בס"ד,

פענוח אוטומטי של ראשי תיבות רב-משמעיים בטקסטים תורניים ללא שיטות עיבוד שפה טבעית מסורתיות

יעקב הכהן-קרנר, אריאל פרץ ואריאל קאס

המחלקה למדעי המחשב, מכון-לב

Department of Computer Sciences, Jerusalem College of Technology (Machon Lev)
21 Havaad Haleumi St., P.O.B. 16031, 91160 Jerusalem, Israel
{kerner, akass, arielp}@jct.ac.il

טקסטים עבריים בכלל וטקסטים תורניים-עבריים בפרט משופעים בראשי-תיבות. השפה העברית עצמה עשירה במספר ראשי-התיבות הקיימים בה. חלקם ניתן לפירוש ביותר מאשר אפשרות אחת. עפ"י חישובים שביצענו התברר כי השפה העברית מכילה כ- 17,000 ר"ת כלליים (וזאת ללא ר"ת רבים הייחודים לתחומים מקצועיים שונים). מתוכם לכ- 6,000 ר"ת (כ- 35%) יש יותר מאשר פירוש אחד. דוגמה קיצונית לראשי תיבות שלהן יותר מפירוש אחד הינן ראשי התיבות א"א אשר לה הוצעו כ- 110 פירושים אפשריים, ביניהם: אמר אברהם, אי אפשר, אשת איש, אבות אבותינו, אבי אבי, אבי אמי, אם אבי, אם אמי, אין אוכלים, אין אומרים ו- אם אומרים.

לעיתים קרובות, מתלבט הקורא הבלתי-מומחה בפענוח הנכון לר"ת מבין שלל אפשרויות. במקרים אחרים, לקורא אין שמץ של מושג מהן אפשרויות הפענוח. בעיית הפענוח חריפה במיוחד בטקסטים תורניים בלתי-מנוקדים המשופעים במגוון עשיר וצפוף של ראשי תיבות (מעניין לציין שריבוי ר"ת בטקסטים תורניים בלתי-מנוקדים מקורו בין היתר בסיבות הבאות: חסכון בפחם (דיו) לעתים עקב עוניו של המחבר, האיסור להעלות על הכתב דברים שבעל-פה, קיצור בכתיבה ובזמן הנדרש לה והקלה על זכרון הדברים). בעיית פענוח ראשי התיבות התורניים הינה קריטית עבור מגזרים מיוחדים של קוראים, כגון: עולים חדשים, בעלי תשובה, ילדים וכאלו הקוראים סוג טקסטים שאינו מוכר להם.

פרויקט-גמר הנמצא לקראת סיומו במסגרת המחלקה למדעי המחשב במכון-לב מציג אב-טיפוס של מערכת מעשית המפענחת באופן אוטומטי ראשי תיבות רב-משמעיים בטקסטים תורניים. המערכת מושתתת על מודל מחקר חדשני, שחלקו פורסם לפני מספר חדשים בכנס בינלאומי שפיט במדעי המחשב. מערכת זו למיטב ידיעתנו היא המערכת המעשית הראשונה מסוגה לפענוח ר"ת בשפה העברית. בשפות אחרות קיים מספר מצומצם מאד של מערכות פענוח ר"ת. רובן ככולן בתחום הרפואי: Yu ושות' דיווחו על 84% הצלחה ו- Pakhomov דיווח על 89% הצלחה בפענוח. עבור טקסטים בשפה הלטינית נכתב מודל תיאורטי אך ללא יישום מעשי.

המודל שפיתחנו משתמש בשילובים שונים של שיטות מדעיות שונות לפענוח ר"ת, ביניהן: שיטת המבוססות על תכונותיהם של הפירושים השונים לר"ת, כגון: תחילית, מילה לפני , מילה אחרי, גימטרי-ה ושיטות סטטיסטיות שונות. אין אנו משתמשים בשיטות עיבוד שפה טבעית מסורתיות המבוססות על parsing , tokens וכד'.

בשלב זה המאגר שנבדק היה חלק מהסימנים במשנה ברורה חלק ג' העוסק בהלכות שבת. המאגר הנבנה מכיל בינתיים כ – 38,340 מילים מתוכן 3122 ר"ת, ש- 1206 מהם הינם רב-משמעיים. מספר הפירושים

הרלוונטיים הממוצע לר״ת הרב-משמעיים הוא 2.3. נכון לעכשיו המערכת עומדת על אחוז הצלחה יפה של 89%, גם בהשוואה לאיכותן של מערכות דומות המפענחות באופן אוטומטי ראשי תיבות רב-משמעיים בטקסטים רפואיים בשפה האנגלית.

מודל הפענוח שלנו הינו הראשון מסוגו לשפה העברית והראשון מסוגו לתחום התורני. דומה כי אין צורך להכביר מילים על אודות הפוטנציאל החשוב הטמון בהצלחת יישום מודל מורכב כזה עבור כל טקסט שהוא. דוגמה למורכבותה של הבעיה הנדונה: מ״ב סק״ו באו״ח סימן של״ח הינו משפט בודד ארוך המכיל מס׳ רב של ר״ת. המערכת בגרסתה הנוכחית מטפלת ב- 19 מהן. ממוצע מס׳ אפשרויות הפענוח עבור כל ר״ת כזה הוא 2.9 ולכן סה״כ מספר אפשרויות הפענוח שלהן הוא כ- 2.9^{19} = 610,326,124, כשרק אפשרות אחת מהן נכונה.

אנו כעת עסוקים בהוספת שיטות בסיסיות, בהרחבה ניכרת של בסיס הנתונים ובשילוב שיטות למידה מתוחכמות.