Avraham Asraf : 315774570

Schnaidman Elchanan: 316092436

עיבוד שפה טבעית

1. לאחר הפעלת אלגוריתם ויטרבי נקבל את המסלול הבא: LHHHLHHL

A screen shot of a computer

Description automatically generatedבגרף המצורף מתואר האלגוריתם, המסלול המקסימלי מצורף בירוק:

2. נצייר איך זה נראה בתצורה גרפית קודם:

נניח שיש , כלומר ממקסם את הסתברות ה-emission עבור . נסמן את זה . כלומר, ה- שממקסם את הסתברות ה-emission עבור יהיה ועבור יהיה וכן הלאה.

הרציונל שעומד מאחורי לקיחת רק קשתות עם משקלים ולא להוסיף מצבים וקשתות עם , זה מפני (אם נסתכל רק על המעבר השני, בה"כ) יהיו לנו 2 מצבים שהם זהים אך עם מעברי משקלים שונים אליהם: וקשת אחת מ- עוברת עם המשקל אל **הראשון** וקשת נוספת מאותו מקור אל **השני** עם משקל .

נניח בשלילה שקיים מסלול אופטימלי שעובר דרך . מכיוון שמעברי ההמשך אחרי תלויים במצב (כי המצב שהוא מייצג את התג הנוכחי ועוד 3 תגי עבר הוא זהה), לכן נוכל לבנות גם מסלול שעובר את אותה דרך עד אך ממשיך ל- ואז ממשיך למצבים מהמסלול האופטימלי של ההנחה.

מכיוון שלפי הגדרה שלנו *אזי*

*שזה אומר שהמשקל של הקשת שבחרנו להגעה למצב*  *הוא יותר גדול. את הקשת הבאה לחזור למסלול האופטימלי היא תהיה עם אותו משקל כי הסתברות ה-emission נוכל לבחור זהה והסתברות המעבר זהה כי המצבים הקודמים זהים (מה שהשתנה זה הסתברות ה-emission.*

*קיבלנו מסלול עם הסתברות יותר אופטימלית, לכן בהכרח נוכל להוריד מצבים עם הסתברות emission אליהם מהצורה .*

*אזי, קיבלנו את הייצוג הגרפי הנ"ל ובעצם נצטרך להריץ ויטרבי עד שנאסוף מספיק מילים שנגיע ל-. (כלומר, איטרציות).*

*אם כן, האלגוריתם יהיה באופן הבא:*

**Input**: an integer , parameters and .

**Definitions**: Define to be the set of possible tags. Define , and for . Define to be the set of possible words.

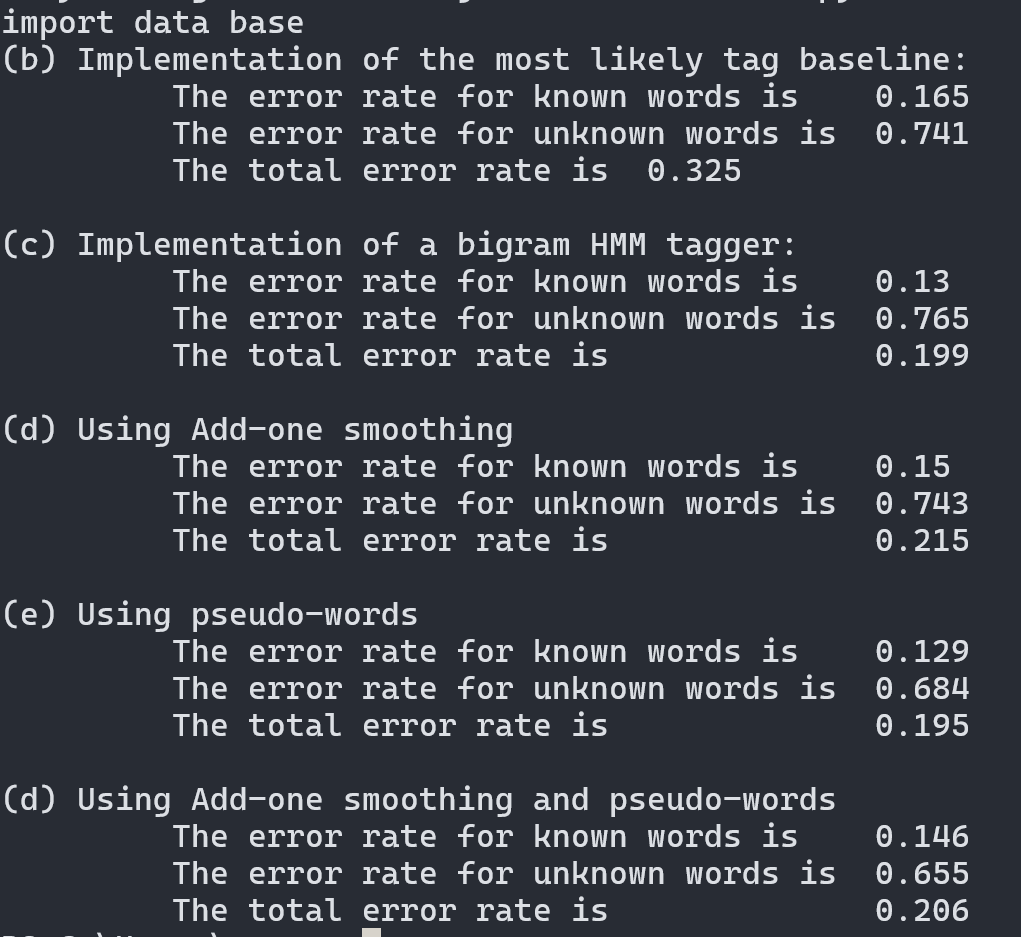
**Initialization**: set

**Algorithm**:

* For :
  + For :

**Return**:

תוצאות קטע הקוד:



המגמות אותם ניתן לראות:

1. מודל הHMM משפר משמעותית את החיזוי.

2. add-one-smooting משפר את החיזוי עבור מילים שלא ניראו, אך מוריד את הדיוק עבור מילים שכבר נראו, ואף באופן כללי.

3. pseudo-words משפרים באופן משמעותי יותר את השגיאה עבור מילים שלא ראינו, אך השפעתם על המילים שכבר ראינו היא אפסית. זאת גם הגישה שמשיגה את השגיאה הכללית הנמוכה ביותר.

4. שימוש גם בAdd-one smoothing וגם pseudo-words משיג את התוצאה הטובה ביותר עבור מילים שלא ניראו, אך לא את באופן כללי.

נעיר שקיימות טכניקות בסיסיות שהיו יכולות לשפר משמעותית את המודל עבור מילים שלא קיימות באוצר המילים, כמו למשל – הורדת תחילית וסופית של מילה, חיפוש שלא תלוי בדיקת כל המילים בlowercase. אך אנחנו נשארנו במסגרת הטכניקות שהוצגו בתרגיל.