פיתוח מדדים לאיתור חיבורים משורבטים (חיבורי ג'יבריש) ותיקוף המדדים מול חיבורי "פסיכורשת"

יעל שפרן, ענת בן-סימון

פרויקט השפה העברית (HLP)

תמצית

מאז הטמעת החיבור בבחינה הפסיכומטרית ושילובו ב'פסיכורשת' (בחינה פסיכומטרית ממוחשבת ואדפטיבית להתנסות עצמית) נאספו במסגרת זו מעל 5,000 חיבורים. אוסף חיבורים זה כולל מגוון רחב של חיבורים שחלקם משקפים ניסיון לכתיבה מיטבית וחלקם נראים כשרבוט אקראי של מילים ומשפטים. קורפוס חיבורים זה סיפק הזדמנות לפיתוח מדדים לאיתור טקסטים שנראה בעליל שנכתבו כלאחר יד ואינם עומדים בשום קריטריון של כתיבה תקנית או חיבורים שאינם לנושא, להלן, חיבורי ג'יבריש. מטרתו של המחקר הנוכחי הייתה לפתח מדדים תקפים לאיתור חיבורי ג'יבריש.

לצורך זיהוי ממוחשב של חיבורי ג'יבריש גובשו עשרה קריטריונים (מדדים) עבור חיבורים בשפה העברית ושמונה קריטריונים (מדדים) עבור חיבורים בשפה הערבית ובוצעו ניתוחים שונים לחקירת הקריטריונים ומאפייניהם ולבדיקת יעילותם בזיהוי חיבורי ג'יבריש. כל המדדים מבוססים על מאפייני טקסט כמותיים המופקים באופן ממוחשב מהטקסט הכתוב.

לצורך בדיקת הרגישות של המדדים לעיל בזיהוי חיבורי ג'יבריש פותח אלגוריתם המפיק מדד ג'יבריש לכל חיבור ומסווג אותו כחיבור תקין או כחיבור ג'יבריש. דיוק הזיהוי נבדק על 4,108 חיבורים בשפה העברית ו-1,088 חיבורים בשפה הערבית שנכתבו במסגרת פסיכורשת, תוך שימוש בסיווג של מעריך אנושי כקריטריון. הבדיקה יושמה רק עבור חיבורים שאורכם עולה על 100 מילה.

התוצאות מלמדות כי רק 66.2% מ- 4,108 החיבורים שנכתבו בשפה העברית ו- 60.4% מ-1088 החיבורים שנכתבו בשפה הערבית נמצאו על ידי המערכת הממוחשבת כחיבורים תקינים. מרביתם של החיבורים הנותרים היו קצרים מהנדרש (28% ו- 34.3% בהתאמה), ומיעוטם (6.1% ו- 5.3% בהתאמה) חיבורי ג'יבריש.

בבדיקת דיוק הסיווג של החיבורים כחיבורי ג'יבריש או חיבורים תקינים נמצא כי 99.4% מהחיבורים בשפה העברית ו-96.7% מהחיבורים בשפה הערבית שאורכם עמד בדרישות סווגו באופן מדויק כחיבורי ג'יבריש או חיבורים תקינים.

בהתייחס לטעויות הסיווג נמצא כי בהתייחס לחיבורים שפה העברית אף חיבור תקין לא סווג כחיבור ג'יבריש ורק 0.4% מכלל החיבורים שסווגו על ידי האלגוריתם כחיבורים תקינים הם למעשה חיבורי ג'יבריש. בהתייחס לחיבורים בשפה הערבית נמצא כי 2% מהחיבורים שסווגו על ידי המחשב כחיבורי ג'יבריש הם למעשה חיבורים תקינים ו- 4% מהחיבורים שסווגו על ידי המחשב כחיבורים תקינים הם למעשה חיבורי ג'יבריש.

תוצאות המחקר מלמדים כי המדד הכולל לזיהוי חיבורי ג'יבריש מדויק דיו וניתן להשתמש בו בגרסה התפעולית של פסיכורשת ובהקשרים נוספים לפי הצורך.