

# בעיות ואתגרים

## סער יהלום\*



לשנות חישובית היא תחום מדעי שמשלב את מדעי המחשב עם בלשנות. התחום עוסק בתוכנות שמסוגלות להבין שפות אנוש, ומתבטא בתוכנות שמתרגמות טקסטים שלמים ו/או מסוגלות להבין פקודות קוליות ולבצען.

תוכנות תרגום של טקסטים שלמים (כמו גוגל תרגום) אינן "מבינות" את הטקסט, אלא רק ממירות מילים משפה לשפה וממקמות אותן במשפט. האתגר הוא להמיר כל מילה אל המקבילה הנכונה שלה בהתאם למשפט, ולמקם אותה במקום הנכון במשפט לאחר המרת כל המילים.

לדוגמה, אם ניתן לתוכנה לתרגם את המשפט 'הכלב של דנה לאנגלית, הוא יתרגם את הידיעה ל-the, 'כלבי למילה 'dog' וישלי למילה 'of'. את המילה 'דנה' הוא לא יזהה ויניח שזה שם, ולכן יתרגם את ההברות של 'דנה' לצורה פונטית באנגלית- 'Dana', לפי מאגר הנתונים שלו. לאחר תרגום שלוש המילים לאנגלית, הוא יסדר אותן לפי הסדר הנכון באמצעות שימוש בחוקי הדקדוק ויתרגם את זה ל-The dog of Dana, או, בתוכנה קצת יותר מתוחכמת, ל-Dana's dog. ובכן, הוא אמנם תרגם את המילים נכון וסידר אותן בצורה הנכונה, אבל הוא אינו יודע מה משמעות המילה 'כלבי', או איזה שיוך מייצגת המילה 'שלי'. תוכנות שקולטות פקודות קוליות, צריכות יותר מזה. הן צריכות להבין את משמעות כל המילים ויילהבין את משמעותן, כבודדות וכמשפט, כדי להיות מסוגלות לעבד את המשפט לפעולה שעל המחשב לבצע.

## עיבוד פקודות קוליות

קיים בלבול עקב קיום תוכנות שכביכול מסוגלות לעבד פקודות קוליות. מתייחסים אליהן כאל תוכנות שמעבדות פקודות קוליות מהסוג שהבלשנות החישובית שואפת אליו, אך בעצם הן עובדות על משפטים מוגדרים מראש המזוהים לפי רצף צלילים.

לדוגמה, אם נותנים למחשב את הפקודה הקולית "פתח דפדפן", הוא לא מבין את משמעות המילים. הוא מזהה את רצף הצלילים במשפט ויודע שהוא צריך לפתוח את הדפדפן.

הבעיה במשפטים שהוגדרו מראש היא בכמות המוגבלת האפשרית של פקודות והצורך לדיוק במילים. התוכנות שהבלשנות החישובית שואפת אליהן הן תוכנות שיהיו מסוגלות להבין את המשמעות של המילים עצמן ואת הקשר ביניהן, וכך אפשר יהיה להשתמש בכמות אינסופית של פקודות, בניסוח שבחורים, ובלי צורך לזכור בעל-פה ובמדויק את הפקודות שהוגדרו.

אחד הדברים שתוכנות התרגום מסתמכות עליו הוא הדמיון בין המבנים הדקדוקיים בכל השפות: פעלים, שמות עצם, תארים

ומבנים מקבילים של משפטים.

דבר נוסף שמקל על הבלשנות החישובית הוא משפחות של שפות. כלומר, קבוצת שפות שהתפתחו במקומות קרובים גיאוגרפית או שהייתה ביניהם אינטראקציה רבה והן מאופיינות במילים ובמבנים דקדוקיים דומים. כזו היא למשל קבוצת השפות השמיות או הלטיניות. אך יש גם השפעות בין-משפחתיות. היום, למשל, רוב השפות האיחפאיות הן שילוב של לטינית ושפה של שבט ברברי שכבש חלק מהאימפריה הרומית. גם הלטינית עצמה התפתחה מהשפה היוונית, שהייתה שפה שמית.

## סוגיות הקשר והמשמעות

**הבעיה המרכזית** בכלשנות חישובית היא בעיית ההקשר: בשפות אנוש יש פעמים רבות מילים בעלות מספר משמעויות, שהקשר המשפט קובע איזו מהן תשמש באותו משפט. יש סיפור ישן המדגים את הבעיה, אך כי כיום תוכנות התרגום קצת יותר מתוחכמות: נתנו לתוכנת תרגום לתרגם מאנגלית לרוסית משפט מהברית החדשה "The spirit is willing but the flesh is weak" ("הרוח חפצה אך הבשר חלש"). כדי לבדוק את התוצאה ביקשו מהתוכנה לתרגם את התוצאה מרוסית בחזרה לאנגלית, התוצאה הייתה "The vodka is good but the meat is rotten" ("הוודקה טובה אבל הבשר רקוב").

**בעיה שנייה** המפריעה לתוכנות התרגום, היא במילים הייחודיות. אנשים רבים מחשיבים את הפועל BE (am, is, was וכיו) לפועל ייחודי, אך הוא איננו כזה. פועל זה ניתן למצוא בשפות רבות, אך בגלל ההתפתחות המוקדמת של העברית והקרבה שלה לערבית ולארמית אין בה את הפועל "BE". למרות זאת, בשפות מסוימות ישנו סוג של מילים ייחודיות שאינן מפריעות לפעולת התרגום.

**בעיה שלישית** שכדאי להתייחס אליה הוא טון הדיבור: כדי לעבד שפה בצורה של קול ולא של טקסט נדרש המחשב לנתח את טון הדיבור ואת שינוי הטון. דבר זה יכול להוות בעיה בקליטת משמעות, אבל לדעתי הוא יותר עוזר לכניית תוכנה לעיבוד דיבור משהוא מפריע: השפה היא בעלת חוקים 'מוזרים' שקשה להגדיר אותם באופן מתמטי. דבר זה הגיוני לחלוטין משום שהמוח שלנו לא עובד כמו מחשב, והוא מפתח שפה שמתאימה לדפוסי הפעולה שלו. בניגוד לכך, שינויי הטונים שלנו בעת דיבור עובדים בדפוסים מתמטיים. לכן אפשר לבנות די בקלות תוכנה שמשמשת בגובה הצליל ובמאפיינים נוספים שלו.

כאשר התוכנה מזהה את תבנית הטונים, היא משווה את המשמעות הכללית של המשפט שנרמזת מאותה תבנית ומשווה אותה עם המשמעויות שנרמזות מהצלכות כל המשמעויות השונות של מילים במשפט, וכך היא מסוגלת למצוא את ההקשר הנכון.

**בעיה נוספת**, פחות משמעותית היא הסלנג. אמנם אפשר

# בבלשנות החישובית



מכירה אך אינה מבינה את הקשר שלה למשפט, היא תנסה להחליף את המילה בכל הייקרובים שלה שאותם תזהה בעזרת הקטגוריות (שורש, הטיה, גוף וכיו.). המשתמש יוכל להגדיר שהמחשב יתקן אותו, או שפשוט ימצא את המילה למטרות הבנה. אם המחשב לא מצא את המילה המתאימה, הוא ישאל את המשתמש.

**לסיכום**, בלשנות חישובית היא תחום שעוסק בתוכנות שמסוגלות לעבד שפה טבעית (שפה שמשתמשת כשפת אם ושמשתמשת לדיבור בין בני אדם) בצורה גולמית שלה.

על אף כמה נקודות תמיכה, סובל התחום ממספר רב של בעיות. מיטב החוקרים עובדים על הבעיות הללו, ולדעתי אנו יכולים לצפות עד סוף המאה הנוכחית לפריצות דרך חשובות בתחום, ולפיתוח תוכנות מתוחכמות שמסוגלות להתקרב מאוד להשיג את מטרת הבלשנות החישובית.

תודה לאירית לוי שהדריכה אותי בכתיבת המאמר.

\* סער יהלום - תלמיד במרכז המחוברים חופיים בעמק חפה, בהנהלת אהובה פאר

לפתור את רובה של הבעיה ביצירת מאגר נתונים של סלנגים שונים, וכשהמחשב אינו מזהה מילה או צירוף מילים כלשהו הוא יכול להשוות עם מאגר הנתונים. אך הבעיה אינה נפתרת מכיוון שהסלנג הוא לא כמו השפה עצמה. כל יום מומצאות מילים חדשות וביטויי סלנג חדשים. לעקוב אחרי כל זה יהיה בלתי אפשרי. הפתרון הוא ליצור מערכת שתשווה את המילה עם מאגר הנתונים שלה ותשתמש בידע שלה על תופעות פיתוח סלנג שיוגדרו לתוך התוכנה בעזרת בלשנים. אבל אינני חושב שזה הפתרון הנכון. לדעתי ישנו פתרון פשוט בהרבה מהפתרון הבלשני, וזהו הפתרון שרוכ בני-האדם משתמשים בו כשהם נתקלים בבעיה כזו: ברגע שהתוכנה תזהה מילה או צירוף מילים שאינם מתאימים לשאר המשפט, היא תשאל את המשתמש למה כוונתו במילים אלו.

**הבעיה האחרונה** שברצוני להתייחס אליה היא בעיית הטעויות אנשים אינם מדברים בצורה מושלמת, הם עושים טעויות דקדוקיות שיכולות לפגום בפעולת המחשב. המחשב תמיד יכול לבקש מהאנשים לחזור על עצמם, אבל לדעתי זה יהיה פגם משמעותי בתוכנה. אני חושב שאפשר ליצור מערכת של זיהוי שורש, הטיית, גופים וכיו, וברגע שהתוכנה מוצאת מילה שהיא