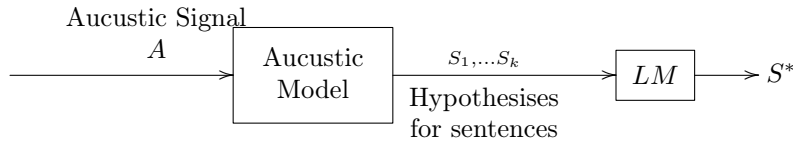


## דוגמאות לשימוש במודל שפה במערכות יישומים

### דוגמה 1 - זיהוי דיבור



$$P(S|A) = \frac{P(A|S)P(S)}{P(A)}$$

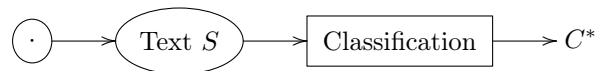
$$S^* = \operatorname{argmax}_S P(S|A) = \operatorname{argmax}_S \underbrace{P(A|S)}_{\text{Acoustic Model}} \cdot \underbrace{P_{LM}(S)}_{LM}$$

כ"ל לתרגום:  $A \equiv$  שפת המקור

$P(A|S) \equiv$  מודל תרגום משפה  $A$  לשפה  $S$

צריך לבחור מודל יחיד - תצפית לא וודאית ממספר היפותזות

### דוגמה 2 - סיווג טקטים



## החלקות - Smoothing

### מודל n-gram: קירוב מרקובי

הנחת קירוב מרקובי - ההסתברות למאורע תלויה רק במספר מובל של מאורעות אחורה:

$$p(w_i | w_1^{i-1}) = p(w_i | w_{i-n+1}^{i-1})$$

במודל n-gram יש פרמטר לכל ה- $n$  סדורה של מילים שמגדיר את ההסתברות המותנית המתאימה (של מילה בהינתן  $n-1$  המילים הקודמות)

לדוגמה: ב-3-gram מסתכלים על שתי מילים אחורה - כלומר במודל יש פרמטר לכל שלשת מילים - לדוגמה עבור הטקסט I ate a peach יש פרמטרים כמו  $p(\text{peach} | \text{ate}, a)$   $p(a | I, \text{ate})$

נסמן:  $V$  - לקסיקון. מספר הפרמטרים:  $|V|^n$

הערות מימוש: • בפועל נשמור פרמטרים רק ל- $n$  מילים שצפנו במדגם, וזה תת-לינארי בגודל המדגם.

• יוריסטיקה - בד"כ כאשר השפה מאוד גדולה, נתעלם מרצפים שראינו רק פעם אחת.

נשים ♡: המודל מגדיר התפלגות מולטינומית נפרדת מעל  $V$  לכל סדרת מילים מתנה באורך  $n-1$  (היא -  $n-1$ ) - שה"כ  $|V|^{n-1}$  (התפלגויות שונות)

### אומדן MLE לפרמטרים

$$p_{MLE}(w_i | w_{i-n+1}^{i-1}) = \frac{\#(w_{i-n+1}^i)}{\#(w_{i-n+1}^{i-1})}$$

(# - מונה)

לדוגמה - עבור I ate a peach

$$p_{MLE}(\text{peach} | \text{ate}, a) = \frac{\#(\text{times seen "ate a peach"})}{\#(\text{times seen "ate a *"})}$$

נשים ♡ אם ה- $n$  לא נצפתה, אזי נשערך לה הסתברות 0 (נניח בשלב זה שה- $w_{i-n+1}^{i-1}$  נצפו). מצב זה יוצר בעיה כאשר יתכנו מאורעות יחסית נדירים, שלא נצפו במדגם, אבל ההסתברות האמיתית שלהן עדיין חיובית.

בגלל הצורה הכפלית של מודל n-gram מספיק שלא נצפה ב- $n$  אחת ונקבל  $P(S) = 0$

הפתרון לבעיה הזו - החלקה

### החלקה - Discounting

במרחב מסוג מודל שפה, שבו יש הרבה מאורעות נדירים שמאפשרים אך לא ייצפו במדגם נתון, נרצה לבצע Discounting - לתת אומדן חיובי למאורעות שלא נצפו, ולמאורעות שנצפו לתת אומדן נמוך מאומדן ה- $MLE$ .

נזכור: באופן יחסי - האומדנים אמינים יותר למונים גבוהים, ועבורים יהיה נכון לעשות הפחתה יחסית קטנה יותר לאומדן  $MLE$

## מודל החלקה ראשון - Lidstone

נקודת מבט של הוספת קבוע  $\lambda$  למונה של כל מאורע אפשרי, ונרמול בהתאם:

$$p_{Lid}(x) = \frac{C(x) + \lambda}{|S| + \lambda|X|}$$

( $C$  - מונה)

- עבור  $\lambda = 1$ : נקרא החלקת Laplace
- כללית:  $\lambda$  היא פרמטר של השיטה

ההצדקה: אומדן לידסטון מוצדק כאומדן בייסיאני אם מניחים הסתברות א-פריורית אחידה לכל המאורעות

מתקיים: האומדן הוא אינטרפולציה לינארית בין התפלגות אחידה לאומדן MLE:

$$p_{Lid}(x) = \mu \frac{C(x)}{|S|} + (1 - \mu) \frac{1}{|X|}$$

עבור

$$\mu = \frac{|S|}{|S| + \lambda|X|}$$

$\mu$  הוא הפרופורציה של  $|S|$  מתוך "המדגם המוגדל"

שימושיות: שיטה מאוד פשוטה ונפוצה לשימוש

מוצדקת: כשיש prior אחיד

מצד שני: תיתן אומדנים מוטים במקרים אחרים - שיכולים לגרום נזק(בפועל - כתלות ברגישות המערכת הכוללת להחלקה)

בפרט - נראה בעייתיות של מודל Lidstone ליישומים שבהם נרצה התנהגות Discounting

ובפרט נראה שהחלקת Lidstone מגדילה את האומדן למאורעות מסויימים ש**נצפו** ביחס לאומדן MLE. כיוון שמתנהגם כאינטרפולציה לינארית - ערך  $P_{Lid}$  יהיה "באמצע"(מוטה ע"י מקדם האינטרפולציה) בין  $P_{MLE}$  להתפלגות יוניפורמית.

נשים  $\heartsuit$ : אם  $C(x) = \frac{|S|}{|X|}$  - כלומר שכיחות  $x$  שווה לשכיחות הממוצעת במדגם - אזי  $p_{MLE}(x) = \frac{1}{|X|}$   
לכן: עבור  $C(x) < \frac{|S|}{|X|}$  נקבל אומדן MLE קטן מאומדן התפלגות יוניפורמית:

$$C(x) < \frac{|S|}{|X|} \implies p_{MLE}(x) = \frac{C(x)}{|S|} < \frac{\frac{|S|}{|X|}}{|S|} = \frac{1}{|X|}$$

מסקנה - ל $x$ ים כנ"ל יתקיים:

$$p_{Lid}(x) > p_{MLE}(x)$$

בניגוד להתנהגות מצופה משיטת discounting