

מודלים ושיטות ליישום מורפולוגיה עברית במנועי אחזור / אפרים מרגלית

עבודה זו עוסקת בבחינת מודלים שונים לניתוח מורפולוגי עברי במנועי אחזור. בתחילה נסקור את ההשתלשלות ההיסטורית של מנועי האחזור, החל מהשימוש במילות מפתח ועד למנועי אחזור בני ימינו. במסגרת סקירת הספרות ננסה לעמוד על ייחודיות של השפה העברית מבחינת עושר מורפולוגי והאתגרים הטכנולוגיים הנובעים מכך.

כידוע, השפה העברית, כשפות שמיות נוספות, עומדת בראש הפירמידה של המורכבות המורפולוגית. מורכבות זו נובעת ממספר היבטים. בראש ובראשונה מספר הצורות הרב בשפה העברית, בה ישנן בין 70-100 מיליון צורות חוקיות ותקינות.

גם בשפות אחרות, קיים צורך בניתוח מורפולוגי. בשפה האנגלית, המורכבות היא פשוטה יחסית וקיימים מנגנונים פשוטים יחסית לניתוח מורפולוגי. המנגנון עבור אנגלית, נקרא גידום או Stemming והוא מתמודד עם קידומות וסיומות לערכים במילון. לדוגמה, הצורה Misunderstanding מכילה הן קידומת והן סיומת לערך המילוני. באנגלית, הקידומות והסיומות הן קבועות כך שניתן לממש אלגוריתם של Stemming בקלות יחסית. קוד לכתובת Stemming מפורסם באינטרנט ומדובר בעשרות בודדות של שורות קוד.

השפה העברית, שאף היא שפה שמית, היא בעלת מורכבות די רבה ועושר דומה של צורות לשוניות חוקיות. בעבודה, מוצגת עבודתו של קמיר בנושא זה. צורת הטיפול המוצגת על ידו דומה לזו הממומשת עבור השפה העברית.

ריבוי הצורות נובע מכך שלמרות שבעברית יש מספר קטן יחסית של שורשים, כחמשת אלפים, כל אחד מהם יכול להתפתח למספר רב של צורות. ישנם פעלים היוצרים מעל עשרים אלף צורות חוקיות.

אתגר נוסף בתחום זה הוא העמימות הלקסיקלית. רק 40 עד 45 אחוז מהמילים העבריות הן חד משמעיות, לשליש מהמילים יש יותר משתי משמעויות. דהיינו, למחרוזת המופיעה בטקסט, יש בממוצע יותר ממשמעות אחת. דבר המקשה על ניתוח הלשוני באופן ממוחשב.

עד כה עסקנו בניתוח מורפולוגי של מילים תקינות בשפה. חלק ממהפכת המידע העולמית היא היכולת של כל אדם לפרסם מידע באינטרנט. במסגרת מהפכה זו אנו עדים לשינויים בסגנון הכתיבה. לא עוד המפורסמים באינטרנט כתובים כיום בעברית עכשווית.

המשמעות הישירה היא שעל המנגנונים המורפולוגיים של מנועי האחזור בני ימינו, להתמודד גם עם סלנג לצורתיו וגם עם טקסטים עבריים המכילים שגיאות.

בעבודה זו נציג באופן מפורט חמישה מודלים לניתוח מורפולוגי:

הראשון שבהם הוא מודל סטטיסטי, המציע שיטה המשלבת שלוש רמות של ניתוח לשוני ובחירה של הניתוח הסביר ביותר באמצעים סטטיסטיים. המודל הוא פרי עבודתו של סגל. ורמת הזיהוי

שלו בטקסטים תקינים היא גבוהה מאוד.

המודל השני הוא של אורגן, שחידש ויצר מילון משמעות עברי. הרעיון המרכזי הוא שהמילון מכיל מלבד הערכים גם מאפיינים סמנטיים. כך ניתן לממש גם חוקיות של בדיקת משמעות סמנטית ברמת המשפט ולהפיג חלק ניכר מהעמימות הלקסיקלית. לדוגמא, לצירוף המילים "הרכבת לימונים" יכולים להיות מספר ניתוחים לשוניים חוקיים. אך הניתוח הנכון של הרכבת ענף בעץ לימונים נעשה בעזרת המילון המושגי המכיל את התכונה "צומח" בשני הערכים.

המודל השלישי הוא מודל יוריסטי, פרי עבודתו של פנקס, המיישם מעין חוקה של דקדוק העברי. מודל זה פועל ללא כל מילון והאלגוריתם מממש את ניתוח הלשוני רק על החוקה.

המודל הרביעי הוא מודל מילון טהור, של שויקה, המיישם יכולת ניתוח על בסיס מילון מלא של השפה העברית הכולל מידע לגבי ההרחבות החוקיות של כל ערך.

המודל החמישי הוא מודל המשלב יוריסטיקה ומילון. מדובר בשיטתם של כרמל ומארק המיישמת מילון עברי חלקי עם יכולת ניתוח יוריסטית.

בפרק המתודולוגיה, נגדיר קריטריונים לבחינת המודלים השונים במספר מימדים. החל מניתוח מילה בודדת, עבור דרך ניתוח משפטים שלם ולבסוף בחינה של צורת הניתוח של השפה העברית המדוברת.

הבדיקה ההשוואתית נעשתה על בסיס הקריטריונים שהוגדרו. התוצאות הן שקיימים הבדלים בין השיטות השונות עבור כל אחד מסוגי הניתוח (טקסטים תקינים, סלנג וטקסט משובש). לאלגוריתמים המבוססים על מילונים (המודל הסטטיסטי, המודל המושגי והמודל המילוני) יש יתרון בטקסטים תקינים ואילו לאלגוריתמים היוריסטים יש יתרון בטקסטים המיכלים שגיאות וסלנג.

בפרק המסקנות ניסנו להציע שיפורים קלים באלגוריתמים הקיימים. עם זאת, נראה שרמת הדיוק של המודלים השונים היא גבוהה יחסית, כך שהשיפורים, אם ייושמו, לא יגרמו לפריצת דרך. לכן, נחתום את העבודה בהצעה של שיטה חדשה.

השיטה המוצעת על ידנו היא הוספת מנגנון שיבצע ניתוח מקדים לטקסטים. ניתוח זה יזהה את "אופיו" של הטקסט המנותח. על פי האופי, יקבע המנגנון את האלגוריתם המתאים ביותר לניתוח הטקסט ואל מנגנון זה הטקסט ישלח לניתוח.

שיטה זו תאפשר, להערכתנו, ניתח מורפולוגי מדויק יותר למגוון רחב של סוגי טקסט.

מספר מיון : 025.04 מרג.מו תשס"ח

מספר מערכת : 1153573