלט למודל.

3. תכונת המקביליות של טרנספורמרס רלוונטית גם באימון וגם בהסקה.
   a. בשלב האימון, מודלי RNN מעבדים כל את ה-embeddings ואת ה-hidden state של כל טוקן ביחד עם ה-hidden states הקודמים. כלומר פעולת העיבוד של כל טוקן תלויה בעיבוד של כל הטוקנים שלפניו, ולכן לא ניתן למקבל את התהליך. לעומת זאת בטרנספורמרים מנגנון ה-self-attention מאפשר לעבד את ה-attention של כל טוקן ביחס לטוקנים האחרים, באופן שהוא בלתי תלוי ב-attention של טוקנים אחרים לכל השאר. לכן את התהליך הזה אפשר לבצע באופן מקבילי.
   b. גם בשלב ההסקה, התלות של מודלי RNN בעיבוד הטוקנים הקודמים לא מאפשר לעבד טוקן כלשהו בלי שקודם עובדו כל הקודמים, לכן לא ניתן לבצע תהליך זה באופן מקבילי. בטרנספורמרים לעומת זאת, גם בזמן ההסקה בדומה לשלב האימון, החישוב של ה-attention של טוקן כלשהו ביחס לטוקנים האחרים הוא ב"ת בחישוב ה-attention של אותם טוקנים אחרים, לכן ניתן למקבל את התהליך.
4. עבור הבעיות המתוארות:
   a. בהתחשב באילוצי המשאבים, אין ספק שלא נרצה לבצע tuning למודל כ"כ גדול כמו T5 XXL עם B11 פרמטרים. מאחר שהמודל InstructGPT הוא מודל סגור, שימוש בו וביצוע in-context learning איתו יאלצו רכישת רשיון, ובהתחשב במשאבים המוגבלים שלנו, גם האופציה הזו לא ריאלית. לכן האופציה הסבירה ביותר היא לבצע fine-tuning ל-ELECTRA-base, שכמו שניתן לראות בחלק התכנותי של התרגיל, מגיע לביצועים לא רעים בכלל.
   b. בשביל לפתור את הבעיה המתוארת בעזרת BERT אוכל לשרשר את הטוקנים של שני המשפטים באמצעות טוקן מפריד (כמו <sep>) ולהוסיף בהתחלה טוקן מיוחד (כמו <cls>), ואז אוכל לבצע fine-tuning על BERT למשימה הזאת, ביחד עם שכבת קלאסיפיקציה מעליו – השכבה תפעל על הקידוד שיוצא מ-BERT לטוקן ההתחלתי המיוחד, שאמור לייצג את כל הרצף. למשל יהיה אפשר להוסיף שכבה לינארית ומעליה רגרסיה לוגיסטית בשביל לתת ערך בין 0 ל-1 שמייצג עד כמה סביר ששני המשפטים הם בעלי אותו סגנון
   c. בקשר לכלי ה-prompting המגניב:
      i. סיבות למה כן להשתמש ב-ChatGPT בתור baseline:
         1. מדובר במודל חזק שאומן על כמות דאטא עצומה והוא נגיש לשימוש (גם אם מוגבל בזמינות במידה מסוימת) בלי צורך להקצות משאבי חישוב יקרים מהצד שלנו
         2. מדובר במודל שעורר המון שיח בחודשים האחרונים ברחבי העולם, לכן נצבר ידע רב וזמין שיכול להכווין אותנו לדעת אילו סוגי פרומפט עובדים יותר טוב עבורו ואילו מכשילים את המודל
      ii. סיבות למה לא להשתמש בו:
         1. מדובר במודל סגור, שאיננו חשופים לתהליך האימון שלו, ולכן לא נוכל לדעת מה הביצועים שלו מלמדים על הפרומפטים שלנו – ייתכן שהוא כבר נחשף בתהליך האימון לדאטא דומה או אפילו זהה.
         2. מדובר במוצר של חברה מסחרית, ואיננו חשופים לכל המודל ולפילטרים שהדאטא עובר לפני שהוא מוצג לנו בתור פלט. לכן ייתכן כי הביצועים יושפעו מהנחיות או הטיות שבעלי המוצר הכניסו לו בשביל שיתאים לצורך שלהם.

## החלק התכנותי

ה-repo שלי זמין בקישור: https://github.com/niveck/ANLP-ex1/tree/main