תאריך : 3.4.2024

מבחן מועד א' בקורס עיבוד שפה טבעית (67658)

שנת לימודים התשפ"ד 2023/2024

משך המבחן : שעתיים

מרצה הקורס : פרופ' עמרי אבנד

השימוש בכל חומר עזר אסור, נא כתבו על הכריכה על אילו שאלות עניתם

חלק א' (80 נקודות): ענו על <u>בדיוק שתי שאלות</u> מבין שאלות 1-3

<u>שאלה 1 (40 נקודות):</u>

בשאלה זו נעסוק ב-Graph-based Dependency Parsing.

א. (12 נק') הגדירו באופן פורמלי את מודל ה-MST Parser עבור הבעיה. אין צורך לתאר את אלגוריתם האימון וההיסק.

תשובה:

Define a scoring function over all possible directed trees over $V = \{w_1, ..., w_n, ROOT\}$ where $ROOT$ is the root of the tree. Let $\Phi : V^2 \times L \times S \to \{0,1\}^d$, where L is the label set and S is the set of sentences (feature values can also be real numbers if needed), be a feature function over possible edges.

Let θ be the weight vector (the parameters of the model):

$$score_\theta(v_1, v_2, l \,|x_{1:n}) = \theta^t \cdot \Phi(v_1, v_2, l, x_{1:n})$$

For a directed tree T define:

$$score_\theta(T \,|x_{1:n}) = \sum_{(v_1,v_2,l)\in T} score_\theta(v_1, v_2, l \,|x_{1:n})$$

ב. (16 נק') עבור כל אחת מן התכוניות (features) הבאות, ציינו אם ניתן לייצג אותה במודל ה-MST Parser.

a. התכונית: מספר המילים (tokens) הכולל במשפט.

כן. חלק מהקלט הוא סדרת ה-x.

b. נגדיר אורך צלע בעץ בין ה-token ה-i ל-token ה-j במשפט להיות |i-j|. התכונית: אורך הצלע הממוצע בעץ.

לא. ה-score מוגדר על כל צלע בנפרד.

c. התכונית: חלק הדיבר הנפוץ ביותר במשפט (אם יש כמה נפוצים ביותר, בוחרים אקראית ביניהם).

כן. חלקי הדיבר הם לרוב חלק מהקלט, ומכיוון שפונקציות התכוניות מקבלת את כל המשפט, ניתן לבצע חישוב זה.

d. התכונית: העומק של העץ (כלומר, האורך המקסימאלי של מסלול מכוון בין השורש לקודקוד כלשהו).

לא. ה- score מוגדר על כל צלע בנפרד.

ג. (12 נק') נעסוק כעת במודל הבא המוגדר (גם) באמצעות תכוניות מסדר גבוה. בהינתן עץ T ללא תגיות (unlabeled) על קבוצת הקודקודים V, ובהינתן משפט $x_{1:n}$, וקטורי משקולות $\theta_1 \in R^d$, $\theta_2 \in R^l$

ופונקציות תכוניות ϕ_1, ϕ_2, נגדיר:

$$score(T) = \sum_{(i,j,k)\in V^3 | (i,j)\in T \& (i,k)\in T} score(i,j,k) + \sum_{(i,j)\in T} score(i,j)$$

$$score(i,j,k) = \theta_1^t \cdot \phi_1(i,j,k,x_{1:n})$$

$$score(i,j) = \theta_2^t \cdot \phi_2(i,j,x_{1:n})$$

נניח לצרכי שאלה זו שידוע לנו אלגוריתם יעיל (העובד בזמן פולינומיאלי בגודל הקלט) המקבל משפט $x_{1:n}$, וקטורי משקולות $\theta_1 \in R^d$, $\theta_2 \in R^l$, ופונקציות תכוניות ϕ_1, ϕ_2, ומוצא את העץ הפורש בעל ה-score המקסימלי תחת ההגדרה לעיל. הציעו וריאציה לאלגוריתם הלמידה שהצגנו עבור ה-MST Parser (כלומר ה-Perceptron) אשר מאפשר למידה יעילה של הפרמטרים של מודל זה. אין צורך להוכיח את אופטימליות האלגוריתם.

תשובה:

1. $\theta_1 \leftarrow 0$
2. $\theta_2 \leftarrow 0$
3. for $r = 1, ..., N_{iterations}$
4. for $i = 1, ..., N$
5. $T' \leftarrow \text{argmax}_T \quad score(T)$
6. $\theta_1 \leftarrow \theta_1 + \eta \left(\sum_{T_i} \phi_1(v_1, v_2, v_3) - \sum_{T'} \phi_1(v_1, v_2, v_3)\right)$
7. $\theta_2 \leftarrow \theta_2 + \eta \left(\sum_{T_i} \phi_2(v_1, v_2) - \sum_{T'} \phi_2(v_1, v_2)\right)$
8. return averaged θ_1 and θ_2

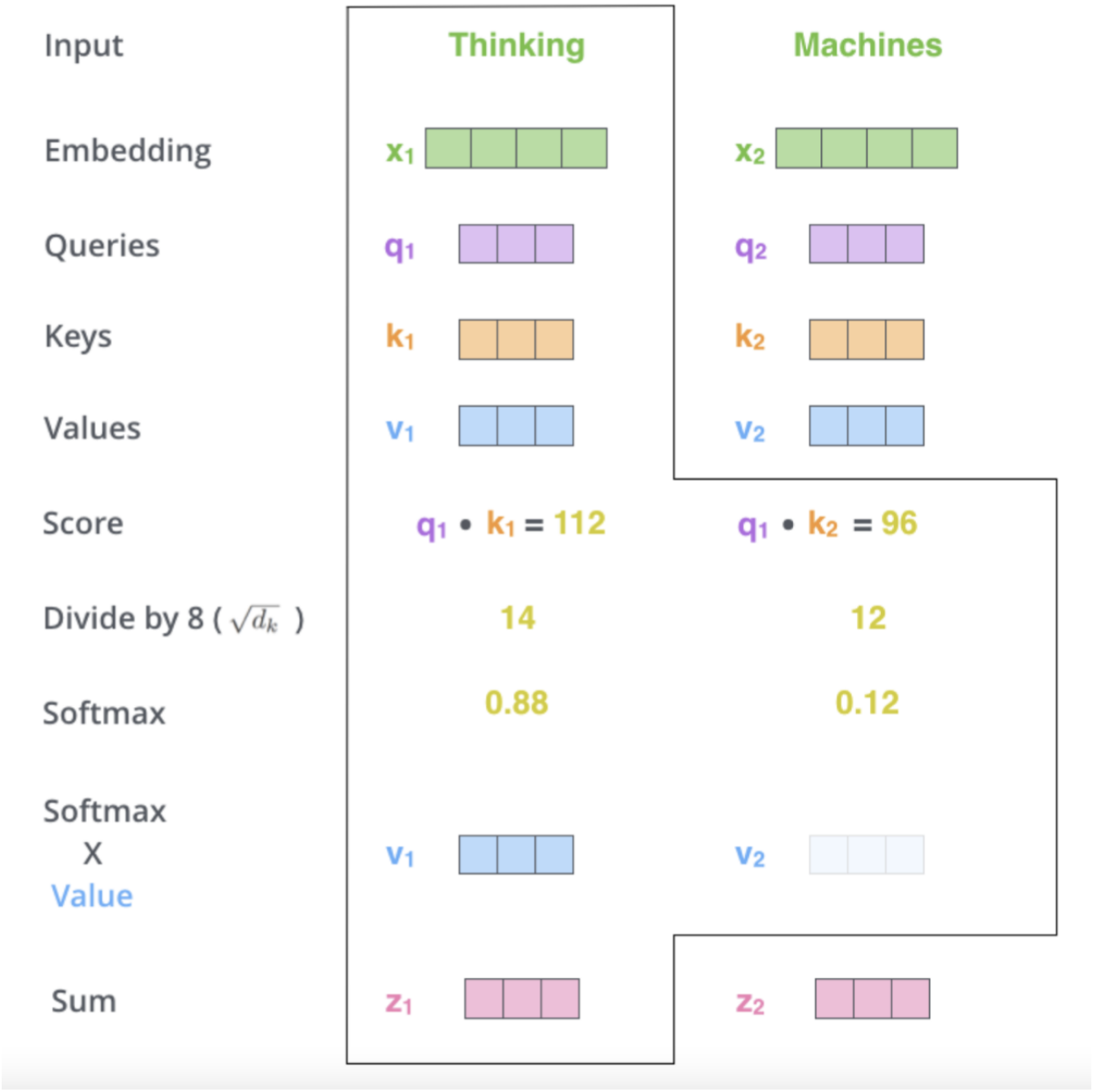
הערות:

- שורה 5 מתייחסת להפעלת אותו אלגוריתם שנתון לנו.

<u>שאלה 2 (40 נקודות):</u>

א. (20 נק') הסבירו בפירוט את החישוב המבצע מנגנון self-attention עבור מקרה של שני tokens. הכוונה למנגנון כפי שנלמד בכיתה עבור צד ה-encoder בארכיטקטורת ה-Transformer (אין צורך להגדיר multi-head attention). ציינו בתשובתכם מהו הקלט, מהו הפלט וכיצד הפלט מחושב מן הקלט.

תשובה:



ב. (20 נק') נניח שאנו עוסקים ב-Machine Translation באמצעות ארכיטקטורת ה-Transformer. על מנת לחסוך במספר פרמטרים, הוחלט לבטל את ה-positional encodings ב-encoder (אבל להשאיר אותם ב-decoder). נקרא לארכיטקטורה זו Transformer2. הסבירו מהו החיסרון ב-Transformer2 לעומת ה-Transformer הרגיל. נמקו באמצעות מתן דוגמא לזוג משפטים שלפחות עבור אחד מהם התרגום אמור לצאת לא מוצלח ב-Transformer2.

תשובה:

Transformer2 לא יהיה רגיש לסדר המילים בקלט. דוגמא: "כלב נשך אדם" ו-"אדם נשך כלב": לפחות אחד לא יתורגם נכון ב-Transformer2 כי מבחינתו הם שקולים.

שאלה 3 (40 נקודות):

א. (10 נק') הגדירו באופן פורמלי את מודל MEMM מסדר ראשון (bigram) ומודל CRF מסדר ראשון (bigram). אין צורך לתאר את אלגוריתם האימון וההיסק.

נניח כעת כי נתון מודל bigram CRF מאומן ומודל bigram MEMM מאומן. על מנת לחסוך בחישובים, הוצע לא לכלול את גורם הנרמול (ה-partition function) בחישוב סדרת התגיות הסבירה ביותר (inference). כלומר, בחישובים הנערכים על ידי אלגוריתם Viterbi לא נכלול את המכנה בחישוב משקלי הצלעות, אלא רק את המונה (כלומר את האקספוננט של ה-score).

ב. (10 נק') האם ההיסק יעבוד בצורה נכונה במקרה של MEMM? ובמקרה של CRF? כלומר, האם הוא יחזיר את סדרת התגיות בעלת ההסתברות המקסימאלית? נמקו את תשובתכם.

תשובה:

כן במקרה של CRF אבל לא במקרה של MEMM. הסיבה לכך היא שהנרמול ב-MEMM הוא שונה לכל בחירה של מצב, כך שהסרת גורם הנרמול יגרום לכך שקודקודים שסכום המשקלים של הצלעות היוצאות מהם גבוה יועדפו.

ג. (10 נק') במקרים שבהם ההיסק יעבוד בצורה נכונה, האם השינוי לעיל ישנה את סיבוכיות זמן הריצה האסימפטומטית של אלגוריתם Viterbi להיסק? נמקו את תשובתכם.

תשובה:

לא. או שמניחים מראש שלא צריך את גורם הנרמול, או שגם אם מחליטים שצריך, עדיין תכנון דינאמי באותו זמן ריצה עם חיבור במקום מקסימום יחזיר את ה-partition function.

ד. (10 נק') טענה: ניתן לייצג תכונית (feature) במודל bigram CRF אם ורק אם ניתן לייצג אותה במודל bigram MEMM.

האם הטענה נכונה או לא? נמקו את תשובתכם.

תשובה:

נכון וזאת בגלל שפונקצית התכוניות מקבלת בדיוק את אותם הארגומנטים בשני המקרים.

חלק ב' (20 נקודות): ענו על <u>בדיוק שאלה אחת</u> מבין שאלות 4-5

שאלה 4 (20 נקודות):

א. (10 נק') הגדירו באופן פורמלי מהי רשת נוירונים רקורנטית (recurrent neural network). התייחסו בתשובתכם לפונקציה הרקורנטית (recurrent function) ופונקצית הפלט (output function).

תשובה:

- Maps a sequence of vector inputs to a sequence of vector states

$$x_1, \ldots, x_n \rightarrow s_1, \ldots, s_n$$

- A (parametrized) transition function *R* does the mapping:

$$s_i = R(s_{i-1}, x_i)$$

- An output function *O* maps states to (vector) outputs, which are often viewed as a distribution over the possible labels

ב. (10 נק') הציגו מגבלה של מודלי שפה מרקוביים אשר למודלי שפה רקורנטי (recurrent language models) עשוי להיות יתרון יחסי בהתמודדות איתה. כללו בתשובתכם דוגמא.

תשובה:

מודלי שפה מרקוביים לא יכולים להתמודד עם תלויות מעבר ל-n צעדים (בתלות בסדר של המודל). מודלים רקורנטיים יכולים מכיוון שה-hidden state שלהם מבוסס על כל ה-context. דוגמא:

The meeting I was supposed to go to yesterday _____ (was/were)

The meetings I was supposed to go to yesterday _____ (was/were)

<u>שאלה 5 (20 נקודות):</u>

א. (10 נק') הגדירו באופן פורמאלי את גישת ה-linear interpolation :back off להחלקה של מודלי שפה מרקובים מסדר שני (trigram model).

תשובה:

$$\begin{aligned} q(w_i \mid w_{i-2}, w_{i-1}) = \quad & \lambda_1 \times q_{\mathsf{ML}}(w_i \mid w_{i-2}, w_{i-1}) \\ & +\lambda_2 \times q_{\mathsf{ML}}(w_i \mid w_{i-1}) \\ & +\lambda_3 \times q_{\mathsf{ML}}(w_i) \end{aligned}$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$, and $\lambda_i \geq 0$ for all i.

ב. (10 נק') הגדירו את בעיית האופטימיזציה אותה צריך לפתור על מנת למצוא את הערכים של מקדמי האינטרפולציה ($\lambda_1,\lambda_2,\lambda_3$). אין צורך לפרט כיצד מוצאים את הפתרון אלא רק את הגדרת הבעיה באופן מלא. הניחו בתשובתכם שיש רק סט אחד של מקדמים ($\lambda_1,\lambda_2,\lambda_3$). איזה מגבלה הנחה זו משיתה על המודל?

תשובה:

- That is, maximize *L*, the log-likelihood, where $c'(w_1,w_2,w_3)$ is the number of times the trigram (w_1,w_2,w_3) appeared in the validation set

$$L(\lambda_1, \lambda_2, \lambda_3) = log \left(\prod_{(w_1,w_2,w_3) \in V^3} q(w_3 | w_1, w_2)^{c'(w_1,w_2,w_3)} \right)$$

- Which yields this (concave) optimization problem:

$$L(\lambda_1, \lambda_2, \lambda_3) = \sum_{w_1,w_2,w_3} c'(w_1, w_2, w_3) \log q(w_3 \mid w_1, w_2)$$

such that $\lambda_1 + \lambda_2 + \lambda_3 = 1$, and $\lambda_i \geq 0$ for all i, and where

$$\begin{aligned} q(w_i \mid w_{i-2}, w_{i-1}) = \quad & \lambda_1 \times q_{\mathsf{ML}}(w_i \mid w_{i-2}, w_{i-1}) \\ & +\lambda_2 \times q_{\mathsf{ML}}(w_i \mid w_{i-1}) \\ & +\lambda_3 \times q_{\mathsf{ML}}(w_i) \end{aligned}$$

המגבלה היא שייתכן שהצורך באינטרפולציה שונה לפי השכיחות (לדוגמא, נרצה לתת יותר משקל לטריגרם עבור שלישיות שיש עבורן הרבה דאטה).

בהצלחה!