

יעקב הכהן־קרנר, אריאל קאס, אריאל פרץ

מערכת לומדת המפענחת ראשי־תיבות רב־משמעיים בכתבים תורניים

ראשי־תיבות (ר"ת) נפוצים מאוד בשימוש בשפה העברית בכלל ובכתבים תורניים בפרט. חלק ניכר מר"ת אלו ניתנים לפירוש במספר אופנים. במאמר זה, אנו מציגים מערכת לפענוח ר"ת רב־משמעיים שפותחה במכון לב (בית הספר הגבוה לטכנולוגיה, ירושלים) ע"י הכותבים השני והשלישי תחת הנחייתו של הכותב הראשון. הפענוח התמקד בכתבים תורניים הכתובים בעברית־ארמית. פותחו שמונה־עשרה שיטות בסיסיות לפענוח: חמש־עשרה מהן מבוססות הקשר, שתיים סטטיסטיות ושיטה אחת ייחודית לשפה העברית. נבנה מאגר נתונים המכיל כמאתיים ושלושים מסמכים בהלכה יהודית ובהם מעל חצי מיליון מילים, מתוכן 42,687 ר"ת רב־משמעיים. יכולתן של השיטות השונות לפענח ר"ת רב־משמעיים נמדדה עפ"י הפענוחים הנכונים. שיטות בסיסיות אלו שולבו באמצעות שימוש בשיטת למידה ממוחשבת בשם C4.5 שהביאה לכ-97% הצלחה בפענוח ראשי התיבות הנ"ל. כיווני מחקר עתידיים אפשריים הם: פיתוח שיטות מבוססות עיבוד שפה טבעית, יצירת שיטות משולבות נוספות מהשיטות הבסיסיות, שימוש בשיטות למידה מוצלחות אחרות, בניית מאגרים נוספים מתחומים שונים בשפה העברית ובשפות אחרות, התאמת מודל הפענוח עבורם וביצוע ניסויי למידה שונים עבורם ופיתוח אלגוריתמים לזיהוי ותיקון ר"ת משובשים.

א. מבוא

אחד מנושאי המחקר האטרקטיביים בתחום המחקר של עיבוד שפות טבעיות הינו פענוח מילים רב־משמעיות (Word Sense Disambiguation). לפתרון בעיה רחבת היקף זו, תוכננו ובוצעו מחקרים רבים במספר שפות זרות, כגון: באנגלית [1], בצרפתית ואנגלית [2], בסנינית [3], בצ'כית [4] ובתאילנדית [5].

בפרוייקט מחקר זה, שאיפתנו הייתה לטפל בתת־בעיה — פענוח ראשי־תיבות רב־משמעיים בטקסטים תורניים בעברית. עד עתה נבנו מעט מודלים לטיפול בתת־בעיה זו ואף לא אחד מהם בשפה העברית.

המחקר המועט שכבר נעשה בתחום זה לא התמקד במציאת מודל כללי. המחקרים השונים ניסו לחקות את צורת החשיבה האנושית בתחום מסוים, כדוגמת מאמרים רפואיים או ספרות לטינית. מחקר זה, מחד גיסא, ייחודו בטיפול בפענוח ר"ת רב־משמעיים בשפה העברית בכלל

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

ובטקסטים תורניים בפרט, ומאידיך גיסא, שואף הוא לחקור ולבנות מודל פענוח כללי לכל תחום ושפה.

במודל כללי זה נעשה שימוש בניחות מאפיינים הקשריים וסטטיסטיים של הר"ת ובנוסף נעשה שימוש בשיטות למידה ממוחשבות לחקירת קשרים בין מאפיינים אלו. ניצול קשרים אלו עשוי להביא לשיפור משמעותי נוסף של ביצועי המערכת.

במערכות עיבוד שפה טבעית רבות נעשה שימוש במנתח תחבירי ממוחשב, המזהה את חלקי הדיבור השונים במשפט ואת תפקידיהן התחביריים של מילים שונות מתוך ההקשר המידי בטקסט. מערכות כאלו נבנות בד"כ בצורה ייחודית לשפה או לתחום מסוים ולכן הן לא ניתנות להרחבה או לבנייה של מודל כללי לפתרון בעיה. המערכת במחקר זה שומרת על כלליותה ע"י חלוקת הפענוח לשלבים מוגדרים היטב מבלי להשתמש בעקרונות הבנת שפה טבעית כזו או אחרת. באופן זה, המערכת איננה תלויה בשפה או בתחום מסוים של טקסטים. מגבלת המערכת נקבעת רק ע"י נתוני הקלט השונים מהם היא לומדת ומפתחת שיטות פענוח עבור הר"ת שהוגדרו לה. בנוסף, לא נעשה שימוש במנתח תחבירי, משום שמערכת כזו בשפה העברית לא הייתה זמינה למחקר זה באופן מלא ופיתוח מקורי של מערכת מושלמת כזו היה חורג מהמטרות המחקריות שהוצבו.

המשך המאמר מכיל את הפרקים הבאים: פרק ב עוסק בראשי-תיבות בכתבים תורניים. פרק ג מתאר מודלים קודמים ומערכות קודמות לפענוח אוטומטי של ר"ת. פרק ד מתאר את המודל שבנינו לפענוח ר"ת רב-משמעיים בטקסטים תורניים. פרק ה מציג את התוצאות הניסוייות וניתוחן. פרק ו מסכם את העבודה, מסיק מסקנות ומציע מחקר עתידי אפשרי. הנספח מציג מדגם של ר"ת הרב-משמעיים שנבחנו במאגר על פירושיהם השונים.

ב. ראשי-תיבות בכתבים תורניים

ר"ת נפוצים בשפה העברית בכלל ובטקסטים תורניים בפרט. חלק בלתי מבוטל מהם הינו רב-משמעי. השפה העברית מכילה כ-17,000 ר"ת ידועים [27] ומתוכם כ-35% הינם רב-משמעיים; כלומר, כ-6,000 ר"ת רב-משמעיים. זאת מלבד ר"ת רבים הייחודיים לתחומים מקצועיים שונים, כגון: צבא, רפואה וכו'. לדוגמה, ר"ת 'א' יכולים להתפרש עפ"י [27] ב-110 (!) אופנים שונים, וביניהם: אברהם אבינו, אי אפשר, אי אפשר, אמר אברהם, אין אומרים, אחרים אומרים, אשת איש, אם אמי, אם אבי, אבי אמי, אבי אבי.

סיבות רבות הביאו לשימוש בר"ת, ובהן:

1. עוני – בנו של הרב אברהם גומבינר, בעל המגן אברהם, העיד על אביו [28] כי עקב עוניו לא היה לו די כסף לפחם לכתובה ולכן נאלץ לכתוב בקיצור רב אשר כלל שימוש בר"ת.
2. תורה שבעל-פה – 'לעולם ישנה אדם לתלמידו דרך קצרה'. מכאן למדו רבנים בכל התקופות כי יש עניין חשוב לקצר ולכתוב לעיקרו של עניין. קצרנות זו הביאה לשימוש רחב בר"ת, נוטריקונים ורמזים [28, 29].

3. הלכה יהודית – מרן הרב יוסף קארו פסק בשו"ע [30] כי "אסור לכתוב ג' תיבות מפסוק בלא שרטוט אם הוא כתב אשורית", כלומר שכל שלוש מילים רצופות מפסוקים הכתובים בכתב עברי מרובע, יש להן גדר של כתיבת ספר תורה, תפילין ומזוזה (סת"ם) וקדושה שורה עליהן. לכן יש לכותבן עפ"י הלכות כתיבת סת"ם, כלומר עם שרטוט. לכן, נהגו חלק מהמחברים להביא חלק מציטוטים מפסוקים בר"ת כך שלא יתחייבו בכתובה מהסוג הנ"ל. כך נהג הרב אברהם יצחק הכהן קוק באזכרו פסוקים (ראה למשל ב-[31] בעמ' נו-נז).
4. כתיבה בקיצור לשם חיסכון בזמן כתיבה – נעשה שימוש רב בר"ת ובקיצורים במקום ביטויים מסוימים החוזרים על עצמם פעמים רבות או באותו הטקסט, כגון: מושגים ארוכים, שמות וביטויים שכיחים מאוד (כמו למשל: ע"מ (על-מנת), ר"ת (ראשי תיבות), וע"י (על ידי)).
5. זיכרון – פעמים שרצה המחבר להעביר מסר שייזכר גם לאחר זמן, ולכן "המציא" ר"ת למושג שלו, ובכך להקל על הקורא את זכרון הדברים. כגון: דצ"ך עד"ש באח"ב שהומצאו ע"י רבי יהודה והובאו בהגדה של פסח. ר"ת אלו מייצגים את עשר המכות (דם, צפרדע, כינים, ערוב, דבר, שחין, ברד, ארבה, חושך ובכורות).
- במשך הדורות הייתה התייחסות לבעייתיות שבריבוי תופעה זו. הרב חיים חזקיה מדיני, בספרו "שדי חמד" [30], בולט בהתייחסותו השלילית לתופעת ריבוי השימוש בר"ת בכך שהציע לבטל או למעט עד כמה שניתן בשימוש בר"ת ונוטריונים שאינם מוכרים לכל. הוא מתריע מפני הסכנות שבר"ת שאינם מפורסמים: (א) השקעת עמל רב וזמן רב בפענוח ו-(ב) פענוח מוטעה או אי הצלחה בפענוח ר"ת. זה יכול לקרות משום שהקוראים לא מצליחים לפענח או משום שהר"ת עצמו שגוי, למשל משום שהוחלפה אות באות, הושמטה אות או שנוספה אות.
- חשיבות רבה נודעת לפתרון בעיית פענוח ר"ת רב-משמעיים כאשר לקורא חסרים הכלים או הניסיון הנדרשים. דוגמאות לציבורים כאלו הם: עולים חדשים, ילדים, חוזרים בתשובה והציבור הכללי בעוסקים בטקסטים מתחומים מקצועיים או ייחודיים. לכן, המטרה העיקרית של המחקר הייתה להקל על הקוראים ולפענח בצורה נכונה ר"ת רב-משמעיים ובכך לחסוך עמל וזמן רב בפענוח. מטרה משנית הייתה האפשרות להציע מספר פירושים רלוונטיים עם דירוג רמת מידת רלוונטיות ביניהם עפ"י שיטות שונות.
- בפרוייקט מחקר זה הונח כי הר"ת בכתבים שנחקרו הינם נכונים ואין בהם שגיאות. זיהוי והצעת תיקונים לר"ת משובשים הם משימות מכובדות בפני עצמן המוצעות למחקר עתידי.

ג. מודלים קודמים ומערכות קודמות לפענוח ר"ת רב-משמעיים

ישנם מספר מודלים לפענוח ר"ת רב-משמעיים. להלן נתאר חלק מהם.

מודל "הפתרון היחיד העקבי בהקשר"

מודל "הפתרון היחיד העקבי בהקשר" נוסח ע"י Yarowsky [6]. מודל זה משער כי יש נטייה בשפות טבעיות לעקביות בסגנון הדיבור והכתיבה באותו הקשר. לפי השערה זו, ביטויים בשפה באותו הקשר חוזרים על עצמם רבות כאשר ההבנה של ביטויים אלו עקבית בכל שימושיהם. ביטוי מוגדר כמספר מילים המופיעות בקרבה זו לזו, ולא דווקא כרצף של מילים. השערה זו יכולה להיות נכונה גם לפענוח של ר"ת. כלומר, הפירוש הנכון לכל מופע של אותו ר"ת ימצא ע"פ ההקשר בו נמצא המופע המסוים, למשל בהתבסס על כל המילים הנמצאות בסמוך לפני מופע הנדון ולאחריו. מופע אחר של אותם ר"ת כעבור משפט או שניים, למשל, יכול להתפרש אחרת משום שהוא נמצא בהקשר אחר.

לדוגמה: ע"ש יכול להתפרש במספר אופנים, ביניהם 'ערב שבת' ו'עייני שם'. מילים בעלות זיקה חזקה לפירוש הראשון יכולות להיות 'הכנה', 'מבעוד' ו'יום' ולכן יכולות לרמוז על פענוח זה. שמות של ספרים או מחברים יכולים להיות בעלי זיקה חזקה דווקא לפירוש השני.

מודל "הפתרון היחיד העקבי בדיון"

מודל "הפתרון היחיד העקבי בדיון" נוסח ע"י Church, Gale, ו-Yarowsky [7]. מודל זה משער כי יש נטייה בשפות טבעיות לעקביות באותו דיון. ע"פ השערה זו, אם בדיון מסוים יש אמירה דו-משמעית ייחודית אשר כוונתה מובנת מתוך הדיון עצמו, הרי שכל שימוש באמירה שכזו בהמשך הדיון, היה מובן ע"פ הבנה ראשונית זו. השערה זו יכולה להיות נכונה גם לענייננו וייתכן כי מחברים ישתמשו בר"ת מסוים מספר רב של פעמים במאמר כלשהו, ויתכוונו תמיד לאותו פירוש בדיון אף אם הם נמצאים בהקשרים שונים.

מודל זה נוסח בהקשר של פענוח ר"ת רב-משמעיים במאמרים רפואיים ע"י Tsurouka, Yu ו-Tsujii [8]. על אף השימוש בתאוריה זו, אחוזי השיפור שהוצגו במערכת הזאת היו בשיעור של 2% בלבד לעומת הגירסה הבסיסית.

מודלים אלו ואחרים שימשו כבסיס למערכות קודמות, שביצעו כל אחת בשיטתה פענוח של ר"ת רב-משמעיים. להלן יוצגו מערכות אלו.

מערכות קודמות העוסקות בפענוח ר"ת רב-משמעיים בשפות זרות (רובן בתחום הרפואי) Tsurouka, Yu ו-Tsujii [8] פיתחו מערכת לפענוח אוטומטי של קיצורים ור"ת רב-משמעיים במאמרים רפואיים בשיטת הלמידה LIBSVM [9] תוך שימוש בתאוריית "הפתרון היחיד העקבי בדיון". במערכת זו נעשה שימוש בהקשר ע"פ שתי מילים לפני הר"ת ושתי מילים אחריה. מאגר התקצירים הרפואיים נאסף מתוך מאגר מקוון בשם MEDLINE [10]. בניסוי ראשון נבחנו

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

6 קיצורים ור"ת לפענוח ובניסוי שני נבחנו 10 קיצורים ור"ת אחרים לפענוח. מערכת זו השיגה 87.47% ו-84.31% הצלחה בהתאמה.

Min-Yen [11] פיתח מערכת בשם Meurlin (Metadata Extraction from URLs) לכריית מידע-על מתוך כתובות של אתרי אינטרנט. המאגר שנבחן הכיל כ-1.6 מיליון כתובות כאלו. המערכת פענחה 10 קיצורים ור"ת, כחנה ופענחה תוך שימוש בשיטת הלמידה הממוחשבת Boosting המופיעה במערכת BoosTexter [12]. הערכת הישגי המערכת נמדדה ע"י שימוש במדד ה-F-Measure עם ערך $\alpha=1$. המערכת השיגה תוצאה של 62% באמצעות שימוש בשיטה הבסיסית של בחירה עקבית של "הפירוש הנפוץ במאגר" לר"ת הנדון ותוצאה של 80% באמצעות שימוש בשיטה הבסיסית "כל המילים במשפט" (שיטה 15 בפרק הבא).

Pakhomov [13] פיתח מערכת לפענוח אוטומטי של ר"ת במאמרים רפואיים ע"י שימוש בשיטת הלמידה הממוחשבת Maximum Entropy. מערכת זו מממשת שני מאפיינים של ר"ת וקיצורים ע"י השיטה הבסיסית "הקשר עפ"י שתי מילים לפני ואחרי" וע"י השיטה הבסיסית "הקשר עפ"י רמת פסקה". המאגר שנבחן הכיל כ-10,000 מאמרים רפואיים. בניסוי ראשון נבחנו 6 ר"ת ובניסוי שני נבחנו 69 ר"ת. עבור הקבוצה הראשונה, המערכת השיגה תוצאות של כ-89% ו-90% עבור שני הניסויים.

Pustejovsky ואחרים [14] פיתחו מערכת לפענוח של ר"ת רב-משמעי יחיד, במאגר המונה 52 מאמרים בלבד. המערכת השיגה תוצאה של 97.62% בפענוח מופעי הר"ת, ע"י שימוש בסכמת משקלים הנקראת ATC [15].

Rydberg-Cox [16] פיתח מערכת לפענוח אוטומטי של ר"ת בטקסטים לטיניים מוקדמים בהתבסס על שלושה מאפיינים של ר"ת ע"י שיטה המשלבת ניתוח תחבירי של הר"ת במשפט, עם פענוח בשיטת "הפירוש הנפוץ במאגר" (מקביל לשיטת CC המתוארת בפרק הבא) ועם פענוח על בסיס ההקשר בו מופיע מופע הר"ת. מערכת זו אינה משתמשת בשיטות למידה ממוחשבות. אין מידע סטטיסטי על המאגר שנבחן או על הר"ת הרב-משמעיים שפוענחו.

מחקר סטטיסטי על ר"ת בכלל ור"ת בעלי שלוש אותיות בפרט הוצג ע"י Liu ואחרים ב-[17] ו-[18], בתחום המאמרים הרפואיים בשפה האנגלית. מערכות המייצרות מילוני ר"ת ופירושיהם פותחו ע"י Adar, Yoshida, Fukuda ו-Takagi [19], Yu ו-Hripcsak ו-Friedman [21].

פענוח ר"ת רב-משמעיים בכתבים תורניים בעברית

ראשיתו של מחקר זה פורסמה על-ידינו במאמר [22]. מאמר זה הציג 6 שיטות בסיסיות לפענוח ר"ת רב-משמעיים. נוסו שילובים פשוטים של השיטות ללא שימוש בשיטת למידה ממוחשבת כלשהי. תוצאת הפענוח שם הייתה נמוכה מאוד (בסביבות 60%). המחקר המתואר במאמר הנוכחי מציג הרחבה ניכרת ומאידך שיפור ניכר המתבטאים במובנים הבאים: (1) נוסחו ומומשו שיטות בסיסיות נוספות (18 בסה"כ כמפורט בפרק הבא), (2) מסד הנתונים הנבחן הורחב בהרבה ממאגר המכיל כ-19,000 מילים ובהן כ-1,500 ר"ת רב-משמעיים למאגר המכיל מעל לחצי מיליון

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

מילים, מתוכן 42,687 ר"ת רב-משמעיים ו-3) יושמה שיטת למידה אוטומטית בשם J48 שהביאה לתוצאת פענוח של כ-97% שהינה מצוינת גם בהשוואה למערכות מקבילות בשפות אחרות.

ד. המודל שנבנה לפענוח ר"ת רב-משמעיים בטקסטים תורניים

שיטות בסיסיות לפענוח ר"ת רב-משמעיים

ר"ת רב-משמעיים ניתנים לפענוח במספר פירושים. אולם בהקשר מסוים הם בר"כ בעלי פענוח מסוים אחד בלבד. פענוח ר"ת כאלו ע"י בני-אדם מונחה ע"י שיקולים המופעלים בדרך-כלל באופן אינטואיטיבי, שקשה לעקוב אחריו. אולם, כאשר תהליך הפענוח מומר לתהליך ממוחשב ואוטומטי, הרי שיש להגדיר מגוון רחב של שיקולים אפשריים. תהליך ההחלטה האנושי נתמך ברובו בזיכרון וניסיון קודם ולא מן הנמנע כי הוא משתמש בשילוב של כמה מאפייני החלטה לקבלת החלטה סופית.

להלן מובא פירוט של שמונה-עשר מאפיינים שונים, כאשר כל מאפיין הינו עצמאי. המאפיינים חולקו לשלוש קבוצות: (א) מאפיינים סטטיסטיים, (ב) מאפיין ייחודי לשפה העברית ו-1) (ג) מאפיינים הקשורים.

קבוצת המאפיינים ההקשריים בשונה מהקבוצות האחרות מתייחסת לתכונות עמוקות יותר של הטקסט, ונוגעת ברבדי המשמעות וההקשר של הטקסט. בקבוצת מאפיינים אלו נמצאים מאפיינים העונים לשאלות כמו "האם התחילית של הר"ת חותרת לפענוח מסוים?", וכן "האם המילה שמופיעה לאחר הר"ת יכולה להופיע רק אחרי פענוח מסוים?".

(א) מאפיינים סטטיסטיים

1. Context Common Rule (CC) – הנפוץ במאגר

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י עיבוד סטטיסטי של ההקשר הכולל של הר"ת במאגר הנוכחי הנבחן. הפירוש הנבחר הוא הפירוש הנפוץ ביותר מבין כל פירושי הר"ת במאגר הנוכחי. שיטה זו הוגדרה ע"י Hripcsak, Yu ו-Friedman [21].

2. Language Common Rule (LC) – הנפוץ בשפה

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י עיבוד סטטיסטי של כל המאגרים המשתתפים במחקר. מובן כי ככל שאוסף המאגרים גדול יותר, שיטה זו מייצגת בצורה טובה יותר את השפה כולה. הפירוש הנבחר הוא הפירוש הנפוץ ביותר מבין כל פירושי הר"ת בכל המאגרים.

(ב) מאפיין ייחודי לשפה העברית

3. Gimartia Rule (GM) – גימטריה

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י סכום הערכים המספריים של אותיות

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

הר"ת לפי שיטת הגימטריה הפשוטה ביותר (א=1, ב=2, ..., י=10, כ=20, ..., ק=100, ר=200, ...). הפירוש הנבחר הוא הערך המספרי של הר"ת או תשובה שלילית במקרה שאין ערך מספרי תקף לפי כללי הגימטריה. חוסר תקינות מוגדרת כאשר האותיות של הר"ת אינן מופיעות בסדר מימין לשמאל עפ"י פונקציה מונוטונית יורדת של הערכים המספריים של אותיות הר"ת. חשוב לציין כי פעמים רבות מחברים נמנעים מלערבב בטקסט שלהם אותיות וספרות, ולכן בכואם לצטט מקור מסוים, יעדיפו להשתמש בגימטריה של האותיות שמציינות את מיקום האמרה.

(ג) מאפיינים הקשריים

4. Prefix Counted Rule (PRC) – תחילית ממוספרת

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י התחילית (רצף אותיות) הצמודה לה מלפניו, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על תכונה של השפה העברית שלפיה תחילית מסוימת תבוא דווקא לפני פירושים מסוימים של הר"ת, בגלל הקשרים תחביריים של השפה. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף הקידומות של כל ר"ת לכל פירושיהם. בנוסף נספרים מופעי הקידומות. אחר בניית השיטה, במעבר על טקסט חדש, תשווה התחילית של ר"ת מסוים לכל אוספי התחיליות לכל פירוש של הר"ת. הפירוש עבורו תהיה ההתאמה הטובה ביותר, כלומר מספור התחילית עבור אותו פירוש הוא הגדול ביותר, ייבחר כפירוש הנכון.

שיטות 5-12 (לקמן) מבוססות על טווח של עד ארבע מילים לפני/אחרי הר"ת. הן מתבססות על עקרון הגבלת הזיכרון לטווח קצר בתחום של שבע פלוס-מינוס שתיים (4 מילים לפני הר"ת, הר"ת עצמו ו-4 מילים אחרי, סה"כ 9 שהוא $(7+2)$ Miller [22]).

5-8. Before K Word Counted Rule (BKWC) – K מילים לפני הר"ת באותו

משפט

קבוצת שיטות זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י K המילים המופיעות לפניו באותו משפט, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על תכונה של שפות טבעיות (אנושיות) שבהן קיימים רצפי מילים באורך שונה בעלי משמעות כוללת. דוגמאות בולטות לכך הן ניבים ומושגים. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף K המילים המופיעות לפני כל ר"ת לכל פירושיהם באותו משפט. אם אין K מילים לפני הר"ת במשפט, יילקחו רק המילים שישנן מתחילת המשפט מבלי "לגלוש" למשפט הקודם. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושוו K המילים המופיעות לפני ר"ת מסוים לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי K המילים באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון.

9-12. **(1,2,3,4) After K Word Counted Rule (AKWC) – K מילים אחרי הר"ת באותו**

משפט

קבוצת שיטות זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י K המילים המופיעות אחריו באותו משפט, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על תכונה של שפות טבעיות (אנושיות) שבהן קיימים רצפי מילים באורך שונה בעלי משמעות כוללת. דוגמאות בולטות לכך הן ניבים ומושגים. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף K המילים המופיעות אחרי כל ר"ת לכל פירושיהם באותו משפט. אם אין K מילים אחרי הר"ת במשפט, ייאספו רק המילים שישנן מבלי "לגלוש" למשפט הבא. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושוו K המילים המופיעות אחרי ר"ת מסוים לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי K המילים באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון.

13. **Before Sentence Counted Rule (BSC) – כל המילים המופיעות לפני ר"ת באותו**

משפט

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים המופיעות לפני ר"ת באותו משפט, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון האנושי שקובע כי משמעות ר"ת נקבעת מתוך המשפט בו הוא מופיע. בנוסף נספרים מופעי המילים. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות לפני כל ר"ת לכל פירושיהם באותו משפט. אחר בניית השיטה, במעבר על טקסט חדש, יושוו כל המילים המופיעות לפני ר"ת מסוים באותו משפט לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון.

14. **After Sentence Counted Rule (ASC) – כל המילים המופיעות אחרי ר"ת באותו**

משפט

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים המופיעות אחרי ר"ת באותו משפט, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון האנושי שקובע כי משמעות ר"ת נקבעת מתוך המשפט בו הוא מופיע. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות אחרי כל ר"ת לכל פירושיהם באותו משפט. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושוו כל המילים המופיעות אחרי ר"ת מסוים באותו משפט לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון.

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

15. All Sentence Counted Rule (AISC) – כל המילים המופיעות באותו משפט

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים הסובבות אותו באותו משפט, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון האנושי שקובע כי משמעות ר"ת נקבעת מתוך המשפט בו הוא מופיע. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות לפני ואחרי כל ר"ת לכל פירושיהם באותו משפט. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושו כל המילים המופיעות לפני ואחרי ר"ת מסוים באותו משפט לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון. שיטה זו הינה איחוד שתי השיטות BSC ו-ASC.

16. Before File Counted Rule (BFC) – כל המילים המופיעות לפני ר"ת באותו קובץ

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים המופיעות לפני ר"ת באותו קובץ, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון האנושי שקובע כי משמעות ר"ת נקבעת מתוך ההקשר הרחב בו הוא מופיע. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות לפני כל ר"ת לכל פירושיהם באותו קובץ. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושו כל המילים המופיעות לפני ר"ת מסוים באותו קובץ לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון. במקרה של שיוויון או חוסר במילים לפני הר"ת (תחילת קובץ), השיטה מחזירה תשובה שלילית.

17. After File Counted Rule (AFC) – כל המילים המופיעות אחרי ר"ת באותו קובץ

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים המופיעות אחרי ר"ת באותו קובץ, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון האנושי שקובע כי משמעות ר"ת נקבעת מתוך ההקשר הרחב בו הוא מופיע. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות אחרי כל ר"ת לכל פירושיהם באותו קובץ. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושו כל המילים המופיעות אחרי ר"ת מסוים באותו קובץ לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון.

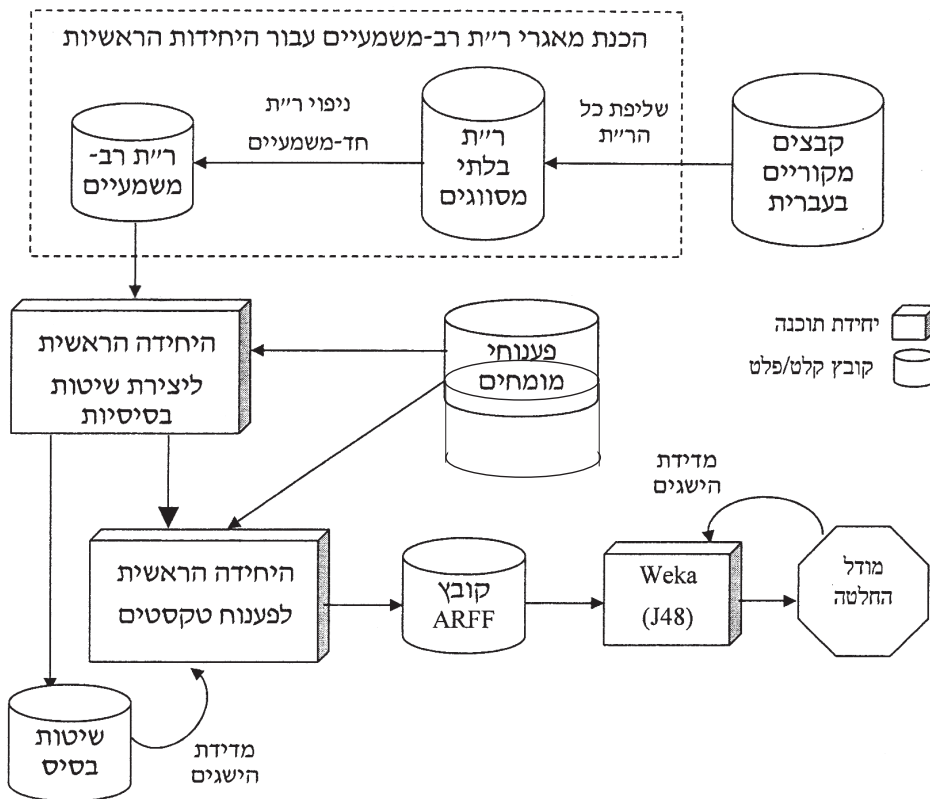
18. All File Counted Rule (AIFC) – כל המילים המופיעות באותו קובץ

שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י כל המילים הסובבות אותו באותו קובץ, בהתאם ליחס הממוספר של הפתרונות כפי שנצפה במאגרי המידע. שיטה זו מתבססת על ההיגיון

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

האנושי שקובע כי משמעות ר"ת נקבעת מתוך ההקשר הרחב בו הוא מופיע. בניית השיטה נעשית ע"י מעבר על קבצי אימון ואיסוף כל המילים המופיעות לפני ואחרי כל ר"ת לכל פירושיהם באותו קובץ. בנוסף נספרים מופעי המילים. אחר בניית השיטה, במעבר על טקסט חדש, יושו כל המילים המופיעות לפני ואחרי ר"ת מסוים באותו קובץ לכל אוספי המילים לכל פירוש של הר"ת ויעשה סיכום של מספורי המילים האלו באוסף. הפירוש שלו סכום מספורי המילים הגדול ביותר, ייבחר כפירוש הנכון. שיטה זו הינה איחוד השיטות BFC ו-AFC.

תהליך זרימת המידע בין יחידות המערכת השונות מתואר באיור 1:



איור 1: זרימת המידע בין היחידות השונות במערכת

שיטת למידה ממוחשבת הינה שיטה המאפשרת לתוכנת מחשב לשפר את ביצועיה תוך כדי ביצוע משימתה. דיון קצר בנושא זה ניתן לראות במאמר [36]. למחקר המתואר במאמר זה נבחרה שיטת הלמידה J48 היות שאופן עבודתה דומה לשיטת פענוח הר"ת האנושית. שיטת

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

הלמידה J48 הינה הגרסה של Weka¹ לשיטת C4.5, שהוגדרה ע"י Quinlan [42], אשר שייכת לקבוצת השיטות המייצרות עץ החלטה. עץ החלטה הינו ניסיון לחקות את צורת המחשבה האנליטית האנושית. להלן תוצג דוגמה ליצירת עץ החלטה. לשם כך נציג תחילה קבוצת מבחן של מאפייני ר"ת י"א בטבלה 1.

מס' פירוש	פירוש נכון	מילה לפני	קידומת	מילה אחרי
1	יש אסורים'	'ז'	'וי'	'ביחד'
2	יש אומרים'	'סתם'	'וי'	'הלכה'
3	י"א לחודשי'	'לקרותה'	'בי'	'וי"בי'
4	יש אומרים'	'בש"עי'	'בי'	'בתרא'
5	'11'	'בש"עי'	'סי'	'ששנה'

טבלה 1: קבוצת המבחן לדוגמה של מאפייני ר"ת י"א

בקבוצת מבחן זו קבוצת המשתנים הקטגוריאליים היא {'11', 'יש אומרים', 'י"א לחודש', 'יש אסורים'}. בקבוצת המבחן לדוגמה לא קיימים משתנים רציפים, שהרי ר"ת תלויים בעיקר בנתונים סטטיסטיים או הקשריים, כלומר מילים, שאינם משתנים רציפים. לכן המשתנים הלא-קטגוריאליים הם רק בעלי ערכים קבועים: מילה אחרי, קידומת ומילה אחרי הר"ת.

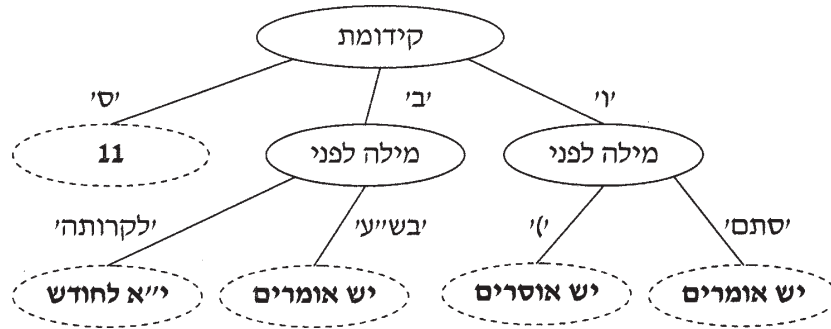
עץ ההחלטה ייבנה כך שכל צומת פנימי בעץ, כולל השורש, הינו שם של אחד מהמשתנים הלא-קטגוריאליים, והעלים בעץ הינם הערכים של המשתנים הקטגוריאליים. על הענפים המקשרים בין הצמתים והעלים בעץ, נמצאים ערכיהם של המשתנים הלא-קטגוריאליים. מבנה זה מתאר את מערכת הכללים שבונה שיטת הלמידה עבור סט הרשומות, כך שבהינתן רשומה חדשה תדע המערכת לסווג אותה בהתאם.

אבל יצירת עץ החלטה איננה טריוויאלית, שהרי ניתן לייצר עצי החלטה רבים עבור כל קבוצה של משתנים קטגוריאליים ולא-קטגוריאליים. כדי לבחור את עץ החלטה הטוב ביותר, דהיינו זה שמספק את המסלולים הקצרים לסיווג רשומה חדשה, יש להשתמש בתאוריית האינפורמטיביות, אשר הוגדרה ע"י Shannon [26].

עפ"י קבוצת המבחן, עץ החלטה לפענוח י"א המתקבל מתואר באיור 2. עץ החלטה זה הוא העץ האופטימלי לפענוח הר"ת. בעץ המתואר, המשתנים הקטגוריאליים {'11', 'יש אומרים', 'י"א לחודש', 'יש אסורים'} מיוצגים ע"י הערכים {'פירוש 1', 'פירוש 2', 'פירוש 3', 'פירוש 4'} בהתאמה.

1 מערכת הלמידה Weka (Waikato Environment for Knowledge Analysis) [24] היא אוסף של אלגוריתמים ללמידת מכונה המתוכנתים בשפת Java עבור משימות של כריית מידע, כגון: סיווג, רגרסיה, אישכול, וכללי היסק. Weka מספקת סביבת מחקר רחבה ונוחה לחוקר המבקש לטפל בכמויות מידע גדולות, ולכן מספר רב של מחקרים בתחום הלמידה משתמשים בה. קובץ הנתונים בו משתמשת Weka הינו מסוגל ARBF (הסברים ב-124).

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ



איור 2: עץ J48. עפ"י קבוצת המבחן לדוגמה לפענוח י"א

אופטימיזציה של שיטות הלמידה בניסויים

במערכת הנידונה נבחנו מספר רב של ערכים עבור פרמטרי שיטת J48 במסגרת Weka, אשר השפיעו במידה רבה על זמן פעולת המערכת ותצורת מערכות ההסקה שהתקבלו. על אף הניסיונות הרבים, התוצאות נשארו דומות בתחום של $\pm 1\%$. היות שערכי האתחול של האלגוריתמים השונים נבחרים בכל הרצה באופן אקראי, הנחת העבודה הייתה כי בכל הניסויים האפשריים, ע"י שימוש בערכי פרמטרים שונים עבור כל שיטה לחוד, מערכות ההסקה שתתקבלנה תשאפנה לאותה רמת הישגים. לכן בכל מדידות ההישגים שתובאנה בהמשך, ההתייחסות תהיה רק עפ"י ערכי ברירת המחדל שנקבעו ע"י מפתחי האלגוריתמים.

ה. תוצאות ניסויים

בפרק זה מפורטות התוצאות שהשיגה המערכת שנבנתה. הנתונים מתייחסים להשוואה ממוחשבת של החלטת המערכת לבין פענוחי המומחים, שהם ההחלטה האנושית.

הטקסטים לפענוח שנבחרו

הטקסטים המהווים את בסיס הנתונים נלקחו משני מקורות: כל כרך ג של המשנה ברורה [33] וכן 130 פסקי שו"ת של הרב עובדיה שנלקטו ממאות פסקיו הנמצאים בספרי פסקי השו"ת שלו "ביע אומר" [34] ו"יחווה דעת" [35]. על אף היותם של מקורות אלו כתבים בנושאים הלכתיים בלבד, ישנם הבדלי סגנון רבים ביניהם. הבדלים אלו נובעים ממספר גורמים:

- שני המחברים פעלו / פועלים בדורות שונים והינם בעלי תרבות שונה. רבי ישראל מאיר הכהן, בעל ה"משנה ברורה" [33], הינו פוסק אשכנזי מובהק, שנפטר לפני כ-70 שנה. הוא חי ופעל בקהילה היהודית שבעיירה ראדין הסמוכה לעיר ווילנה שבליטא. הרב עובדיה יוסף, הראשון לציון והרב הראשי הספרדי לשעבר במדינת ישראל, חי ופועל כיום בירושלים. הרב עובדיה הינו פוסק ספרדי מובהק, ובעל השפעה מכרעת על יהודים ממוצא זה.

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

- החיבור משנה ברורה נכתב כפירוש והרחבה לחלק אורח חיים של ספר "שולחן ערוך" והגהות הרמ"א, אשר נכתבו ע"י הרב יוסף קארו והרב משה איסרליש בסביבות שנת שכ"ה (1565 למנינים). בשונה מכך, מאגר פסקי השו"ת של הרב עובדיה, שבו כל פסק מורכב משאלה ותשובה הלכתית הנובעת ממספר מקורות, שונה בהחלט בסגנון ובאופי הכתיבה.
- מאגר הנתונים שנלקח ממשנה ברורה חלק ג מתמקד בהלכות שבת בלבד. בניגוד לכך, פסקי השו"ת "יביע אומר" ו"יחווה דעת" עוסקים בהלכות ממגוון תחומים (כולל שבת).

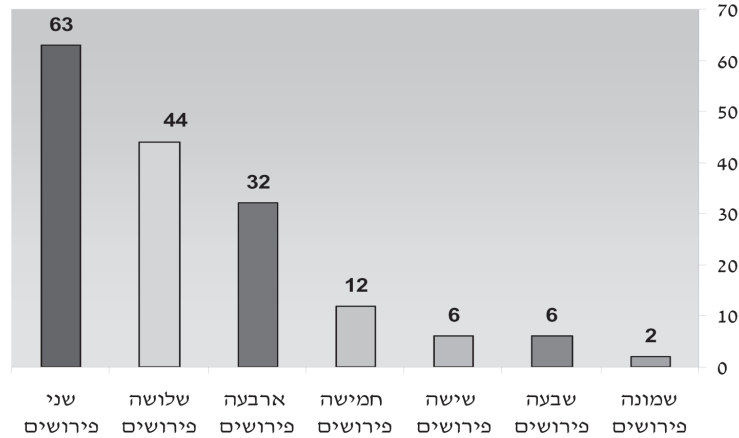
במאגרים אלו ישנם ר"ת רבים, חלקם ר"ת חר-משמעיים וחלקם ר"ת רב-משמעיים. ניתוח סטטיסטי מגוון של המאגרים מוצג בטבלה 2.

שני המאגרים יחד	130 פסקי שו"ת של הרב עובדיה יוסף (יביע אומר ויחווה דעת)	משנה ברורה (מ"ב) חלק ג סימנים רמב-שדמ	
233	130	103	סימנים/קבצים במאגר
564,554	408,200	156,354	מספר מילים כולל ר"ת (A)
114,814	90,431	24,383	מופעי ר"ת (B)
42,687	32,755	9,932	מופעי ר"ת רב-משמעיים (C)
20.34%	22.15%	15.59%	אחוז ר"ת ממילים (% B/A)
37.18%	36.22%	40.73%	אחוז ר"ת רב-משמעיים מכלל הר"ת (% C/B)
7.5%	8%	6%	יחס הר"ת הרב-משמעיים לכלל המילים במאגר (% C/A)
2,423	3,140	1,518	ממוצע מילים בכל סימן (טקסט אחד)
183	252	96	ממוצע הופעות ר"ת רב-משמעיים בכל סימן (טקסט אחד)

טבלה 2: התפלגות מופעי ר"ת במאגרים שנבחנו

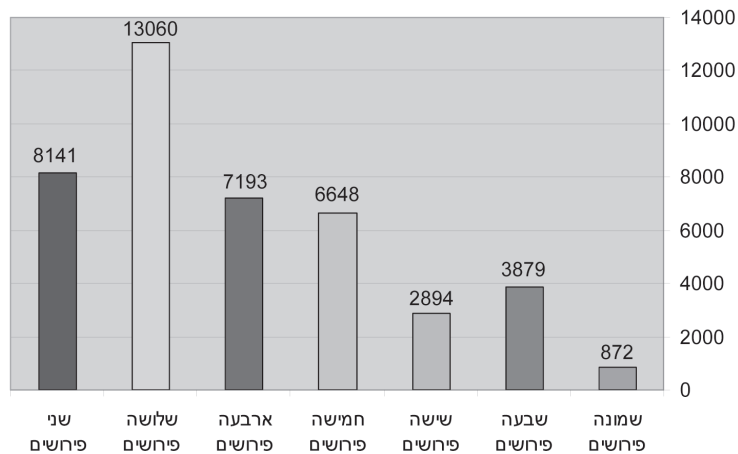
המערכת המוצגת במחקר זה נבחנה על 165 ר"ת רב-משמעיים שונים (מספר רב באופן משמעותי יחסית למחקרים אחרים). לר"ת אלו 42,687 הופעות בשני המאגרים הנבחנים יחד. מספר הפירושים הממוצע לר"ת רב-משמעיים כזה הוא 3.27 ומספר ההופעות הממוצע לכל ר"ת רב-משמעיים הוא 259. התפלגות מספר הפירושים עבור ראשי התיבות הללו מתוארת באיור 3.

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ



איור 3: התפלגות מספר פירושים עבור הר"ת בכל המאגרים

כאשר משווים בין התפלגות מספר הפירושים במחקר המוצג במאמר זה לבין מערכות קודמות (פרק ג), הרי שהמערכת של Yu et al. נבחנה על שני מאגרים של 6 ו-10 ר"ת רב-משמעיים, המערכת של Min-Yen נבחנה על מאגר של 10 ר"ת רב-משמעיים, המערכת של Pakhomov נבחנה על מאגר של 6 ו-69 ר"ת רב-משמעיים והמערכת של Pustejovsky et al. נבחנה על מאגר עם ר"ת רב-משמעי יחיד. לא ידועים נתונים סטטיסטיים על מספר המילים והר"ת, התפלגות מספר הפירושים ואף לא ממוצע הפירושים לכל ר"ת רב-משמעי במאגרים ששימשו במערכות מחקר הנ"ל, למעט עבור 6 הר"ת שנחקרו ע"י Pakhomov שידוע לגביהם כי יש 9 פירושים בממוצע לכל ר"ת.



איור 4: התפלגות מספר פירושים עפ"י מספר הופעות בכל המאגרים

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

יש להדגיש כי במסגרת בחירת הר"ת לבחינה במחקר זה, נלקחו רק ר"ת שלהם לפחות 20 הופעות במאגרי הבחינה כדי לא לגרוע מטיב הלמידה הממוחשבת ותוצאותיה. בנוסף, התפלגויות מספר הפירושים המוצגות באיור 4 רלוונטיות רק לתחום המאגרים שנבחנו. מובן שבשפה העברית כולה, מספר הפירושים לכל ר"ת ומספר ההופעות לכל הר"ת שנבחרו יהיו ברוב המקרים הרבה יותר ממספר הפירושים וממספר ההופעות שנצפו במאגרי הבחינה. אך כאמור, פרוייקט מחקר זה כלל רק פירושים הרלוונטיים לר"ת במאגרי המסמכים התורניים שנבדקו. כדי לאמוד את איכות התשובות שמציעה המערכת, יש לבחור לכל ר"ת פענוח נכון על ידי גורם אנושי, ולהשוות לבחירת פענוח הר"ת של המערכת. הר"ת שפוענחו על ידי גורם אנושי נקראים פענוחי מומחים.

לשם כך היה צורך בפענוחים שהוכנו מראש, על-ידי בני אדם, באמצעות חקירת ההקשר של מופעי הר"ת עבור כל הר"ת במאגרים. ברם, רק למאגר המשנה ברורה נמצאו פענוחים מוכנים מראש עבור רובם המכריע של ראשי התיבות. למרות זאת, לא נמצאו פענוחים מוכנים מראש עבור המאמרים בפסקי שו"ת של הרב עובדיה. אי לכך, לכל הר"ת שלא נמצא להן פענוח מוכן מראש, נעשתה עבודת חקירה ידנית של הקשר הר"ת בכדי למצוא את הפענוח הנכון.

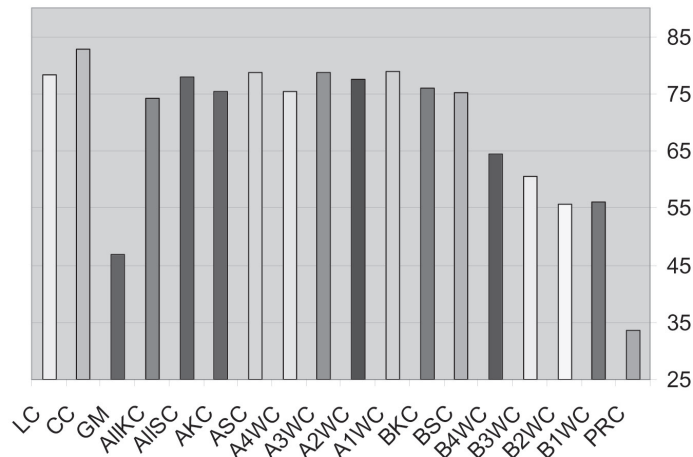
מדידת הישגים

בבואנו לבחון את הצלחת המערכת, יש צורך לקבוע במדויק כיצד לבצע זאת. חשיבות הגדרת המדד היא כפולה. הפן האחד, הוא ההשוואה הפנימית – כדי למצוא האם המערכת משתפרת עם הפיתוח, כדי להשוות בין שיטות עבודה שונות של אותה המערכת, ולשם השוואת תוצאות ניסויים שונים באופן אמין. הפן השני הוא חיצוני – כדי שניתן יהיה להשוות את איכות תוצאות המערכת לתוצאות מערכות מקבילות.

משום שמדד הדיוק (Precision) הוא המדד המקובל והנפוץ ביותר (ובד"כ היחיד) במערכות אחרות לפענוח אוטומטי של ר"ת בטקסטים, ביצועי המערכת שפותחה נמדדו עפ"י מדד זה. מדד הדיוק מוגדר כשיעור פענוחי הר"ת ע"י המערכת הזוהים לפענוחי המומחים (כלומר הפענוחים הנחשבים נכונים), מתוך כלל פענוחי הר"ת (הנכונים והלא נכונים) ע"י המערכת. במילים פשוטות ניתן לומר שמדד זה בודק מהו אחוז ההצלחה היחסי של פענוחי הר"ת הרב-משמעיים ע"י המערכת.

תוצאות השיטות הבסיסיות

תוצאות 18 השיטות הבסיסיות מוצגות באיור 5.



איור 5: תוצאות השיטות הבסיסיות

ניתוח תוצאות השיטות הבסיסיות

שיטת CC (הנפוץ במאגר) הניבה את התוצאה הטובה ביותר – 82.84% הצלחה. שיטה סטטיסטית נוספת, LC (הנפוץ בשפה), הניבה תוצאה גבוהה יחסית גם כן, 78.34% הצלחה. ייתכן כי תוצאות אלו אינן מייצגות, מפני שהשפה הכוללת הוגדרה על שני מאגרים מתחומים קרובים בתוכנם ובנושא התייחסותם, או אולי משום שהמחברים הספציפיים שנבחרו נוטים לייחס פירוש קבוע לר"ת שנבחנו. לכן, אולי כאשר השפה הכוללת תוגדר על תחומים שונים נוספים ו/או ייבחנו מחברים אחרים, תוצאות שיטות אלו עלולות להיפגע.

שיטת A1WC (מילה אחת אחרי הר"ת באותו משפט) הניבה את התוצאה השלישית הטובה ביותר והראשונה בטיבה משיטות ההקשר. שיטה זו קובעת את הפירוש הנכון של ר"ת מסוים עפ"י המילה הבודדת המופיעה אחריו באותו משפט.

מז העבר השני, שיטות PRC (תחילית ממוספרת) ו-GM (גימטריה) הניבו את התוצאות הנמוכות ביותר, עם 33.67% ו-46.82% בהתאמה. מניתוח פענוחי שיטות אלו, עולה כי במרבית המקרים היה חוסר פענוח ואילו במיעוטם הייתה טעות בפענוח. ע"י סקירת מופעי הר"ת בטקסטים התגלה כי לא הייתה תחילית לר"ת או שהר"ת לא ייצג ערך גימטריה תקני. דבר זה מעיד כי מאפיינים אלו אינם מספיקים לפענוח ר"ת בפני עצמם, אלא מוסיפים רובד פענוח לשאר השיטות, או שהם מתבטאים במופעים מסוימים ומשליכים על מופעים אחרים בטקסט.

כמו כן, מתוצאת שיטת GM ניתן לראות כי בטקסטים שנבחרו, כ-50% מר"ת מייצגים דווקא ערך גימטריה כלשהו, ולא קיצור של רצף מילים. ייתכן שיש לחקור טקסטים נוספים כדי להסיק מסקנות חותכות יותר על מאפייני ר"ת בשפה העברית.

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

ניתן לראות כי כעיקרון שיטות ה-After, כלומר AFC (כל המילים המופיעות אחרי ר"ת באותו קובץ), ASC (כל המילים המופיעות אחרי ר"ת באותו משפט), A1WC-A4WC (4-1) מילים אחרי הר"ת באותו משפט), הניבו תוצאות גבוהות בצורה מובהקת, משיטות ה-Before המתאימות להן, וכולן אינן יורדות מגבול ה-75% הצלחה. ניתן להבין כי קיים קשר הדוק יותר בין הפירושים השונים לר"ת לבין המילים העוקבות אותם ולכן השימוש בתאוריית "הפתרון היחיד העקבי בהקשר" איכותי יותר כאשר חוקרים מאפיינים שונים של ר"ת. עבור ר"ת בפרט, תאוריה זו מעידה על קשר חזק בין הר"ת לבין המילים שאחריו, וקשר חלש יותר, אך עדיין חזק בפני עצמו, בין הר"ת לבין המילים שלפניו.

השילוב בין שיטות ה-After וה-Before, המיוצג בשיטות AIIFC (כל המילים המופיעות באותו קובץ) ו-AIISC (כל המילים המופיעות באותו משפט), לא שיפרו את תוצאות השיטות, ולעתים אף להפך. ייתכן כי הסיבה לכך היא שזיקת הר"ת למילים שלפניו או שלאחריו גבוהה יותר מאשר הזיקה שלו לצירוף שלהם.

בנוסף, ראוי לציין כי יש שיפור ניכר בתוצאות שיטות ה-Before, ככל שנכללות יותר מילים בבניית השיטה. על אף זאת, בשיטות ה-After לא ניתן להצביע על שינוי מונוטוני לצד כלשהו בתוצאות השיטות. ייתכן כי הקשר החזק בין הר"ת לבין המילים שאחריו רלוונטי רק עבור המילה הראשונה אחרי הר"ת ואילו עבור המילים שלפני הר"ת, ככל שמוסיפים יותר מילים, כך הזיקה לפירוש מסוים גדלה.

תוצאת הלמידה הממוחשבת במחקר הכללי שהושגה עבור שילוב השיטות הבסיסיות ע"י שיטת J48 הייתה 96.95% (!). סה"כ הושג שיפור של כ-14% בהשוואה לתוצאה שהושגה עבור השיטה הבסיסית הטובה ביותר (82.84%).

1. סיכום, מסקנות ומחקר עתידי

תוצאותיהן של המערכות לפענוח אוטומטי של ר"ת בשפות אחרות שהוצגו בפרק ג מושוות בטבלה 3 לתוצאות המערכת שפותחה במחקר זה.

מאגר ר"ת 2		מאגר ר"ת 1		
מספר ר"ת במאגר	אחוזי הצלחה	מספר ר"ת במאגר	אחוזי הצלחה	
6	87.47%	10	84.31%	Yu et al., 2003
---	---	10	80%	Min-Yen, 2002
69	89.17%	6	89.66%	Pakhomov, 2002
---	---	1	97.62%	Pustejovsky et al., 2001
---	---	165	96.95%	המערכת שפותחה במחקר זה

טבלה 3: השוואת אחוזי הצלחה של מערכות שונות לפענוח אוטומטי של ר"ת רב-משמעיים

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

מהשוואת הישגי המערכות, ניתן לראות כי המערכת שפותחה בפרוייקט מחקר זה משיגה תוצאות טובות באופן משמעותי מרוב שאר המערכות.

אמנם להשוואה זו מספר הסתייגויות חשובות:

1. המערכות האחרות חקרו את בעיית הר"ת הרב-משמעיים בשפה האנגלית בלבד ואילו המערכת שפותחה חקרה ר"ת בשפה העברית בלבד.

2. המערכות השונות חקרו מאגרים שונים וכן ר"ת שונים ומובן כי מספר ההופעות הממוצע ומספר הפירושים השונים בממוצע לכל ר"ת, שונה ממערכת למערכת.

על אף הסתייגויות אלו, במערכת זו מספר הר"ת הרב-משמעיים וכן מספר המאפיינים עבור כל ר"ת שנחקרו עולים על מספרם המתאים במערכות אחרות.

המערכת של Pustejovsky אמנם השיגה תוצאות גבוהות מעט מעל תוצאות המערכת שפותחה בפרוייקט מחקר זה. אך המערכת של Pustejovsky חקרה רק ר"ת רב-משמעי יחיד ובתחום של 52 מאמרים בלבד.

על כן לא ניתן להסיק באופן חד משמעי כי המערכת שפותחה טובה יותר מהמערכות האחרות וכן לא ניתן להסיק באופן חד-משמעי כי בעיית הר"ת הרב-משמעיים בשפה העברית קלה יותר מבעיה זו בשפות זרות אחרות.

אולם, אין בכך כדי להאפיל על הישגיה הגבוהים מאוד יחסית, בכל קנה מידה, של המערכת שפותחה במחקר זה: (1) זוהי מערכת פענוח אוטומטי של ר"ת רב-משמעיים היחידה מסוגה בשפה העברית בכלל ובתחום התורני בעברית בפרט; (2) הושגו תוצאות פענוח גבוהות ע"י שילוב שיטות בסיסיות באמצעות שימוש בשיטת הלמידה J48; (3) מערכת זו השיגה תוצאות טובות ללא השימוש בעקרונות הבנת שפה טבעית ולכן אינה מוגבלת לשפה כלשהי; ו-(4) המערכת בחנה מספר רב של מאמרים (233) ור"ת (165), יותר מכל מערכת קודמת. מספר ר"ת הרב-משמעיים שנחקרו במחקר זה גבוה באופן משמעותי ממספר ר"ת הרב-משמעיים שנחקרו במחקרים האחרים.

מחקר עתידי אפשרי מוצע בנושאים הבאים: (1) מחקר על מאגרים ור"ת נוספים, הן מתחומים תורניים והן מתחומים חדשים בשפה העברית, כגון: חדשות מהאינטרנט, פרוטוקולים של הכנסת, Emails, כהרחבה למחקר הטקסטים התורניים שנעשה; (2) פיתוח שיטות פענוח בסיסיות נוספות; (3) יישום שילוב השיטות הבסיסיות באמצעות שיטות למידה מונחות מוצלחות (Supervised Learning) נוספות; (4) שימוש בשיטות למידה בלתי-מונחות (Unsupervised Learning) בטקסטים בהם הר"ת אינם מפוענחים; (5) מימוש שיטות בסיסיות המושתתות על עיבוד שפה טבעית ובדיקת שילובן לשיפור תוצאות המחקר; (6) יישום כל הרעיונות המחקריים הללו גם עבור פענוח ר"ת רב-משמעיים במסמכים מתחומים מגוונים עבור שפות זרות שונות. צעד ראשון בכיוון זה יכול לכלול שימוש במתודולוגיית פרוייקט מחקר זה על המאגרים הרפואיים שנחקרו במערכות פענוח ר"ת רב-משמעיים אוטומטיים מקבילים ו-(7) פיתוח אלגוריתמים לזיהוי ר"ת משובשים או שגויים, לתיקונם ולהצעת פתרונות עבורם.

נספח: מדגם של ר"ת הרב-משמעיים שנבחנו במאגר על פירושיהם השונים

טבלה 4 מציגה 12 מתוך 165 ר"ת רב-משמעיים שנבחנו במאגר, מספר מופעיהם ופירושיהם השונים שנחקרו בפרוייקט המחקר והיו לבסיס נתונים עבור המערכת לפענוח אוטומטי של ר"ת שפותחה.

מספר הערות:

- כל הפירושים עבור בסיס הנתונים הם הפירושים אשר הופיעו לפחות במופע אחד במאגר הטקסטים שנחקרו. ייתכנו עוד פירושים רבים לחלק מהר"ת ואף פירושים יותר נפוצים מאלו שמוצגים, אך כאמור לפירושים אלו לא היו קיימים מופעים במאגרי הטקסטים שנחקרו.
- במקרים רבים, הפירושים הידועים לא התאימו למופעים של הר"ת ולכן לא היה ברור מהו הפירוש הנכון. הנחת המחקר הייתה כי קיים פירוש נוסף לא ידוע ולכן כתחליף נבחר הר"ת בעצמו כפירוש למופעים אלו. לאופן פעולה זה אין השפעה על המערכת שהרי פירוש זה שונה מהפירושים האחרים הידועים ומסמל פירוש נוסף כלשהו. אין למערכת התייחסות מיוחדת לפירוש כזה ביחס לשאר הפירושים.
- מופעים רבים של הר"ת מופענחים כחלק ממילה כוללת. לדוגמה: הר"ת אח"ז הוא מופע של הר"ת ח"ז. האות 'א' במקרה זה היא חלק מהמילה הראשונה של הפענוח. הוחלט שבמקרה כזה, הפירוש הנכון הוא 'חר זמן' וכאשר המערכת תחליף את הר"ת בפירוש, המילה תהיה מובנת.
- מספור הפירושים המוצג כאן זהה למספור בבסיס הנתונים עבור המערכת שפותחה והוא נבחר באופן שרירותי ולא דווקא עקבי. אין משמעות אמיתית למספור זה, שהרי יצירת השיטות הבסיסיות נעשית עבור כל ר"ת בנפרד וכן השימוש בתאוריית "הפתרון היחיד העקבי בדיון" הוא ביחס לכל ר"ת בנפרד. בנוסף, השימוש בשיטות הלמידה הממוחשבות נעשתה עבור כל ר"ת בנפרד, כלומר במחקר פרטני.

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

שם הר"ת	מספר פירושים במאגר	מספר הופעות במאגר	פירוש 1	פירוש 2	פירוש 3	פירוש 4	פירוש 5	פירוש 6	פירוש 7	פירוש 8
א"א	8	203	אי אפשר	אין אומרים	אשל אברהם	אדוניי אבי	אין אתה	אי אמרת	אין איסור	אם אפשר
בכ"מ	3	16	בכסף משנה	בכמה מקומות	בכל מקום	-	-	-	-	-
בס"ד	2	114	בסייעתא דשמיא	בסעיף ד'	-	-	-	-	-	-
כ"ד	7	145	כמה דוכתי	כתב דברי	כך דברי	כה דברי	כדי דיבור	כאן דבריו	-	-
ע"פ	6	1301	על פי פסח	ערב פסחים	ערבי פניהם	על פני	על פה	-	-	-
ק"ק	2	44	קהילת קודש	קצת קשה	-	-	-	-	-	-
ר"פ	5	54	280 פרשת	ריש פרק	רב פעלים	רבי פלאני	-	-	-	-
שא"כ	3	377	שאין כן	שאם כן	שאינה כשרה	-	-	-	-	-
שא"צ	4	167	שאינו צריך	שאינה צריכה	שאינו צריכות	שאינם צריכים	-	-	-	-
שלי"ה	2	34	שני לוחות הברית	335	-	-	-	-	-	-
שכ"ה	2	35	שכן הוא	325	-	-	-	-	-	-
שי"א	2	29	שיש אומרים	311	-	-	-	-	-	-

טבלה 4: מדגם של ר"ת רב-משמעיים שנבחנו במאגר, מספר מופעיהם ופירושיהם השונים

ביבליוגרפיה

- [1] M. Sanderson, "Word Sense Disambiguation and Information Retrieval," *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag, pp. 142-151, 1994.
- [2] D. Vickrey, L. Biewald, M. Teyssier, D. Koller, Word Sense Disambiguation for Machine Translation. *HLT/EMNLP*, Vancouver 2005.
- [3] Y. Zhang, L. Gong, Y. Wang, "Chinese Word Sense Disambiguation Using HowNet," *Advances in Natural Computation*, Springer Berlin/Heidelberg, pp. 925-932, 2005.
- [4] O. Cihhart, J. Hajic, "Word Sense Disambiguation of Czech Texts," *Text, Speech and Dialogue: Second International Workshop, TSD'99, Plzen, Czech Republic, September 1999 Proceedings*, Springer Berlin/Heidelberg, p. 109, 1999.
- [5] S. Pongpinigpinyo, W. Rivepiboon, "Word Sense Disambiguation of Thai Language with Unsupervised Learning," *Knowledge-Based Intelligent Information and Engineering Systems*, Springer Berlin/Heidelberg, pp. 1275-1283, 2005.

מערכת לומדת המפענחת ראשי-תיבות רב-משמעיים בכתבים תורניים

- [6] D. Yarowsky, "One Sense per Collocation," *Proceedings of the Workshop on Human Language Technology*, pp. 266-271, 1993.
- [7] W. Gale, K. Church, D. Yarowsky, "One Sense Per Discourse," *Proceedings of 4th DARPA Speech in Natural Language Workshop*, pp. 233-237, 1992.
- [8] Z. Yu, Y. Tsuruoka, J. Tsujii, Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using SVM and One Sense Per Discourse Hypothesis. SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics, 2003.
- [9] C.C. Chang, C.J. Lin, LIBSVM: a Library for Support Vector Machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (Downloaded 7.1.07).
- [10] U. S. National Library of Medicine (NLM), <http://medline.cos.com/>. 2003.
- [11] K. Min-Yen, Metadata Extraction and Text Categorization Using Universal Resource Locator Expansions. National University of Singapore Department of Computer Science Technical Report, TR 10/03, 2003.
- [12] R. E. Shapire, Y. Singer, "BoosTexter: A Boosting-Based System for Text Categorization," *Machine Learning*, 39(2/3): pp. 135-168, 2000.
- [13] S. Pakhomov, "Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts," *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 160-167, 2001.
- [14] J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, M. Morrell, A. Rumshisky, *Extraction and Disambiguation of Acronym-Meaning Pairs in Medline*. Unpublished Manuscript, 2001.
- [15] G. Salton, *The SMART Information Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 1971.
- [16] J.A. Rydberg-Cox, "Automatic Disambiguation of Latin Abbreviations in Early Modern Texts for Humanities Digital Libraries," *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 372-373, 2003.
- [17] H. Liu, Y. A. Lussier, C. Friedman, "A Study of Abbreviations in the UMLS," *Proc. AMIA Symp.*, pp. 393-397, 2001.
- [18] H. Liu, A. R. Aronson, C. Friedman, "A Study of Abbreviations in MEDLINE Abstracts," *Proc. AMIA Symp.*, pp. 464-468, 2002.
- [19] E. Adar, S-RAD: A Simple and Robust Abbreviation Dictionary. HP Laboratories Technical Report, September 2002.
- [20] M. Yoshida, K. Fukuda, T. Takagi, Pnad-css: a Workbench for Constructing a Protein Name Abbreviation Dictionary, *Bioinformatics 2000*; 16: pp. 169-75, 2000.
- [21] H. Yu, G. Hripcsak, C. Friedman, "Mapping Abbreviations to Full Forms in Biomedical Articles," *J American Med Inform Assoc.* May-Jun 2002; 9(3): pp. 262-272, 2002.
- [22] Y. HaCohen-Kerner, A. Kass, A. Peretz, "Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents," *Lecture Notes in Artificial Intelligence*, Springer Berlin/Heidelberg, 3230: pp. 58-69, 2004.
- [23] G. A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information*. Harvard University, 1956.

יעקב הכהן-קרנר, אריאל קאס, אריאל פרץ

- [24] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementations*, Morgan Kaufmann, San Francisco 2000.
- [25] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kauffman, 1993.
- [26] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, 27: pp. 379-423 and pp. 623-656, 1948.
- [27] ש' אשכנזי וד"ר ד' ירון, אוצר ראשי תבות, הוצאת קריית ספר בע"מ 1994.
- [28] הרב י' קארו, שולחן ערוך, אורח חיים כרך א', הוצאת מושיב בני הפצת ספרים, הקדמות לפירושים על השו"ע 1975.
- [29] הרב י' באבד, מנחת חינוך, הקדמה של הרב י' בוקסבוים לפירוש מנחת חינוך, עמ' 9, הוצאת מכון ירושלים 1988.
- [30] הרב י' קארו, שולחן ערוך, יורה דעה כרך ג' סימן רפ"ד סעיף ב', הוצאת מושיב בני הפצת ספרים 1975.
- [31] הרב א"י הכהן-קוק, אגרות הראי"ה, חלק ד', הוצאת המכון ע"ש הרצי"ה קוק זצ"ל, ירושלים 1984.
- [32] הרב ח"ח מדיני, שדי חמד קונטרס הכללים, כרך ד' עמ' 6 טור 1, הדפסה מחודשת ע"י הוצאת "בית הסופר", שנת ההוצאה לאור איננה רשומה.
- [33] הרב י"מ הכהן מראדין, משנה ברורה המנוקד, הוצאת מכון דעת יוסף 1995.
- [34] הרב ע' יוסף, יביע אומר, הוצאת המחבר, מהדורה חדשה 1986.
- [35] הרב ע' יוסף, יחוה דעת, הוצאת מכון ירושלים בשיתוף ישיבת פורת ובי"ת המדרש חזון עובדיה 1977.
- [36] י' הכהן-קרנר, ד' מוגהץ, ח' בק, א' יהודאי, "סיווג פסקי שו"ת לפי עדת הפוסק ותקופת החיבור באמצעות מילים", התקבל לפרסום בבר"ד, כתב-עת לתורה ומדע, אונ' בר-אילן.