מבוא

במחקר הבלשני נהוג להבדיל בין מילים לפי חלקי דיבר (POS), אשר משמשים להגדרת התפקיד התחבירי שהמילה יכולה למלא. הגדרה של מילה לפי ה-POS שלה יכולה לסייע במקרים של כפל משמעות, הנגרם לרוב כאשר נוצר בלבול בין שתי מילים המובעות בצורה זהה אך שונות ב-POS שלהן. כך למשל, את כפל המשמעות של המשפט "חולצה מטיילת" ניתן לפתור באמצעות ציון לגבי כל אחת מהמילים האם היא פועל או שם עצם. באנגלית ציון ה-POS של כל מילה שימושי במיוחד, מכיוון שמעבר לפוליסמיה הקיימת בכל שפה תקין ואף מקובל באנגלית לעשות שימוש במילה בעלת POS מסוים בתור POS אחר, כאשר השינוי ב-POS לאו דווקא מתבטא מבחינה מורפולוגית. כך למשל המילה "run" יכולה לשמש כפועל או כשם עצם, והמילה "inspired" יכולה לשמש כשם תואר או כפועל.

מכיוון שקיימים בשפה כפלי משמעות רבים כאלו, וכמובן שה-POS אינם מצוינים במקומות הרלוונטיים, הגיוני להניח כי קיים מנגנון קוגניטיבי המאפשר לבני-אדם ברוב המקרים להצליח להסיק מה הם ה-POS הנכונים של מילה דו-משמעית (באופן לא מודע כמובן). חוקי התחביר של כל שפה הם המגדירים אילו שילובים של POS תקינים בשפה, אך אנו יוצאים מנקודת הנחה כי מנגנון קוגניטיבי שכזה לא כולל ניתוח תחבירי של משפט – ניתוח תחבירי דורש את המשפט המלא, בעוד שבמהלך קריאה או דיבור המשפטים נקלטים באופן הדרגתי מילה אחרי מילה מבלי ליצור בלבול בהבנת ה-POS (אם נתעלם ממשפטי garden path שהם נדירים בשפה טבעית). מתוך כך, ניתן להבין כי לדובר שפה המנסה להכריע לגבי ה-POS של מילה דו-משמעית יש שני מקורות מידע להתבסס עליהם: האחד, ההסתברות היחסית שהמילה תקבל POS מסוים על סמך כלל ה-POS שקיבלה במקרים בהם נחשף אליה בעבר. השני, ההסתברות שהמילה תקבל POS מסוים על סמך התכונות של המילים הקודמות לה במשפט. קל להוכיח שמקור המידע הראשון אינו מספיק, מכיוון שגם במשפט המכיל מילים דמיוניות שהקורא לא נתקל בהן בעבר הוא מסוגל בקלות יחסית למצוא את ה-POS המתאים לכל מילה (לדוגמא שיר ה-Jabberwocky). משום כך, מקובל להניח כי מציאת ה-POS של מילה דו משמעית מתבססת בעיקר על מידע הסתברותי המוכל בחלק המשפט שקדם לה.

בעבודה הנוכחית יצרנו מודל פשוט הבוחן POS של מילה בהתבסס על ה-POS של המילים הקודמות לה במשפט (מודל n-gram), תוך התייחסות ספציפית ל-POS הנפוצים פועל, שם עצם, ושם תואר. כאמור, אנחנו מניחים שלכל POS יש סביבה אינדיקטיבית המסייעת לזהות אותו במשפט, ובעבודה הנוכחית בכוונתנו לבחון האם הסביבות האינדיקטיביות של ה-POS הנפוצים שונות מבחינת הגודל שלהן. לשם כך בכוונתנו להשוות בין מודל ביגרם ומודל טריגרם עבור כל אחד מה-POS האלו, במטרה למצוא האם עבור זיהוי של POS שונים קיים יתרון למודל n-gram מסוים על פני מודלים בגדלים אחרים. מסיבות של היקף העבודה ומסגרת זמנים בחרנו במימוש מצומצם יחסית של שני n-grams ושלושה POS, כמו גם בקורפוס מצומצם של כ-4,000 משפטים.

קורפוס

המשפטים המתויגים נלקחו מקורפוס Brown מספריית NLTK מקטגוריית News. סה"כ הקורפוס הכיל 4,623 משפטים.

כחלק מהניתוח ל-POS שהקורפוס מספק קיימים מספר תגים שהקורפוס מבדיל ביניהם, אך שלצורך העבודה הנוכחית לא מצאנו צורך להבדיל ביניהם (אנו מודעים לכך כי עשויים להיות הבדלים בתפוצה בין אחדים מהתגים האלו, אך במסגרת העבודה החלטנו להתעלם מכך). לשם כך, בשלב ניתוח התוצאות בלבד, יצרנו שלוש קבוצות של "על-תגים" התואמות לשלושת ה-POS בהם התמקדנו בעבודה, כפי שמפורט בטבלה:



חשוב לציין שוב כי חלוקה זאת לעל-תגים לא באה לידי ביטוי מבחינת הקוד, אלא רק כדרך שלנו לבחון את התוצאות בצורה בהירה יותר.

כחלק מהניתוח ל-POS שהקורפוס מספק קיימות מילים מורכבות אשר מקבלות שני תגים (כמו למשל המילה "wanna" שמקבלת תיוג של VB+TO). בחרנו שלא לכלול מילים עם שני תגים בניתוח התוצאות יחד עם התגים שציינו בטבלה לעיל, על מנת להימנע מניתוח כפול של אותה המילה. במילים אחרות, מידת ההצלחה של האלגוריתם בזיהוי מילים עם שני תגים נכללת באחוז ההצלחה הכללית של האלגוריתם, אך לא באחוזי ההצלחה של שלושת ה-POS הספציפיים שבדקנו.

הגבלות

על מנת לצמצם את זמן הריצה של האלגוריתם לזמן סביר עבור מודל הטריגרם בחרנו להריץ אותו רק על התגים הנפוצים, כלומר להתעלם מהתגים שהופיעו פחות מ-20 פעמים בקורפוס האימון (בסביבות ה-80 תגים). על מנת להעריך האם הייתה לכך השפעה משמעותית על התוצאות השווינו את התוצאות של מודל ביגרם המורץ על כל התגים לתוצאות של מודל ביגרם המורץ רק על התגים הנפוצים, כפי שיפורט בחלק התוצאות.

כמו כן, על מנת לצמצם את זמן הריצה של מודל הטריגרם, מידת ההצלחה של האלגוריתם נבחנה על 100 משפטים בלבד. על מנת להעריך האם הייתה לכך השפעה משמעותית על התוצאות השווינו את התוצאות של מודל ביגרם שבו התוצאות נבחנו על 100 משפטים לתוצאות של מודל ביגרם שבו התוצאות נבחנו על 10% מהקורפוס, כפי שיפורט בחלק התוצאות.

תוצאות

להלן מובאות התוצאות הגולמיות של ההרצות השונות של האלגוריתם שערכנו. כל הרצה מאופיינת לפי המודל שלה (ביגרם או טריגרם), האם נעשה שימוש בכל התגים או רק בתגים הנפוצים כמתואר לעיל, איזה אחוז מהקורפוס שימש לאימון, וכמה משפטים שימשו למבחן.

פירוט התוצאות של כל הרצה כולל את אחוזי ההצלחה הכלליים על כל התגים, אחוזי ההצלחה הכלליים על מילים מוכרות (כלומר שהופיעו בקורפוס האימון), אחוזי הצלחה כלליים על מילים לא מוכרות, ואחוזי הצלחה על כל אחד משלושת העל-תגים כפי שפורטו לעיל.

1.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Bigram** | Total: **71.2** |
| Tags: **common** | Known: **79.1** |
| Training: **90%** | Unknown: **18.9** |
| Test sentences: **100** | Verbs: **52.8** |
|  | Nouns: **78.4** |
|  | Adjectives: **56.1** |

2.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Bigram** | Total: **72.9** |
| Tags: **common** | Known: **78.8** |
| Training: **95%** | Unknown: **15.7** |
| Test sentences: **100** | Verbs: **56.2** |
|  | Nouns: **76.4** |
|  | Adjectives: **65.4** |

3.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Trigram** | Total: **58.1** |
| Tags: **common** | Known: **65.9** |
| Training: **90%** | Unknown: **6.9** |
| Test sentences: **100** | Verbs: **31** |
|  | Nouns: **37.5** |
|  | Adjectives: **48.8** |

4.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Trigram** | Total: **59.3** |
| Tags: **common** | Known: **64.9** |
| Training: **95%** | Unknown: **5.2** |
| Test sentences: **100** | Verbs: **33.6** |
|  | Nouns: **33** |
|  | Adjectives: **64.1** |

5.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Bigram** | Total: **71.4** |
| Tags: **all** | Known: **79.4** |
| Training: **90%** | Unknown: **18.6** |
| Test sentences: **100** | Verbs: **52.8** |
|  | Nouns: **78.4** |
|  | Adjectives: **56.1** |

6.

|  |  |
| --- | --- |
| **Variables** | **Success Percentage** |
| Model: **Bigram** | Total: **72.5** |
| Tags: **all** | Known: **79.6** |
| Training: **90%** | Unknown: **17.6** |
| Test sentences: **10%** | Verbs: **53.1** |
|  | Nouns: **79.2** |
|  | Adjectives: **62.4** |

ניתוח

על מנת לוודא שההגבלות הטכניות שאילצו אותנו לצמצם את היקף הבדיקה במודל הטריגרם לא פגעו באמינות התוצאות, התחלנו מהשוואה שלהם במסגרת מודל הביגרם. השוואה של מודל ביגרם בו נבחנו כל התגים (טבלה 5) למודל בו נבחנו התגים הנפוצים (טבלה 1) מראה הבדל מינימלי בתוצאות, כך שהגיוני להניח שגם עבור מודל טריגרם לא הינו רואים הבדל משמעותי אילו היינו בוחנים את כל התגים. השוואה של מודל בו המבחן היה על 10% מהקורפוס כמקובל (טבלה 6) למודל שבו המבחן היה על 100 משפטים בלבד (טבלה 1) מראה גם היא הבדלים קטנים בלבד.

לסיכום, סביר להניח כי האילוצים הטכניים לא גרמו לשינוי משמעותי בתוצאות שהתקבלו למודל הטריגרם. עם זאת, לצורך ההשוואה המדויקת ביותר בין מודל הביגרם למודל הטריגרם, גם עבור מודל הביגרם נתייחס לתוצאות שהתקבלו תוך שימוש רק בתגים הנפוצים ועבור 100 משפטי מבחן.

משתנה נוסף שיש לקחת בחשבון הוא גודל קורפוס האימון. שוב, ממגבלות טכניות נאלצנו להסתפק בקורפוס לא מאוד גדול של כ-4,600 משפטים, שמשמעותו שקורפוס האימון בגודל של 90% מהקורפוס כולו הכיל כ-4,100 משפטים. על מנת לקבל מושג לגבי ההשפעה של גודל קורפוס האימון על התוצאות, עבור כל מודל ביצענו בדיקה נוספת שבה גודל קורפוס האימון היה 95% מהקורפוס כולו, כלומר כ-4,300 משפטים. להוספה הזאת של כ-200 משפטי אימון הייתה השפעה מועטה ברוב המדדים, הן במודל הביגרם (טבלאות 1 ו-2) והן במודל הטריגרם (טבלאות 3 ו-4). ההבדל המשמעותי היחיד שניתן לראות הוא שההוספה של 5% לקורפוס האימון הביאה לשיפור של כ-10% במידת ההצלחה של האלגוריתם בזיהוי שמות תואר, בשני המודלים.

מהשוואה של אחוזי ההצלחה של מודל הביגרם ומודל הטריגרם ניתן לראות כי באופן כללי, מודל הביגרם מדויק הרבה יותר – אחוזי ההצלחה הכלליים שלו גבוהים בערך ב-10% עד 15%, ובאופן ספציפי אחוזי ההצלחה עבור שלושת ה-POS הנבדקים גבוהים אפילו יותר: כ-25%-30% עבור שמות עצם ופעלים. ההבדל אינו משמעותי במידה דומה עבור שמות תואר, שעבור המודלים של 90% קורפוס אימון מראים הבדל של 8% בלבד (לטובת מודל הביגרם), ועבור המודלים של 95% קורפוס מראים הבדל זניח של אחוז אחד בלבד.

שתי ההשוואות האלו, בין שני גדלי קורפוס האימון ובין שני המודלים, מביאות למספר מסקנות בנוגע ל-POS שבחנו. המסקנה הראשונה, העולה כבר מהניתוח לעיל, היא שנראה שהיכולת לזהות שמות תואר תלויה יותר בגודל קורפוס האימון, ופחות במודל שהורץ. במילים אחרות, נראה שהיכולת לזהות שמות תואר תלויה פחות בהסקה הסתברותית על סמך הסביבה בה הם מופיעים, ויותר ביכולת לזהות מילים בודדות כשמות תואר. נראה כי תוצאות כאלו הגיוניות – בשונה משמות עצם ומפעלים, באנגלית הסביבה התחבירית בה מופיעים שמות תואר לא מאוד מגוונת, ברוב הגדול של המקרים אך ורק כחלק מ-NP, וכמעט תמיד בתחילת הצירוף, כלומר אחרי ה-article. בדיקה ישירה של הסביבה בה מופיעים שמות תואר הראתה שהאינטואיציה הזאת נכונה, וכ-35% משמות התואר בקורפוס אכן הגיעו אחרי article ((וזאת לעומת שמות עצם למשל, אשר רק כ-20% מהם הגיעו אחרי article). המשמעות של סביבה "יציבה" בהקשר הזה היא שמספיק מעט מידע לגביה על מנת לזהות אותה, מה שיכול להסביר מדוע לא מצאנו הבדל גדול בביצועים של מודל ביגרם ומודל טריגרם בנוגע לשמות תואר – ייתכן כי המידע הנוסף לגבי הסביבה שמודל הטריגרם מספק אינה נחוצה על מנת לזהות שם תואר. מעבר לכך, העובדה כי הסביבה יציבה יחסית מנעה מהוספת המידע במודל הטריגרם להכניס רעש מיותר ולפגוע בביצועים של האלגוריתם, כפי שכנראה קרה עבור שמות העצם והפעלים.

המסקנה השנייה קשורה להבדל בביצועי מודלי הביגרם והטריגרם, כפי שהתבטאו בהצלחה היחסית בזיהוי שמות עצם ופעלים. כפי שציינו לעיל, הנחנו שההצלחה הפחותה משמעותית של מודל הטריגרם נובעת ממספר שילובי התגים הרב שהוא מאפשר, מה שמקשה על האלגוריתם למצוא סביבה אינדיקטיבית לכל POS (ייתכן כי אם גודל הקורפוס היה 4 מיליון משפטים ולא 4,000 אז מודל הטריגרם היה מצליח ליצור סטטיסטיקה אמינה יותר עבור הסביבות האינדיקטיביות, אך לא היה באפשרותנו להגדיל את הקורפוס על מנת לבחון זאת). על מנת לבחון השערה זאת בדקנו באופן ישיר מה הן הסביבות הנפוצות ביותר בהן מופעים שמות עצם ופעלים, עבור שני המודלים. במודל הביגרם, שתי סביבות נפוצות הופיעו יותר מ-10% כל אחד לפני פעלים, שמות עצם בשיעור של כ-20% והמילית "to" בשיעור של כ-11%. עבור שמות עצם, 4 סביבות הופיעו יותר מ-10% כל אחת, שמות עצם אחרים בשיעור של כ-22%, articles בשיעור של כ-20%, שמות תואר בשיעור של כ-12% ומילות יחס בשיעור של כ-11% (נדגיש שוב כי בהקשר זה כאשר אנחנו מתייחסים ל-"פעלים", "שמות עצם" ו-"שמות תואר" הכוונה היא לעל-תגים שיצרנו, וכי איחוד תגים שכזה לא בא לידי ביטוי מבחינת אופן הפעולה של האלגוריתם). נתונים אלו כשלעצמם עוזרים להסביר את אחוזי ההצלחה הגבוהים יותר בזיהוי שמות עצם במודל הביגרם – באופן כללי רוב שמות העצם מופיעים במספר קטן יותר של סביבות מאשר פעלים, כך שלאלגוריתם פשוט יותר למצוא את הסביבה האינדיקטיבית שלהם. לעומת זאת, במודל הטריגרם שתי הסביבות הנפוצות ביותר לפני פועל הן השילוב “article noun” ו-“noun noun” בשיעור של כ-5% כל אחת, והסביבות הנפוצות ביותר לפני שם עצם הן “preposition article” בשיעור של כ-9%, ו-“noun preposition” בשיעור של כ-6%. מכך ניתן לראות כי מודל הטריגרם אכן מוסיף מספר גדול של סביבות אפשריות, אשר מקשות על האלגוריתם לקבוע מה הוא ה-POS של מילה על פי הסביבה. לצורך ההשוואה, שתי הסביבות הנפוצות ביותר עבור שמות תואר במודל הטריגרם הן “preposition article” בשיעור של כ-15% ו-“verb article” בשיעור של כ-7%, כאשר סביבות אחרות מופיעות בשיעורים נמוכים בהרבה של כ-2% כל אחת.

סיכום

מתוך ההשוואה שערכנו בין תוצאות מודל הביגרם ומודל הטריגרם עבור ה-POS השונים, ניתן להגיע לכמה מסקנות בנוגע לסביבות האינדיקטיביות של ה-POS שנבדקו. ככל הנראה שימוש בקורפוס גדול יותר היה משפר את התוצאות של האלגוריתם, אבל אנחנו מאמינים שגם במסגרת המצומצמת של העבודה ניתן לראות את המגמות הכלליות. אחוזי ההצלחה הגבוהים ביותר התקבלו עבור שמות עצם במודל הביגרם, גבוהים בהרבה מאשר אחוזי ההצלחה המקבילים במודל הטריגרם, מה שיכול לרמז על כך שהסביבות בהן שמות עצם מופיעים נוטות להיות מגוונות למדי, כך שהסביבה האינדיקטיבית ביותר עבור שמות עצם היא המילה הקודמת להם בלבד. אחוזי ההצלחה שהתקבלו עבור פעלים היו נמוכים יותר אך הראו מגמה דומה של הצלחה גבוהה יותר עבור מודל הביגרם, ומכך ניתן להסיק כי גם פעלים נוטים להופיע בסביבות מגוונות, אפילו יותר משמות עצם. שמות התואר לעומת זאת הראו מגמה שונה – נראה כי לא קיים הבדל משמעותי בין הסביבה של שמות תואר ברמת הביגרם לסביבה ברמת הטריגרם. תוצאה זו יכולה להעיד על כך שהסביבה בה שמות תואר נוטים להופיע יציבה במידה דומה הן ברמת הביגרם והן ברמת הטריגרם, כך שאורך הסביבה האינדיקטיבית אינו גורם משמעותי בזיהוי שמות תואר. מהתוצאות שקיבלנו ניתן להסיק כי גודל הקורפוס משפיע יותר על היכולת של האלגוריתם לזהות שמות תואר, כך שעל מנת להגיע לתוצאות חד-משמעיות יותר יהיה צורך לערוך בדיקה דומה על קורפוס גדול יותר.

מהמחקר שערכנו לא בהכרח ניתן להסיק מסקנות לגבי מערכת קוגניטיבית מקבילה המאפשרת להבין POS של מילים דו-משמעותיות, אך הוא מראה כי לפחות ברמה התיאורטית ניתן להסיק מסקנות לגבי ה-POS של מילה מתוך ה-POS שנמצאים בסביבה האינדיקטיבית שלה. מבחינה מעשית יותר, מחקר נוסף באותו הכיוון יוכל לעזור לגלות את ההבדלים בסביבות האינדיקטיביות של POS רבים אחרים, ובכך יוכל לעזור לפתח אלגוריתמים מתוחכמים יותר שמתייחסים באופן שונה לכל POS על מנת למקסם את היעילות שלהם.