דבר אל המסך

איך להבין אם המחשב מבין שפה

ישראל בנימיני

איורים: Cd Banc/clipart.com

טקסט בלי הבנה

חלק גדול מאד מעבודה עם מחשבים מתרכז ביצירת ובעריכת טקסט, בשיתוף טקסט עם אנשים אחרים, בקריאת טקסט ובחיפוש טקסט. בכל השימושים האלה, המחשב הוא פסיבי כמעט לחלוטין: הוא אינו מבין את המילים והמשפטים שבהם עיקר עיסוקו. מבחינה זו אין המחשב שונה מטכנולוגיות קודמות, החל מלוחות החרס שעליהם כתבו השומרים לפני יותר מ-5000 שנה. השומרים לא ציפו מלוחות החרס שיבינו את שנכתב עליהם, כפי שמשתמשי מכונות הכתיבה של המאה ה-19 לא דרשו כי המכונה תגיב למילים שהוקלדו עליה. הופעת המחשב שינתה זאת: הביטויים "מוחות אלקטרוניים" ו-"מכונות חושבות" שליוו את המחשבים בשנותיהם הראשונות (ושמהדהדים עד היום בגזירת המילה "מחשב" מהשורש "חשב" בעברית) יצרו ציפייה לדו-שיח בין אדם ומכונה ושל שיתוף פעולה ביצירת טקסטים, בקריאתם ובניתוחם. בעשרים השנים האחרונות נוצר אפילו הרושם כי מופיעים ניצנים ראשונים של הבנה: לדוגמה, מעבד תמלילים מודרני מתקן בצורה אוטומטית שגיאות איות שכיחות, כאילו הבין מה התכוונו לכתוב; וכאשר מחפשים באינטרנט את צורת המילה ברבים, מנוע החיפוש עשוי "להבין" כי גם צורת המילה ביחיד מתאימה לאותו חיפוש.

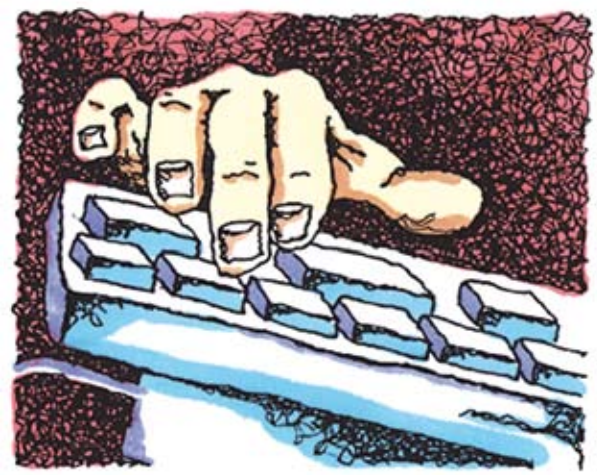
דוגמאות אלו ממחישות שני צדדים של השאיפה להעניק למחשב יכולות גבוהות יותר להבנת "שפה טבעית" (המונח המקובל בעולם הבינה המלאכותית עבור שפות אנושיות, בניגוד לשפת מתכנתים). מצד אחד, אם המחשב מסוגל לכך, ואם יכולות אלה עוזרות לנו, מדוע שלא נדרוש עוד רמות של הבנה? מצד שני, אף על פי שברור לנו בצורה אינטואיטיבית כי קיים הבדל גדול בין תיקון שגיאות כתיב לבין הבנה אמיתית, קשה להפוך אינטואיציה זו להגדרות מדויקות יותר של מאפייני הפער, ולכן קשה להסיק מכך מהן הרמות הבאות שכדאי לשאוף אליהן. צד שלישי לאותה שאיפה קשור כמובן לחלום המקורי של תחום הבינה המלאכותית: למרות העדויות המצטברות לקיום מאפיינים מסוימים של שפה אצל בעלי חיים, השימוש בשפה נשאר אחד מהמאפיינים החשובים ביותר של האנושות, וקשה לנו מאוד לדמיין יצור תבוני שאינו משתמש בשפה. לכן כל מאמץ ליצור מכונה הראויה להגדרה "אינטליגנטית" חייב לפחות לשקול את המשימה של שימוש בשפה טבעית, ולו רק כדי לאפשר לנו, בני האדם, לשכנע את עצמנו בבינתה של המכונה.

שימושים להבנה חישובית

נניח שמישהו היה מציג בפנינו מכונה וטוען כי היא מבינה עברית (כמובן שבאותה מידה היינו יכולים לצפות מהמכונה עצמה להזמין אותנו לשיחה...). איך היינו בודקים טענה זו? תחום הבינה המלאכותית הציע לכך כמה וכמה תשובות, החל מ"מבחן טיורינג" המפורסם, שהוגדר על-ידי אלן טיורינג (Turing) ב- 1950. גרסה אחת של מבחן זה דורשת כי שופט אנושי המנהל שיחות כתובות עם אדם ועם מכונה לא יוכל להחליט מיהו האדם ומיהי המכונה (וראו "איך להתחזות לאדם", "גלילאו" 75). מבחן זה, המשתמש בשפה טבעית רק ככלי עזר כדי לבחון את השאלה הכללית של האם מכונה היא אינטליגנטית, קשה מדי עבור היכולות הנוכחיות של הבינה המלאכותית. אילו מבחנים קלים יותר אפשר להציע כדי לבדוק הבנה? בנקודה זו, ירצו אולי חלק מהקוראים לחשוב כיצד היו בודקים אם אדם כלשהו הבין טקסט שהוצג לו.

אחת התשובות הפוריות ביותר לשאלה זו הוצעה על-ידי ד"ר עידו דגן (Dagan), ראש המעבדה לעיבוד שפה טבעית במחלקה למדעי המחשב באוניברסיטת בר-אילן. לפני שנציג את תשובתו, מעניין לבחון מהם השימושים הפרקטיים של "הבנה חישובית". הדוגמה המקובלת ביותר היא השימוש

במנועי חיפוש, כגון Google ו- Yahoo. כיום היכולת של הבנת המשמעות מצד מנועים אלה מוגבלת ביותר. הם מתייחסים לשאילתת המשתמש כאל אוסף מילים, ומאתרים מסמכים המכילים מילים אלו. המנוע אינו מתייחס למעשה למגוון משמעויות המילים ולקשרי המשמעות ביניהן. לדוגמה, שאילתא על "זיהום מים" תחמיץ כיום מסמך המכיל את המשפט "נפט דלף ממיכלית בים התיכון", מכיוון שהמסמך אינו מכיל בדיוק את שתי המילים הללו, אף על פי שהוא מדבר על דליפת נפט (שמשמעותה "זיהום") בים התיכון (שמשמעותו מאגר של "מים"). לעומת זאת, המנוע עשוי לאחזר מסמך שבו מופיעות המילים "זיהום" ו"מים" גם אם



אינן קשורות אחת לשנייה בטקסט באופן המבוקש בשאילתא (למשל, מסמך המסביר כיצד שטיפה במים עוזרת למנוע זיהום פצעים). דוגמאות נוספות לצורך בהבנת משמעות טקסטים הן מענה אוטומטי על שאלות מורכבות (כגון "אילו מדינות מייצרות מכוניות?"), תמצות אוטומטי של מסמכים (למשל תמצות המידע בכתבות שהתפרסמו ביום מסוים על המלחמה בגיאורגיה), תרגום אוטומטי בין שפות טבעיות, או בדיקה אוטומטית של תשובות תלמידים במבחנים על-ידי השוואת משמעותן ל"פתרון בית הספר" הנתון לשאלה.

לא קשה לדמיין כיצד יובילו מחשבים בעלי יכולות כאלה למהפכה ביכולתנו לאתר מידע רלוונטי, מהפכה הניתנת להשוואה לזו שהתחוללה עם הופעת מנועי החיפוש באינטרנט. מצד שני, ניתן לקוות כי השגת יכולות הבנה כאלה היא אתגר בדיוק בגודל הנכון: קשה מספיק כדי להיות מעניין, אבל קל יותר מהגשמת כל החלום של חלוצי הבינה המלאכותית - מכונה בעלת בינה "אמיתית". כדי להימנע מיצירת ציפיות לבינה אמיתית בהקשר זה, החוקרים אינם משתמשים במונח "הבנה" אלא במונח "הבנה חישובית".

איך לבחון הבנה חישובית

בעולם הבינה המלאכותית, הגישה הקלאסית להבנה חישובית של טקסט דורשת ייצוג של הטקסט המקורי על-ידי רשת של סמלים. לשם כך, יש ליצור אוסף גדול של סמלים. חלק מהסמלים ייצגו שמות עצם מוחשיים או מופשטים כמו "ספל", "אדם", "ג'ורג' וושינגטון", או "אושר". סמלים אחרים ייצגו קשרים בין סמלים, כך שנוכל לייצג את הטענות "ג'ורג' וושינגטון הוא אדם" או "ספל משמש לשתיית נוזלים", וכו'.

כאשר קיים עולם מספיק עשיר של סמלים, אפשר להגדיר הבנה חישובית על-ידי ייצוג נכון של הטקסט בצורה של זיהוי הסמלים שאליהם מתייחס המשפט וייצוג נכון של הטענה שמביע המשפט

כאשר קיים עולם מספיק עשיר של סמלים, אפשר להגדיר הבנה חישובית על-ידי ייצוג נכון של הטקסט בצורה של זיהוי הסמלים שאליהם מתייחס המשפט וייצוג נכון של הטענה שמביע המשפט. ייצוגי הטקסט משתמשים בכלים פורמליים מתחום הלוגיקה כדי לנסח בשפה מתמטית את סוגי הסמלים (למשל במשפט "ספל משמש לשתיית נוזלים", "ספל" ו-"משמש" הם סמלים מסוג שונה), ואת היחסים בין הסמלים המופיעים במשפט. תוכנה העומדת בדרישות אלה תייצג בצורה דומה את המשפטים "יוסי זרק את הכדור" ו-"הכדור נזרק על-ידי יוסי". כאשר נבקש מהתוכנה למצוא בטקסטים העומדים לרשותה תיאורים של סיטואציות של משחק, היא תעקוב אחר הקשרים המובילים בין הסמלים "כדור" ו-"זריקה" אל הסמל "משחק" ותסיק, בסבירות גבוהה, כי משפט זה מתאים לחיפוש.

שאיפה זו שואבת את רעיונותיה מעולם הבלשנות התיאורטית, כמו גם מאחד הענפים הרעיוניים שליוו את

הבינה המלאכותית מתחילת דרכה. למעשה, הרעיון של הגדרת משמעות השפה על-ידי רשת של סמלים קשור לתפישה האינטואיטיבית שלפיה קיימת "שפה של חשיבה" – שפה שבה משתמש המוח, ושאליה מתורגם כל היגד טקסטואלי כדי שיהיה ניתן לעיבוד. שאיפה זו הובילה לפרויקטים ארוכי-שנים כמו Cyc (קישור בסוף המאמר; וראו גם "כמה צריך לדעת", "גלילאו" 62). עם הזמן התברר כי קשה מאד ליצור עולם עשיר דיו של סמלים, עם סדר והיגיון פנימי ועם יכולת לגדול ולכלול עוד ועוד מושגים. רבים בקהילת הבינה המלאכותית אינם מאמינים בסיכויי ההצלחה של רעיונות כאלה, אם מסיבות מעשיות של היקף הקושי, ואם מסיבות תיאורטיות, לפיהן החשיבה אינה ניתנת להסבר על-ידי פעולות לוגיות על רשתות של סמלים (ומכאן שההנחה של "שפת החשיבה" – Language of Thought Hypothesis – אינה נכונה).

ד"ר דגן חיפש דרך אחרת להגדיר הבנה חישובית, מתוך גישה מעשית שלא תהיה תלויה בתיאוריה זו או אחרת על מנגנוני ייצוג משמעות. אחד ממקורות ההשראה שלו היה תרגיל בהבנת הנקרא בספר ללימוד אנגלית ששימש את בנו בבית הספר. בין השאר, הופיע בספר המשפט (באנגלית) "משולש ברמודה משתרע באוקיינוס האטלנטי, ליד חופי פלורידה". התלמידים נדרשו להחליט, על בסיס טקסט זה, האם נכון המשפט "משולש ברמודה נמצא ליד ארצות הברית".

נחזור לשאלה "איך נוכל לבדוק אם אדם כלשהו הבין טקסט שהוצג לו?". לפי גישת רשת הסמלים, עלינו לבדוק בזיכרונו של המחשב את הייצוג הסימבולי שיצר המחשב מתוך המשפט שאותו התבקש להבין. גישה זו אינה אפשרית כשמדובר בבני אדם: אפילו אם קיימת רשת סמלים כזו אצל בני אדם (שאלה שהתשובה עליה רחוקה מלהיות ברורה), אין ברשותנו מכשירים המסוגלים לקרוא את מה שמיוצג בה. כפי שאפשר לראות בדוגמה של התרגיל באנגלית, עבור בני-אדם כולנו משתמשים בשיטה פשוטה הרבה יותר: אנו מבקשים מאותו אדם לנסח את הטקסט במילים אחרות, או שואלים שאלות שהתשובה עליהן מחייבת מה שמקובל כהבנה נכונה של הטקסט. במילים אחרות: מבחן אמפירי של הבנת טקסט הוא על-ידי בדיקת נכונות טקסט אחר שהפיק הנשאל.

אנלוגיה זו שימשה כמוטיבציה המרכזית לטענה כי קיים מכנה משותף ליישומים רבים של הבנת טקסט, וביניהם היישומים שתוארו בתחילת המאמר: אם תהיה ברשותנו שיטה כללית לענות על השאלה "האם אפשר להסיק מטקסט <A> את טקסט <B>?", נוכל לבסס עליה פתרונות למגוון רחב של שאלות אחרות. לדוגמה, חיפוש טקסטים המדברים על "זיהום מים" יבוצע על-ידי בדיקה עבור כל טקסט האם אפשר להסיק ממנו על זיהום מים, אפילו כאשר המילים "זיהום" ו-"מים" אינם מופיעים כלל בטקסט (וגם להפך - אפילו אם טקסט מכיל מילים אלה, יתכן כי הבדיקה תעלה שאי אפשר להסיק כי הוא עוסק בזיהום מים). תמצות של מידע על אירועים בחדשות יסרוק מאמרים ממקורות שונים, ויבדוק האם כל מאמר תורם מידע חדש, או שהוא בר הסקה מטקסט שכבר נכלל בתמצות. כאשר תרגום אוטומטי מאנגלית לעברית נתקל במשפט שקשה לתרגמו ישירות, הוא

יוכל להיעזר במציאת משפט אחר באנגלית שנובע מהמשפט המקורי, ושאותו קל יותר לתרגם. דוגמאות אלה באות להמחיש כי להסקה טקסטואלית יש תפקיד מרכזי ביישומים רבים של עיבוד שפה טבעית והבנתה.

לרעיון זה קרא דגן "גרירה טקסטואלית" (Textual Entailment). הרעיון פורסם לראשונה במאמר שפרסם דגן יחד עם תלמידו, אורן גליקמן (Glickman), בשנת 2004. זו היתה אחת מהתרומות של המעבדה לעיבוד שפה טבעית באוניברסיטת בר-אילן לרשת מחקר אירופית בשם PASCAL (קישור בסוף המאמר), המקשרת בין חברות מסחריות

וקבוצות מחקר בתחומים של זיהוי תבניות ולמידה חישובית. בסדנה הראשונה של PASCAL הציג גליקמן את רעיון הגרירה הטקסטואלית. לאחר הצגה זו פנו מארגני הסדנה לגליקמן והציעו לו להפוך את הבעיה של מידול חישובי של גרירה טקסטואלית לאתגר שיועמד בפני קבוצות המחקר (דוגמאות למרכיבי האתגר יתוארו בהמשך). דגן ותלמידיו קיבלו עליהם את המשימה, והיו אחראים לארגון שלוש תחרויות מדעיות כאלו במסגרת PASCAL. קבוצות מחקר רבות התעניינו בגרירה הטקסטואלית, פרסמו מאמרים בנושא ופיתחו גישות ורעיונות משלהן. העניין האקדמי בנושא הלך וגבר, ומשך גם את תשומת ליבם של מנהלי NIST (National Institute of Standards and Technology) המכון הלאומי האמריקאי לסטנדרטים וטכנולוגיות, שקיבל עליו את ארגון התחרות הרביעית. אחד מתפקידיו של המכון הוא לזהות תחומי מחקר מבטיחים ולפעול כדי לקדם אותם. במדור זה (וראו מדור בינה, שבו הוזכרו בעבר תחרויות שארגן NIST בתחום זיהוי פנים, "האח הגדול מזהה אותך", "גליליאו" 108). השנה כולל NIST במסלול המחקר של ניתוח טקסט (קישור בסוף המאמר) שלוש תחרויות, שתוצאותיהן יוכרזו לקראת סוף 2008. התחרויות הן בתחומים של מענה על שאלות, סיכום טקסט, וזיהוי של גרירה טקסטואלית, ד"ר דגן הוא חבר בוועדה המייעצת של מסלול זה. זוהי רק אחת העדויות למהירות המרשימה שבה אימצה קהילת המחקר את הרעיון. נכון להיום, ארבע שנים לאחר הצגת הרעיון של הגרירה הטקסטואלית, הוא הפך לאחד מתחומי המחקר החמים בתחום עיבוד שפה על-ידי מחשב ויצר מומנטום לקידום המחקר בתחום הקשה של הבנת משמעות טקסטים.

טקסט גורר טקסט

קבוצות המחקר הניגשות לתחרות נדרשות לפתח תוכנה שתוכל לענות על שאלות הבוחנות הצלחה בזיהוי של גרירה טקסטואלית. כל שאלה מורכבת מטקסט נתון וממשפט קצר שיש לבדוק האם הוא נובע מהטקסט. לדוגמה, אם הטקסט הנתון הוא "(1) התנגשות של אוטובוס ומשאית באוגנדה הביאה למותם של לפחות שלושים אנשים והותירה עוד 21 אנשים פצועים", ומשפט הבוחן הוא "תאונה באוגנדה גבתה שלושים קורבנות", על התוכנה להסיק כי המשפט נובע ("נגרר") מהטקסט הנתון. לעומת זאת, על התוכנה להשיב בשלילה לשאלה האם הטקסט "(2) דימום במערכת העיכול עשוי להופיע כתופעת לוואי של תרופות כמו אספירין ואיבופרופן" גורר את הטענה "אספירין מונע דימום במערכת העיכול". "(3) הנשיא רייגן נכח בטקס בוושינגטון לציון הפלישה לנורמנדי" אינו גורר את המסקנה "וושינגטון נמצאת בנורמנדי", בעוד שהטקסט "(4) הצעת החוק זכתה ב-21% תמיכה במשאל העם, בעוד שהצעת החוק של מפלגות האופוזיציה קיבלה 40% תמיכה" גורר את המסקנה "הצעת החוק הובסה על-ידי האופוזיציה".

דוגמאות אלה, המובאות בשינויים קלים (ובתרגום חופשי מאנגלית לעברית) מתוך תחרויות קודמות בגרירת טקסט, ממחישות את המגוון הרחב של היכולות הנדרשות מתוכנה המוגשת לתחרות. יכולות אלה כוללות זיהוי של משמעויות דומות ("התנגשות" מול "תאונה" בדוגמה 1), סיבה ותוצאה ("התנגשות גרמה למוות" היא בעלת משמעות דומה ל- "תאונה גבתה קורבנות" בדוגמה 1, אבל "עשוי להופיע כתוצאה" הוא ההפך מ- "נמנע על-ידי" בדוגמה 2), סידור מחדש של משפט ("דימום נמנע על-ידי אספירין" מול "אספירין מונע דימום"

אנו זקוקים לניתוחים של יחסי גרירה כדי להעשיר את המאגרים של יחסים לשוניים ושל מידע על העולם, שבלעדיהם לא תיתכן הבנת טקסט. מצד שני, כדי לבצע ניתוחים אלה כשלעצמם אנו זקוקים ליכולות בסיסיות של עיבוד טקסט

בדוגמה 2), יחסים מספריים (40% שבהם זכתה האופוזיציה הם יותר מ- 21% שבהם זכתה הקואליציה, ולכן אפשר להסיק בדוגמה 4 כי הקואליציה הובסה), ציוני מקום (מתוך "הנשיא רייגן נכח בטקס באולם בית הנבחרים בוושינגטון" אפשר להסיק כי אולם בית הנבחרים נמצא בוושינגטון, אך לא כך בדוגמה 3) ועוד.

הדוגמאות גם מראות כי יחס הגרירה אינו סימטרי: למשל, אפשר להסיק מדוגמה 3 את הטענה "הנשיא רייגן היה בוושינגטון", אבל אם היה ידוע לנו רק כי הוא היה שם, לא היינו יכולים להסיק כי הוא נכח בטקס. זהו יתרון של משימת הגרירה בהשוואה למשימות סימטריות באופיין, כמו תרגום או ניסוח מחדש (paraphrasing – "אמירה במילים אחרות"): מנקודת המבט של המפתחים, המשימה מאפשרת לתוכנה לחפש בתוך טקסט אותן מסקנות שביכולתה להסיק

ולהתעלם מהאחרות. מנקודת המבט של המשתמשים, הרבה מהצרכים שעבורם אנו מעוניינים בהבנת טקסט דורשים למעשה יחס א-סימטרי כזה – למשל, תמצות טקסט מכמה מקורות, או החלטה האם דף אינטרנט מסוים עוסק בנושא שאנו מחפשים.

יש גם לשים לב כי בעולם האמיתי רוב ההסקות אינן וודאיות: אם ידוע לנו שיוסי הוא ציפור, סביר להסיק כי יוסי יכול לעוף, אך זה אינו נכון אם יוסי הוא פינגווין או יען, וגם אם יוסי הוא תוכי ייתכן שנוצותיו נקצצו (שלא לדבר על גורלו המר של התוכי יוסי בסוף שירו של אברהם חלפי).

אנטומיה של גרירה

הגדרת המשימות בתחרות אינה מכתיבה שיטה כלשהי שבה יש להשתמש כדי לעמוד במשימות. תכונה זו מאפשרת לבחון מגוון רחב של רעיונות ושיטות. קבוצת המחקר באוניברסיטת בר-אילן לא רק הציעה את האתגר, אלא גם מפתחת גישה יישומית לעמידה בו. גישה זו מבוססת על פיתוח מנוע היסק, במסגרת עבודת הדוקטורט של רועי בר-חיים (Bar-Haim), המפעיל סדרה של טרנספורמציות שהופכות טקסטים לטקסטים אחרים.

ניקח לדוגמה את המשפט "ירד גשם כאשר ג'ון ומרי עזבו". לפני שנוכל להתחיל בתהליך הגרירה, אנו זקוקים לכלים חישוביים המפרקים את המשפט למרכיביו ומנתחים את התחביר שלו. זוהי משימה לא פשוטה, אבל כבר קיימים עבורה כלים מוכנים (מטבע הדברים, כלים אלה זמינים יותר עבור אנגלית מאשר עבור עברית, והדוגמאות המובאות כאן הן תרגום מאנגלית של הדוגמאות במחקרים המקוריים).

לא נציג את תוצאת הניתוח בצורה הפורמלית המקובלת בבלשנות, אלא בצורה הפשוטה (והלא-מדויקת) הבאה:

משפט ראשון: "ירד (פועל) גשם (שם עצם, נושא)"
קישור: "כאשר"
משפט שני: "ג'ון (שם עצם, נושא) ו- (חיבור) מרי (שם עצם, נושא) עזבו (פועל)"

ממשפט זה אפשר להסיק מסקנות רבות. ראשית נראה תהליך פשוט של הפעלת כללי טרנספורמציה. שנית, נפעיל את הכלל כי כאשר שני משפטים מחוברים על-ידי "כאשר", אפשר להסיק את כל אחד מהמשפטים, ולכן נוכל לדעת כי:

"ג'ון (שם עצם, נושא) ו- (חיבור) מרי (שם עצם, נושא) עזבו (פועל)"

עתה נפעיל את הכלל כי כאשר פועֶל מתייחס לשני נושאים המחוברים על-ידי ו' החיבור, אפשר להסיק כי הוא נכון עבור כל אחד מהנושאים, ולהסיק:

"מרי (שם עצם, נושא) עזבה (פועל)"

יש כמובן צורך בכללי טרנספורמציה תחביריים נוספים, כמו הפיכה ממשפט סביל לפעיל: למשל מַעבר מ- "דני זרק את הכדור" ל- "הכדור נזרק על-ידי דני" ומכאן ל- "הכדור נזרק".

כאשר המשפט הופך למורכב יותר וקשה יותר לניתוח נדרשים מנגנונים נוספים, וכן כאשר כללי הטרנספורמציה דורשים גם רמה מסוימת של הבנה סמנטית, שבה חשוב לא רק זיהוי תפקידה התחבירי של מילה מסוימת אלא גם משמעותה. לדוגמה:

(1) "ג'ון יודע שמרי עזבה" גורר "מרי עזבה"

(2) "ג'ון יודע שמרי לא עזבה" גורר "מרי לא עזבה"

(3) "ג'ון לא יודע שמרי עזבה" גורר "מרי עזבה"

(4) "ג'ון חושב שמרי עזבה" אינו גורר "מרי עזבה" (אי אפשר לדעת אם היא עזבה או לא)

כדי לפעול נכון על משפטים אלה, צריך בין השאר להבדיל

בין הקישור "יודע ש-" לבין "חושב ש-". כמו כן, יש צורך בכללים שונים לטיפול במילה "לא" במשפטים (ב) ו- (ג).

כללים נוספים לטרנספורמציה קשורים להתאמה בין משמעויות מילים. לדוגמה, אפשר להסיק מהמשפט "חברת טויוטה מייצרת מכוניות" את המשפט "חברת טויוטה מספקת כלי תחבורה". אם נוסיף גם כללי טרנספורמציה המתבססים על מידע כמו "חברת טויוטה היא חברה יפנית", נוכל להסיק את המשפט "חברה יפנית מספקת כלי תחבורה". מסקנה זו אולי אינה נראית מעניינת, אבל היא חיונית לצורך מענה על שאלות כמו "איזה חברות יפניות מספקות כלי תחבורה?". מנועי החיפוש הנמצאים בשימוש נרחב כיום אינם מסוגלים לענות על שאלה כזו, אבל אם היו נעזרים בגרירה טקסטואלית הם היו יכולים להפוך משפטים הנמצאים באינטרנט לצורה המאפשרת התאמה למבנה השאלה.

למידת כללים

צוות המחקר יצר רבים מכללי הטרנספורמציה התחביריים (כמו פירוק משפטים והפיכה מסביל לפעיל) והסמנטיים (כמו ההבדל בין "יודע ש-" לבין "חושב ש-") בצורה ידנית. מכיוון שכנראה אין צורך במספר עצום של כללים כאלה, הגישה של קידוד ידני של ידע לשוני היא סבירה. לעומת זאת, אין זה מעשי להזין בצורה ידנית את כל המידע הנדרש כדי להגיע למעברים כמו "מכונית" ל- "כלי תחבורה" (הגדרה לשונית) או "טויוטה" ל- "חברה יפנית" (ידע על העולם).

כדי ליצור את הכללים בצורה אוטומטית, פנו החוקרים למאגרים קיימים, כמו WordNet (קישור בסוף המאמר) המכיל מאות אלפי קישורים בין מילים בשפה האנגלית: חלון הוא חלק מבניין, נמר שייך למשפחת החתוליים, לחישה וצעקה הן סוגים של דיבור, ל-"בלון" (באנגלית) יש שתי משמעויות – כדור פורח או שק גומי דק הניתן לניפוח... לכל סוג של קישור אפשר להתאים כללי טרנספורמציה מתאימים: מילה ניתנת להחלפה במילה נרדפת לה, פועל או שם עצם ניתנים להחלפה במילה המבטאת קטגוריה יותר רחבה ("נחירה" היא "השמעת קול") או שניתן להסיק ממנה (אם אתה נוחר אפשר להסיק שאתה ישן), וכו'.

למרות גודלם של מאגרים כאלה, הם אינם מכילים מספיק מידע. כדי להרחיב את המידע, פנו החוקרים להסקה סטטיסטית מתוך טקסטים קיימים. בהקשר זה הראו תלמידיו של דגן, עידן ספקטור, מעין ז'יטומירסקי-גפת ושחר מירקין (Szpektor, Zhitomirsky-Geffet, Mirkin) כי אפשר לגלות מתוך ניתוח סטטיסטי של טקסטים רבים כי במקום שבו מופיעה המילה "company", למשל, יכולות להופיע גם המילים "firm, bank, group, subsidiary, unit, business, supplier, agency, division, entity, financial institution, manufacturer, corporation, joint venture"; וכי התבנית "X prevent Y" נגררת מתוך תבניות כמו "X reduce Y, X protect against Y, X eliminate Y, X stop Y, X for prevention of Y, X lower risk of Y, X be cure for Y, X treat Y, X in war on Y, X eliminate the possibility of Y". ברור כי אין אלה רשימות של מילים נרדפות, אבל מכיוון שמטלת הגרירה אינה דורשת

סימטרייה, ניתן לזהות את המקרים שבהם מותר להחליף מילה במילה שאינה נרדפת - למשל מילה כללית יותר.

מקור נוסף למידע על העולם הוא האנציקלופדיה הפתוחה Wikipedia. תלמיד אחר של דגן, אייל שנרך (Shnarch), עסק בעבודת המסטר שלו במציאת יחסי גרירה מתוך הגדרות ב-Wikipedia: למשל, מתוך ההגדרה "עט הוא מכשיר המשמש לכתיבה בדיו" אנו רוצים להסיק "עט הוא מכשיר (יחס הכללה), "עט משמש לכתיבה" (יחס שימוש - כלומר שאם מישהו השתמש בעט סביר להסיק כי הוא כתב משהו), ו- "עט מכיל דיו" (יחס הכלה).

אנו זקוקים לניתוחים כאלה כדי להעשיר את המאגרים של

יחסים לשוניים ושל מידע על העולם, שבלעדיהם לא תיתכן הבנת טקסט. מצד שני, כדי לבצע ניתוחים אלה כשלעצמם אנו זקוקים ליכולות בסיסיות של עיבוד טקסט. כאמור, לפחות חלק מיכולות אלה הן כבר בסל הכלים שפותחו בעשרות השנים האחרונות במסגרת המחקר בעיבוד שפה טבעית, אבל עדיין יש צורך ברעיונות ובפיתוחים חדשים (למשל, אילו תכונות טקסטואליות של "עט הוא מכשיר המשמש לכתיבה בדיו" מצביעות על כך שהעט מכיל דיו?).

יצירת מאגרים עשירים דיים אינה האתגר היחיד עבור התקדמות בגרירה טקסטואלית. דוגמה מעניינת נוספת היא מילים שיש להן יותר ממשמעות אחת, כאשר כל משמעות עשויה להוביל למסקנה אחרת. ברוב המקרים אין בני-אדם מתקשים בכך, מכיוון שההקשר מכתיב את המשמעות הרלוונטית. לדוגמה, כנראה שלא נהיה מודעים להבדלים בעיבוד המשפטים "התרנגולת הטילה ביצה" מול "המנהלת הטילה על דן לארגן את הכנס". ההקשר מכתיב את המשמעות השונה של המילה "הטילה". קבוצתו של דגן מציעה ליצור כללי גרירה כך שהם שומרים על התאמה בין ההקשרים המתאימים לכל חלק במשפט, כך שהמושא "ביצה" מחייב התאמת המשמעות הנכונה עבור הפועל "הטילה" (ולהפך – ברור לנו כי לפי המשמעויות האפשריות לפועל "הטילה", המילה "ביצה" אינה מתייחסת לשטח אדמה המכוסה במים רדודים, בִּיצה).

בשבחי הגרירה

ברור כי הצלחה בזיהוי גרירה טקסטואלית, לפחות במצב הנוכחי של המחקר, אינה בת-השוואה להבנה אנושית של טקסט כשלעצמה. לשם כך חסרים חלקים רבים הנחשבים כרגע מחוץ לתחומה של הגרירה. דוגמה אחת מרבות: אחד הכלים החשובים המשמשים אותנו כדי להבין טקסט הוא השלמת פרטים מתוך הידע שלנו על העולם: מתוך "אבנר אכל במסעדה אבל היה מאד לא מרוצה מהשירות ומהאוכל" נוכל להסיק כי דן כנראה שילם על הארוחה אבל לא השאיר תשר. אתגרים אלה אינם חדשים בעולם הבינה המלאכותית[1], וכנראה כי יש צורך בהתקדמות מקבילה בערוצים רבים כדי להתקרב לרמת ההבנה האנושית. מצד שני, סביר להניח כי הערוצים השונים יכולים לתרום זה לזה: למשל, מי שעוסק בהשלמת פרטים מתוך ידע על העולם יוכל להיעזר בגרירה טקסטואלית כדי לזהות את המרכיבים שכבר נמצאים בטקסט ולקשור אותם לידע רלוונטי, ולהפך: המסקנות של השלמת הפרטים יוכלו לתרום למציאת עוד מסקנות הנובעות מהטקסט.

מדוע זכתה גישת הגרירה בפופולריות רבה במהירות כזו? יש לכך כמה סיבות: ראשית, כפי שכבר הוזכר, היא מהווה אתגר "בגודל הנכון". שנית, יש בה תרומה תיאורטית: הגרירה היא שכבה חדשה של ניתוח, מעל הרמה המורפולוגית (כלומר צורת המילה, כמו כאשר המחשב נדרש לזהות את המילה "וחתוליה" כנגזרת משם העצם "חתול" יחד עם ו' החיבור, סיומת של ריבוי וסיומת של שייכות) והרמה התחבירית (זיהוי של מרכיבי המשפט – נושא, נשוא, מושא, שם תואר וכו'). שכבה חדשה זו נושאת את הפוטנציאל לשימוש כמודול היסק מרכזי עבור יישומים שונים.

גם אם גרירה אינה מטפלת בכל הנושאים הסמנטיים, כך שיידרשו כלים נוספים, הרעיון הוא שמנוע גרירה יהווה את השלד להיסק סמנטי כללי. שלד זה יאפשר להוסיף יכולות סמנטיות בלי להידרש "להמציא את הגלגל" בכל פעם מחדש. מצד שני, רעיון הגרירה הוא בלתי-תלוי, ואינו נוגד תיאוריות ספציפיות מתוך תחומי הבלשנות, הפילוסופיה של החשיבה או הבינה המלאכותית. לכן ענפים תיאורטיים רבים יכולים להיעזר בו. הסיבה השלישית היא אולי החשובה ביותר: כאמור, התקדמות בהבנת טקסט עשויה להוביל מהפכה בדרך שבה אנו עובדים עם טקסט ועם מחשבים. נדע כי המהפכה אכן הגיעה כאשר תוכנה תוכל להסיק כי הטקסט "רעיון הגרירה הטקסטואלית נראה כבעל פוטנציאל לתרום בחיפוש, בתרגום, בסיכום, במענה על שאלות, בהוראה ממוחשבת, ובשימושים חשובים נוספים" גורר את המסקנה "רעיון זה ימשיך לזכות בתשומת לב רבה גם באקדמיה וגם בתחום המסחרי". ❖

ישראל בנימיני עובד בחברת ClickSoftware בפיתוח שיטות אופטימיזציה מתקדמות.

לקריאה נוספת:

פרויקט Cyc - http://www.cyc.com

רשת המחקר PASCAL - http://www.pascal-network.org

מסלול ניתוח הטקסט במסגרת NIST - http://www.nist.gov/tac

WordNet - http://wordnet.princeton.edu,

1. הדוגמה של המסעדה, למשל, לקוחה מעבודתו של רוג'ר שנק (Schank) שהחל בשנות השבעים להדגיש את התפקיד של "סיפורים" או "תסריטים" בהבנת טקסט, על-ידי שיבוץ הפרטים החסרים מתוך תסריט מוכר: במקרה זה, ביקור במסעדה כולל הזמנת אוכל, אכילה, תשלום והוספת תשר התלוי בשביעות הרצון.