פרוייקט סיום בקורס "חישוב ומשמעות"

הרחבת המודל במאמר

"Nouns are Vectors, Adjectives are Matrices"

Participle Verbs עבור

יובל שטיינמץ 204209233 ואלה נאמן 204106918

רקע

כמה מחקרים בשנים האחרונות ביקשו לייצר גישה קומפוזיציונלית לסמנטיקה בעולם דיסטריבוטיבי-לקסיקלי. לפי הגישה הקומפוזיציונלית משמעותם של ביטויים מורכבים נקבעת ע"י המשמעות של תתי-הביטויים המרכיבים אותם, יחד עם החוקים המשלבים ביניהם. לפי הגישה הדיסטריבוטיבית-לקסיקלית, מייצגים מילים במרחב על סמך התפוצה שלהן, ולא עוסקים בייצוג משמעות של משפט. כל מילה מיוצגת בתור וקטור של תפוצת המילים שמופיעות בסמיכות לה.

במאמר "Nouns are Vectors, Adjectives are Matrices" ביקשו החוקרים, מרקו בארוני ורוברטו זאמפרלי, לייצר שיטות להרכבת משמעות מוקטורים של מילים, באמצעות גישה קומפוזיציונלית. גישה זו מחפשת אחר מטריצות מעבר ופעולות חיבור וכפל שהן "החוקים" שמבארים את המשמעות המשותפת. אלו, מביאות לאחר הפעלתן לוקטור המייצג את אותה משמעות. גישה זו מבקשת לברר אילו פעולות מתמטיות עשויות להסביר את הקשר בין וקטורים של שם תואר ושם העצם שמופיע אחריו. המאמר התמקד בתארים אטריבוטיביים המקדימים למילה (attributive adjective).

החוקרים הציעו גישה מקורית לייצוג שמות תואר כפונקציות לינאריות אשר חוזות את התפוצה המשותפת של שם עצם ושם תואר. בהינתן וקטורים המייצגים שמות עצם (N) לפי הסמנטיקה הדיסטריביוטיבית, ו-וקטורי תפוצה משותפת (AN), לומדים את המטריצה A שמהווה ייצוג לשם התואר.

$$\begin{pmatrix} y_{1,1} & \cdots & y_{1,n} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}^T$$

$$A \cdot N = AN_{approx}$$

השערת המחקר

באנגלית, participle verb הוא פועל המשמש במשפט כדי לשנות שם עצם, ביטוי שמני, פועל או ביטוי פועלי וממלא תפקיד דומה לשם תואר או שם פועל. בספר *A dictionary of linguistics and phonetics.* של David Crystal משנת 1991 מופיעה ההגדרה הבאה:

"a word derived from a verb and used as an adjective".

לדוגמא, במשפט "the painted wall" הפועל painted הוא participle verb המשמש לתיאור שם העצם wall.

המחקר שלנו מבקש לבחון האם אותו קשר בייצוג הזוג שם עצם ושם תואר אטריביוטיבי מתקיים גם בייצוג הזוג שם עצם ו-participle verb. כלומר, נבקש לברר האם ניתן לחזות בעזרת המודל של בארוני וזאמפרלי את המשמעות המשותפת של מילים כאלו באמצעות פונקציה לינארית שנלמדת מדוגמאות. הפונקציה הלינארית מתוארת בדומה למאמר המקורי בצורת מטריצה, אותה נסמן V. היא תופעל על וקטור המייצג שם עצם N. וקטורי ההופעה המשותפת של participle verb ו-noun יכונו VN.

$$\begin{pmatrix} y_{1,1} & \cdots & y_{1,n} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}^T$$

$$V \cdot N = VN_{approx}$$

כלים וקורפוס

פרויקט DISSECT הינו פרויקט קוד בתחום עיבוד השפה הטבעית. הוא מעניק כלים לבניית מרחבים סמנטיים המיוצגים על ידי מטריצת הופעות משותפות, לביצוע פעולות קומופוזיציונאלית על מרחבים אלו, ולמדידת המרחק הסמנטי בין מילים או ביטויים. הפרויקט מתמקד במשמעות הקומופוזיציונלית של המילים, כלומר הוא מכיל פונקציות אשר גוזרות את המשמעות של ביטויים ומשפטים מתוך המשמעות של חלקי המשפט.

הפרויקט מכיל בין השאר כלים למימוש הרעיון אשר הוצג במאמר שהזכרנו לעיל ונבנה בשיתוף כותביו. החוקרים עשו שימוש בכלי הפרויקט בשביל למדוד את הקומפוזיציונליות של שמות עצם ושמות תואר. בעבודה זו השתמשנו בפונקציות שסופקו בפרויקט, והתאמנו אותן לעבודה עם participle verbs בתפקיד של שמות התואר על מנת לשחזר את הניסוי מהמאמר.

איך ניגשנו לבעיה?

בחירת הקורפוס

רצינו לשחזר את הניסוי באמצעות אותו קורפוס עתיר משפטים שחוקרי המאמר עשו בו שימוש. מדובר בקורפוס מאוחד המורכב ממספר מקורות, ביניהם עותק של ויקיפידה משנת 2009 ו- British National Corpus. כדי להשתמש במאגר המשפטים יצרנו קשר עם החוקרים וקיבלנו הרשאות לשימוש מחקרי בו. קיבלנו גישה לשני מאגרים:

- ukWack הוא קורפוס בן 2 מיליארד מילים אשר מורכב ממילים עם תדירות בינונית שמופיעות באתרים עם הסיומת uk. בכתובת האתר. משקלו 7.6 GB.
- Wackypedia EN הוא עותק של ויקיפדיה האנגלית המכיל בערך 800 מיליון מילים. משקלו 6.0 GB.

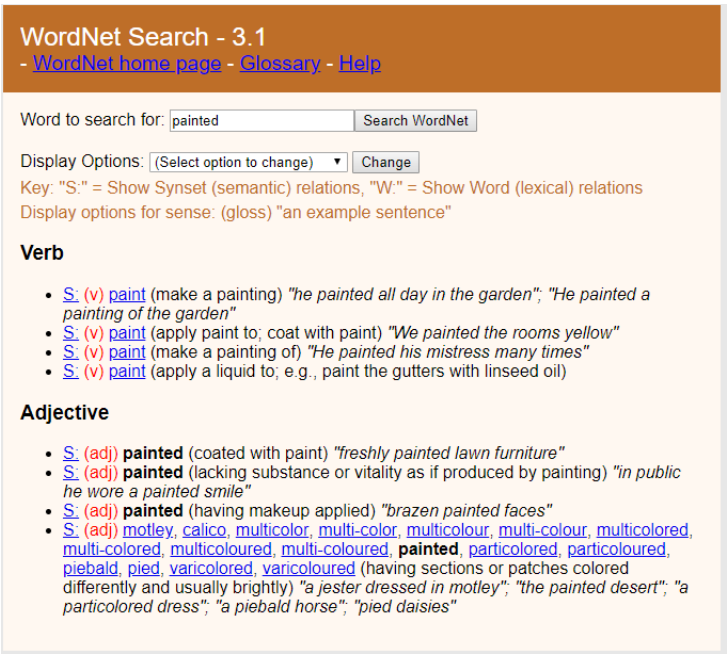
שני הקורפוסים מתויגים לפי חלקי דיבר. גילינו שעבודה עם מאגרים אדירי-מימדים כגון אלו הינה מסובכת עד בלתי-אפשרית באמצעים שעמדו לרשותנו ובחרנו לעבוד במקום זאת עם ספריית nltk ולעשות שימוש בקורפוס בראון.

קורפוס בראון מכיל מיליון מילים בשפה האנגלית ונוצר בשנת 1961 באוניברסיטת בראון בארה"ב. הוא מכיל טקסט מ-500 מקורות, והמילים בו מתויגות לפי חלקי דיבר ולפי קטגוריות.[1] במהלך העבודה החלטנו להגביל את כמות המשפטים מהקורפוס ל-16 אלף, בשל קשיי הרצה ויעילות שהתמודדנו איתם לכל אורך המחקר. את צמדי המילים ממשפטים אלו (לשם ספירת "הופעות משותפות") הצלבנו רק עם מילים מתחום הסיפורת (קטגוריית fiction), סה"כ 8,795 מילים. נציין כי 16 אלף המשפטים מכילים בתוכם את קטגוריית הסיפורת.

בחירת הפעלים

ביקשנו לייצר רשימה של participle verbs. לשם כך, השתמשנו בספריית nltk אשר מציעה ממשק לעבודה עם WordNet, מסד-נתונים לקסיקאלי עבור השפה האנגלית המספק בין השאר הגדרות קצרות ודוגמאות שימוש. מתוך כל המילים במאגר סיננו מילים שהן גם פועל וגם שם תואר, אך לא משמשות כשם פועל או כשם עצם. למשל, נביט בצילום מסך מתוך הממשק האינטרנטי של WordNet עבור תוצאת החיפוש של painted:

[1] https://www.nltk.org/book/ch02.html



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: painted | Search WordNet

Display Options: (Select option to change) | Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Verb

- S: (v) paint (make a painting) *"he painted all day in the garden"; "He painted a painting of the garden"*
- S: (v) paint (apply paint to; coat with paint) *"We painted the rooms yellow"*
- S: (v) paint (make a painting of) *"He painted his mistress many times"*
- S: (v) paint (apply a liquid to; e.g., paint the gutters with linseed oil)

Adjective

- S: (adj) **painted** (coated with paint) *"freshly painted lawn furniture"*
- S: (adj) **painted** (lacking substance or vitality as if produced by painting) *"in public he wore a painted smile"*
- S: (adj) **painted** (having makeup applied) *"brazen painted faces"*
- S: (adj) motley, calico, multicolor, multi-color, multicolour, multi-colour, multicolored, multi-colored, multicoloured, multi-coloured, **painted**, particolored, particoloured, piebald, pied, varicolored, varicoloured (having sections or patches colored differently and usually brightly) *"a jester dressed in motley"; "the painted desert"; "a particolored dress"; "a piebald horse"; "pied daisies"*

אספנו את המשפטים לדוגמא שמופיעים עבור כל ערך כדי ליצור מאגר דוגמאות (בתמונה: מופיעים תחת השימוש של ה-participle verb בתור שם תואר). מתוך דוגמאות אלו רצינו ללמוד על משמעותם המשותפת של צירופי שם עצם ו-participle verb. כדי ליצור את מדגם האימון לקחנו את ה-participle verbs אשר יש להם ארבע דוגמאות שונות או יותר, כלומר ישנם לפחות ארבעה צמדים שונים של שם עצם ו-participle verb.

רצינו לבדוק בקורפוס בראון את ההופעות של אותם זוגות שבודדנו מ-WordNet. לשם כך ביצענו חיתוך של קבוצה זו עם המילים בקורפוס בראון המתויגות 'VBN'. אבל לא צפינו את העובדה כי מילים אלו מופיעות מספר מועט של פעמים, ולא לצד אותם שמות עצם. לכן נאלצנו לוותר על הרעיון ולזנוח את המילים שבודדנו מ-WordNet. לאור זאת, השתמשנו רק במילים אשר תויגו בתור participle verbs בקורפוס בראון, אשר הופיעו לצד מילים שתויגו כשמות עצם בקורפוס זה.

יצירת מרחבים סמנטיים

יצירת מרחב הליבה

מרחב סמנטי נוצר על ידי ספירת ההופעות המשותפות של המילים. המרחב מיוצג על ידי מטריצה, כאשר העמודות מייצגות את המילים בקורפוס, ואילו השורות את שמות העצם וה- participle verbs. כפי שצוין לעיל, עבדנו לבסוף עם קורפוס מוגבל של כ-9,000 מילים מקטגוריית הסיפורת. סה״כ אספנו כ-8,543 שמות עצם ופעלים שהיוו את שורות המטריצה.

כל שורה היא למעשה וקטור המכיל את מספר ההופעות המשותפות עם כל מילה בקורפוס המוגבל. ספירת ההופעות המשותפות נעשתה בעזרת הפונקציה bigram של nltk אשר מחלקת את הטקסט לצמדי מילים סמוכות, כך שניתן לספור הופעות בחלון בגודל 1 עבור כל מילה. בתמונה ניתן לחזות בחלק קטן ממטריצת ה-co-occurrence ולהבחין באופי הדליל שלה.

	A	Aaron	Abbe	Abbey	Abbot	Abel	Ablard	Academy	Ada	Ada's	Adam	Adonis	Africa	African	Afternoon	Ages	Aggie	Aj	Al	Alabama	Alastor	Albany
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aaron		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Abbe		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Abbey		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Abbot		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Abel		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ablard		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
About		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Above		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Academy		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Accounts		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Across		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acting		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ada	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ada's		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Adam		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Adonis		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Afraid		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
African		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
After		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Afternoon		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Afterwards		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Again		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ages		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aggie		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ahead		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ain't		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Al	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Al's		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alabama		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alas		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alastor		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Albany		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Albright		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Albright's		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alex		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alex's		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alexander		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alfred		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alix		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Alix's		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
All	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Allen		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

יצירת Peripheral Space

מרחב ה-peripheral מכיל צמדים של מילים אשר מקבלות משמעות שונה כאשר הן מופיעות יחדיו. במקרה שלנו מטריצת ה-peripheral תכלול את הביטויים המורכבים. לדוגמא, במטריצת הליבה תהיה שורה עבור המילה wall, ואילו במטריצת ה- peripheral תהיה שורה עבור painted wall כביטוי מורכב. ספירת ההופעות המשותפות נעשתה בעזרת הפונקציה trigram של nltk אשר מחלקת את הטקסט לשלשות של מילים סמוכות. לאחר מכן שרשרנו את שתי המילים הראשונות בכל שלשה לביטוי אחד. בצורה זו יצרנו למעשה צמדים של ביטוי ומילה. כך ספרנו הופעות בחלון בגדול 1 עבור כל ביטוי.

החלטנו להרחיב את מרחב הperipherals גם לזוגות של שם עצם ושם תואר לצורכי הערכת המודל, ובכך העשרנו את המרחב הסמנטי. לבסוף יוצרו 3,689 שורות במטריצת ה-peripheral המכילות שמות עצם ולפניהם שם תואר או participle verb.

בתמונה ניתן לחזות בחלק קטן ממטריצת ה-co-occurrence שמשמשת לבניית מרחב ה-peripherals. נבחין כי עמודות המטריצה זהות לעמודות מרחב הליבה ואילו שורות המטריצה מכילות צמדי-מילים של VNים ו-ANים.

	A	Aaron	Abbe	Abbey	Abbot	Abel	Ablard	Academy	Ada	Ada's	Adam	Adonis	Africa	African	Afternoon	Ages	Aggie	Aj	Al
British_column	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
British_ships	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brown_eyes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fine_Rector	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
French_coffee	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Funny_thing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Good_man	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jewish_section	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Naked_girls	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Precious_right	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rich_people	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Then_Rector	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
alien_water	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
anchored_ships	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
appointed_table	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
approached_Rector	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
armored_vest	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
arranged_things	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
asked_know	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
asked_watching	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
awful_good	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bad_things	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bandaged_wound	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
barbed_wire	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bare_trees	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bashful_boy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
befogged_loneliness	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
big_body	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
big_trees	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
black_silk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ראוי לציין כי לפני בניית ה-bigrams וה-trigrams סיננו את כל סימני הפיסוק. לכן, ייתכן ששני חלקים ממשפטים שונים יתחברו לכדי ביטוי אחד, אשר לא קיים בקורפוס.

במסגרת העבודה עם כלי DISSECT נדרשנו ליצור קובץ לכל מטריצת הופעות משותפות. קובץ זה נשמר בצורה דחוסה (dense), כלומר עבור כל צמד מילים מצוין מספר ההופעות המשותפות (אלו הדוגמאות המופיעות מעלה בתמונות החלקיות). ניסינו לעבוד גם בצורה דלילה (sparse) אבל ניסיונות אלו לא צלחו. DISSECT דורש קובץ נתונים בפורמט אחיד (דחוס או דליל) והצלחנו לשמור את המידע באחת מהצורות המתאימות.

שלב הלמידה

אנו מסתמכים על ההשערה כי המשמעות של צמדי המילים אינה נגזרת מהתפוצה של הפעלים בנפרד, אלא לפי איך שהפעלים (בדומה לשמות התואר) משנים את התפוצה של שמות העצם. המודל לומד את הפעלים כפונקציות שפועלות על שמות העצם.

ביקשנו ללמוד שמונה participle verbs אשר נבחרו על-פי כמות שמות-העצם הייחודיים שהם מופעים לפניהם (שלושה שמות עצם לפחות). הלמידה נעשתה באמצעות הכלים ש-DISSECT מספקת, באמצעות הורדת מימד (SVD) והפעלת ridge regression עם פרמטר 2. כך למדנו שמונה מטריצות, אחת לכל participle verb.

הערכת המודל

ביקשנו לבדוק האם הקירוב במשמעות עובד גם על צמדי מילים חדשים, כלומר כאלו שהמודל לא ראה במסגרת האימון. קיווינו כי צמד מילים סינתטי יתנהג דומה ככל האפשר לצמד מילים "טבעי", שיוצר מספירות "אמתיות" שנלקחו מהשפה.

חוקרי המאמר השתמשו במספר שיטות להערכת המודל שלהם. אחת מהן הייתה בחינה של סביבת התוצרים במרחב הסמנטי. תחילה הם ייצרו את וקטורי ה-AN הסינתטיים באמצעות מטריצת שם התואר. לאחר מכן, בנו מוקטורים אלו צנטרואיד עברו כל שם תואר ומצאו את השכנים הקרובים ביותר של כל צנטרואיד, כאשר החישוב נעשה באמצעות קוסינוס הזוית בין שני הוקטורים. כך ביקשו החוקרים לוודא כי הצנטרואיד נמצא בסביבת ה-ANים הטבעיים במרחב הסמנטי. למשל, עבור הצנטרואיד של American N (שם התואר "אמריקאי" ואחריו שם עצם כלשהו), נמצאו השכנים הטבעיים "נציג אמריקאי", "טריטוריה אמריקאית" ו"מקור אמריקאי". נוסף על כך, הגרילו החוקרים צמדים של AN וביקשו להביט בשלושת השכנים הקרובים ביותר שלהם. כך הם בדקו כי לפחות ברמה האינטואיטיבית הוקטורים הסינתטיים דומים בהתנהגותם לוקטורים הטבעיים.

התלבטנו רבות כיצד למדוד את הצלחת המודל. נתקלנו בקשיים משום שהמטריצות שהחזקנו דרשו זמן ריצה רב ולעיתים לא עמדו במגבלות הזיכרון של המחשב. החלטנו לפשט את צורת ההערכה מהמאמר ולהציע דרך משלנו – בחינת הסביבה של הוקטורים הסינתטיים בתוך עולם של וקטורי VN ו-AN טבעיים.

במסגרת הבדיקה, לאחר שלמדנו את מטריצת ה-V, כפלנו אותה בשמות עצם N שלא נראו במדגם האימון וקיבלנו ביטוי מורכב VN. לבסוף, בדקנו מי הם השכנים הקרובים ביותר במרחב הצמדים של הוקטור הסינתטי שיוצר. גם אצלנו בדיקת הדמיון נעשית על ידי חישוב קוסינוס הזווית בין שני וקטורים. את מרחב הperipherals עיבינו ע״י הוספת זוגות של שם עצם ושם תואר כדי שהסביבות יהיו אינפורמטיביות יותר.

לכל מטריצת V הצענו חמישה שמות-עצם, שלא נראו במדגם האימון, המתאימים להופיע לאחר הפועל המוטה המשמש כשם תואר. בדקנו האם התוצאות תואמות את האינטואיציה לקרבה סמנטית בין המילים. אלו שמות העצם שהתאמנו לכל פועל:

Broken	Changed	Closed	Colored	Detailed	expressed	given	improved
table leg hand window car	weather circumstances country shape case	farm group way bottle store	man baby table car shirt	explanation path book figure magazine	words feelings disappointment shame sorrow	car book condition goals facts	machine car work manner position

תוצאות

בטבלה הבאה מופיעים הצמדים שהמטריצות שלמדנו היטיבו לחזות את המשמעות שלהם. הם מופיעים יחד עם כשלושה שכנים שמשמעותם תואמת יחסית (מתוך עשרת השכנים הקרובים ביותר).

broken hand	changed country	closed way	improved position
broken nails	changed appearance American families American speech	closed bedroom closed fields human freedom	improved equipment aroused interest
colored shirt	detailed explanation	given facts	expressed feelings
cheap clothing rosy mouth	important role detailed study essential part	available supplies great number given tickets	English-Dutch manors cold annoyance

אולם לא תמיד התוצאות תואמות את המשמעות הרצויה. בטבלה הבאה מופיעים צמדים שהמטריצות שלמדנו לא הצליחו לחזות את משמעותם. השכנים הקרובים ביותר אליהם אינם קשורים אליהם מבחינה סמנטית.

broken table	changed weather	closed store	improved work
black silk Big ones Christian thought	changed appearance long knife rosy mouth	Certain features intellectual activity smooth fashion	improved equipment old days hot water
colored man	detailed figure	given car	expressed feelings
improved equipment old days hot water	essential part important part large part	far end Black strips Jewish prisoners	blasphemous rites cool clothes completed schedules

את פלט התוצאות המלא יחד עם ציוני הדמיון ניתן למצוא בנספחים.

נבחין כי יש מילים שחזרו כשכנים עבור תוצאות הכפלת המטריצות השונות, ולא מאד ברור כיצד הן קשורות מבחינת דמיון סמנטי. למשל, השכן American families הופיע לצד לצמדים שיוצרו על ידי ארבעה מהפעלים. השכן Cheap clothing חזר גם עבור תוצאות של changed וגם עבור colored כמו גם השכן long knife אשר לא מתאים במשמעות לאף אחד מהפעלים ושמות העצם.

עוד נציין כי רשימות עשרת השכנים הקרובים ביותר של שמות העצם שהוכפלו בפועל expressed היו זהות. גם ב-broken וב-improved הרכב הרשימות דומה מאד. בשאר הפעלים, תוצאות ההכפלה של חלק משמות העצם היו רשימות דומות.

ניסינו לנמק זאת בכך שעבור expressed רשימת שמות העצם מכילה רגשות, אשר מגיעים מאותה קטגוריה סמנטית. לכן, הגיוני שהמטריצה שנלמדה תשלח אותם לאותו איזור במרחב הסמנטי. ייתכן גם כי התופעה כולה ניתנת להסבר על-ידי כך שהמטריצות מאפסות חלקים בוקטור, כלומר חלק ממרכיבי הוקטור המקורי לא נשמרים לאחר הפעלת המטריצה. מסיבה זאת ייתכן כי וקטורי ה-VN נמצאים בסביבות דומות.

מסקנות

התוצאות שהתקבלו עשויות להתפרש בכמה מובנים ואינן משכנעות באופן חד-משמעי בדבר קיום של פונקציה לינארית שפועלת על שמות עצם לבניית משמעות מורכבת. זאת בניגוד לתוצאות שהוצגו במאמר שעליו התבססנו.

ייתכנו מספר סיבות לכך. השערה אחת היא ששחזור הניסוי התבצע על קורפוס קטן שלא כולל הרבה הופעות של צמדים מהסוג שרצינו לבחון. ייתכן כי המרחב שבנינו אינו מספיק מייצג, ועקב כך קשה להכריע האם ה-participle verbs מתנהגים כמו שמות עצם. השלכה נוספת של שימוש במרחב סמנטי שאינו עשיר הייתה על שיטת ההערכה: חיפוש אחר שכנים קרובים ובדיקת דמיון בין וקטורים עלולים להיות לא מדויקים במרחב דל, ואנחנו חוששים שזה היה המקרה בשחזור הניסוי שלנו. גם כאשר ניסינו לעבות את המרחב הסמנטי באמצעות צמדים של שמות עצם ושמות תואר, כדי להגדיל את היצע השכנים, לא קיבלנו את התוצאה הרצויה והמרחב היה דל מידי.

אנחנו חושדים כי התוצאות שקיבלנו נובעות מהיותו של המרחב שייצרנו בלתי-ניתן להפרדה. זאת משום שהספירות שהגענו אליהן מועטות יחסית. כבר ציינו כי הוקטורים מאד דלילים, אולם יותר מזה, כמות הספירות בכל קואורדינטה חיובית מסתכמת במספר חד-ספרתי בלבד. יתר על כן, הערכים החד-ספרתיים הגדולים אינם שכיחים כלל.

המכשול המרכזי שעמד בפנינו היה יכולת החישוב המוגבלת שעמדה לרשותנו. חישוב ההופעות המשותפות של כל זוג מילים דרש זיכרון מעבר ליכולות המחשבים האישיים שלנו והמחשבים במעבדת המחשבים באוניברסיטה. כאשר ניסינו לצמצם את כמות המשפטים מהקורפוס שעליהם מתבצעת הספירה, קיבלנו תוצאות לא מייצגות ומאכזבות מבחינת כמות. ככל שניסינו להגדיל בחזרה את כמות המשפטים, הריצה דרשה זמן ממושך יותר (בשעות), כאשר הרצה מלאה פשוט קורסת. אם כן, ניסינו לייצר טרייד-אוף בין עושר המרחב הסמנטי ליכולת להריץ בזמן סביר (או בכלל) את המימוש שלנו.

בעבודת המשך כדאי לבחון שימוש מושכל יותר במבני נתונים ספרסיים ובהרצה מקבילית של הקוד שמבצע את הספירות. הרצה במקביל תאפשר חיסכון בזמן ריצה; בעוד ששמירת הנתונים בצורה מושכלת עשויה לחסוך בזיכרון. תוך שילוב הפתרונות הללו ייתכן ונוכל להתגבר על המכשולים שציינו.

לסיום, עדיין מוקדם לשלול את ההשערה כי המשמעות של צמדי המילים נגזרת מהצורה שהפעלים, בתפקידם כשמות התואר, משנים את התפוצה של שמות העצם.