Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Αλληλεπίδραση Ανθρώπου – Υπολογιστή, 2024–25
Διδάσκων: Ι. Ανδρουτσόπουλος

Ασκήσεις μελέτης της ενότητας Β2 (γλωσσικά μοντέλα *n*-γραμμάτων, διόρθωση και πρόβλεψη κειμένου)

1. Επιβεβαιώστε τους υπολογισμούς της διαφάνειας «Υπολογισμός απόστασης Levenshtein». Δοκιμάστε και άλλα ζεύγη λέξεων. Υλοποιήστε τον αλγόριθμο. Υπόδειξη: Χρειάζονται απλά δύο φωλιασμένοι βρόχοι επανάληψης.

2. (α) Χρησιμοποιώντας τις παρακάτω προτάσεις ως (μικροσκοπικό) σώμα κειμένων εκπαίδευσης:

<start> he plays football
<start> he plays cricket
<start> she enjoys good football
<start> she plays good music
<start> he prays to god
<start> please buy me the other ball
<start> he pleases the other players by playing good football

εκτιμήστε τις πιθανότητες $P(t_1^4)$ που θα επέστρεφε ένα γλωσσικό μοντέλο διγραμμάτων με εξομάλυνση Laplace για κάθε μία από τις δύο παρακάτω προτάσεις t_1^4:

t_1^4: <start> he please god football
t_1^4: <start> he plays good football

Υποθέστε ότι το λεξιλόγιο V περιέχει όλες τις λέξεις του σώματος κειμένων (εξαιρώντας το <start>), οπότε $|V| = 21$. Δείξτε λεπτομερώς τους υπολογισμούς σας, χωρίς να εκτελέσετε τις τελικές αριθμητικές πράξεις.

Απάντηση:

Το γλωσσικό μοντέλο διγραμμάτων εκτιμά την *P*(<start>, he, please, god, football) ως εξής:

P(he | <start>) *P*(please | he) *P*(god | please) *P*(football | god) =

(4+1)/(7+21) (0+1)/(4+21) (0+1)/(1+21) (0+1)/(1+21)

όπου χρησιμοποιήσαμε εξομάλυνση Laplace.

Ομοίως για την *P*(<start>, he, plays, good, football). (Κάντε τους υπολογισμούς μόνοι σας.)

Σημείωση: Στην πράξη αποφεύγουμε διαδοχικούς πολλαπλασιασμούς πιθανοτήτων, γιατί συχνά οδηγούν σε πολύ μικρούς αριθμούς που δεν μπορούν να παρασταθούν με ακρίβεια στον υπολογιστή. Γι' αυτό υπολογίζουμε συνήθως τον λογάριθμο της πιθανότητας μιας ακολουθίας λέξεων, δηλαδή θα υπολογίζαμε το log *P*(<start>, he, please, god, football), αντί του *P*(<start>, he, please, god, football), οπότε θα καταλήγαμε στο παρακάτω άθροισμα τεσσάρων λογαρίθμων, αντί του παραπάνω γινομένου τεσσάρων πιθανοτήτων.

log[(4+1)/(7+21)] + log[(0+1)/(4+21)] + log[(0+1)/(1+21)] + log[(0+1)/(1+21)]

(β) Υποθέστε ότι ένας χρήστης έγραψε στο πληκτρολόγιο του κινητού του την παρακάτω ακολουθία λέξεων w_1^4:

w_1^4: <start> he pls gd ftball

Εκτιμήστε τις πιθανότητες $P(t_1^4|w_1^k)$ των δύο υποθέσεων (ακολουθιών λέξεων που ίσως ήθελε να γράψει) t_1^4 του σκέλους (α), χρησιμοποιώντας ένα μοντέλο θορυβώδους καναλιού (βλ. σχετικές διαφάνειες) και το γλωσσικό μοντέλο διγραμμάτων του σκέλους (α). Θεωρήστε ότι $P(w_i|t_i) \cong \frac{1}{LD(w_i,t_i)+1}$, όπου $LD(w_i, t_i)$ η απόσταση Levenshtein από τη λέξη w_i στην t_i. Δείξτε λεπτομερώς τους υπολογισμούς σας, χωρίς να εκτελέσετε τις τελικές αριθμητικές πράξεις και χωρίς να υπολογίσετε τις αποστάσεις Levenshtein.

Απάντηση: Χρησιμοποιώντας το θορυβώδες κανάλι των διαφανειών της διάλεξης, έχουμε:

$P(t_1^4|w_1^4) = P(t_1^4)\ P(w_1^4|t_1^4)\ /\ P(w_1^4)$

Για t_1^4= <start> he please god football:

P(<start>, he, please, god, football)
P(<start> he pls gd ftball | <start>, he, please, god, football) / $P(w_1^4)$

Η πιθανότητα *P*(<start>, he, please, god, football) εκτιμάται από το γλωσσικό μοντέλο, όπως στο σκέλος (α).

Χρησιμοποιώντας την προσέγγιση $P(w_i|t_i) \cong \frac{1}{LD(w_i,t_i)+1}$ της εκφώνησης (βλ. και διαφάνειες), η πιθανότητα *P*(<start> he pls gd ftball | <start>, he, please, god, football) γίνεται:

P(he, he) *P*(pls, please) *P*(gd, god) *P*(ftball, football) =

1/(*LD*(he, he)+1) 1/(*LD*(pls, please)+1) 1/(*LD*(gd, god)+1) 1/(*LD*(ftball, football)+1)

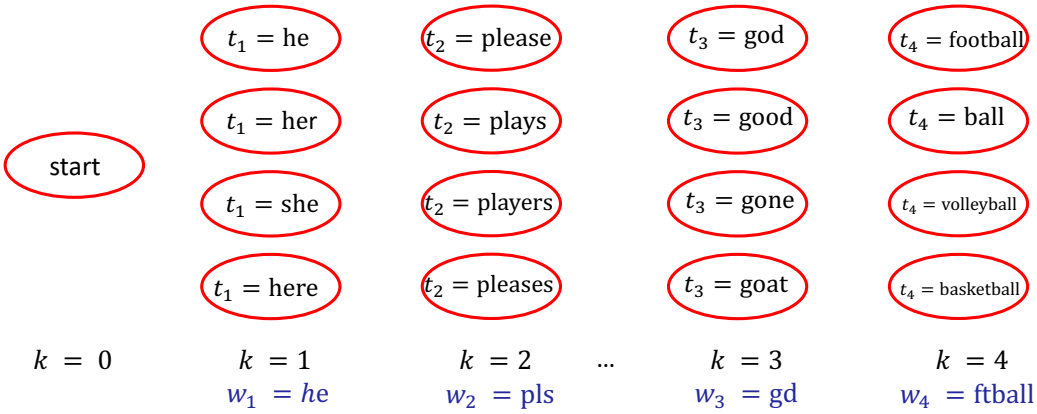
Η πιθανότητα $P(w_1^4)$ δεν χρειάζεται να εκτιμηθεί, γιατί είναι ίδια και για τις δύο υποθέσεις t_1^4 = <start> he please god football και t_1^4 = <start> he plays good football.

Ομοίως εκτιμούμε την πιθανότητα $P(t_1^4|w_1^4)$ για την υπόθεση t_1^4 = <start> he plays good football. (Κάντε τους υπολογισμούς μόνοι σας.) Επιλέγουμε τελικά την υπόθεση t_1^4 με το μεγαλύτερο $P(t_1^4|w_1^4)$.

Σημείωση: Και πάλι στην πράξη θα υπολογίζαμε τον λογάριθμο κάθε γινομένου πιθανοτήτων, οπότε θα καταλήγαμε σε αθροίσματα λογαρίθμων πιθανοτήτων, αντί γινόμενα πιθανοτήτων.

(γ) Εξηγήστε αναλυτικά πώς θα γινόταν η αποκωδικοποίηση με beam search (διαφάνειες «Beam search decoder»), αν ο χρήστης γράψει στο πληκτρολόγιο την ακολουθία λέξεων w_1^4 του σκέλους (β). Χρησιμοποιούμε πάλι το γλωσσικό μοντέλο διγραμμάτων λέξεων του σκέλους (α), εκπαιδευμένο στο μικροσκοπικό σώμα εκπαίδευσης εκείνου του σκέλους, μαζί με το μοντέλο θορυβώδους καναλιού του σκέλους (β). Θεωρήστε ότι το πλέγμα (lattice) αναζήτησης είναι το ακόλουθο, δηλαδή περιλαμβάνει 4 κοντινές (κατά απόσταση διόρθωσης) υποψήφιες σωστές λέξεις (του λεξικού), για κάθε λέξη w_i που έχει γράψει ο χρήστης. Σε κάθε βήμα του beam search, κρατάμε τα $b = 2$ καλύτερα μονοπάτια.

Beam search decoder



Για $k = 1$, τα υποψήφια μονοπάτια t_1^k είναι τα ακόλουθα. Δίπλα στο καθένα, υπολογίζουμε την ποσότητα $P(t_1^k)P(w_1^k|t_1^k)$. (Στη πραγματικότητα θα υπολογίζαμε τον λογάριθμο αυτής της ποσότητας.) Τα δύο καλύτερα μονοπάτια σημειώνονται με «**».

<start, he>: P(he|start) P(he|he) = (4+1)/(7+21) 1/(0+1) = 5/28 = 0.179 **
<start, her>: P(her|start) P(he|her) = (0+1)/(7+21) 1/(1+1) = 1/28 1/2 = 0.018
<start, she>: P(she|start) P(he|she) = (2+1)/(7+21) 1/(1+1) = 3/28 1/2 = 0.054 **
<start, here>: P(here|start) P(he|here) = (0+1)/(7+21) 1/(2+1) = 1/28 1/3 = 0.012

Για $k = 2$, τα υποψήφια μονοπάτια t_1^k είναι τα ακόλουθα. Δίπλα στο καθένα, υπολογίζουμε πάλι την ποσότητα $P(t_1^k)P(w_1^k|t_1^k)$. Συμπληρώστε μόνοι σας τους υπολογισμούς. Σημειώστε πάλι τα δύο καλύτερα μονοπάτια με «**».

<start, he, please>: P(he|start) P(he|he) P(please|he) P(pls|please) = …
<start, he, plays>: P(he|start) P(he|he) P(plays|he) P(pls|plays) = …
<start, he, players>: P(he|start) P(he|he) P(players|he) P(pls|players) = …
<start, he, pleases>: P(he|start) P(he|he) P(pleases|he) P(pls|pleases) = …
<start, she, please>: P(she|start) P(he|she) P(please|she) P(pls|please) = …
<start, she, plays>: P(she|start) P(he|she) P(plays|she) P(pls|plays) = …
<start, she, players>: P(she|start) P(he|she) P(players|she) P(pls|players) = …
<start, she, pleases>: P(she|start) P(he|she) P(pleases|she) P(pls|pleases) = …

Για $k = 3$, τα υποψήφια μονοπάτια t_1^k είναι τα ακόλουθα. Δίπλα στο καθένα, υπολογίζουμε πάλι την ποσότητα $P(t_1^k)P(w_1^k|t_1^k)$. Συμπληρώστε μόνοι σας τους υπολογισμούς. Σημειώστε πάλι τα δύο καλύτερα μονοπάτια με «**».

Συμπληρώστε και τους υπολογισμούς για $k = 4$. Τελικά ποια υπόθεση $< \text{start}, t_1, t_2, t_3, t_4 >$ επιλέγεται;