תאריך : 6.3.2019

מבחן מועד ב׳ בקורס עיבוד שפה טבעית (67658)

2018/2019 שנת לימודים התשע״ח

משך המבחן : שעתיים

מרצה הקורס : ד״ר 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-Relation Extraction, וספציפית בזיהוי אזכורים במסמך נתון D של שמות אנשים הנשואים או היו נשואים זה לזה.

א. (10 נק׳) הציעו מודל לזיהוי כל השמות הפרטיים של אנשים בטקסט המבוסס על Bigram Conditional Random Field Model (Bigram CRF). יש לרשום את מודל ה-CRF, ואת מרחב התוויות אליו ממפה המודל כל מילה.

ב. (10 נק׳) נאמן את האלגוריתם שהצעתם בסעיף 1, ונריץ אותו על D. נרצה כעת לייצר מודל תיוג (classifier) שיהיה מסוגל לקבל משפט, ושני שמות פרטיים של אנשים בתוך המסמך, ולהשיב ׳כן׳ אם האנשים מוזכרים במשפט כנשואים, ו-׳לא׳ אחרת. נניח גם שלא נתון לנו מידע מתויג שיכול לשמש כ-training data לבעיה זו. כיצד ניתן לייצר training data מקורב עבור הבעיה באמצעות Wikipedia?

ג. (10 נק׳) נשתמש בטכניקה לייצור training data מקורב שהצעתם בסעיף ב׳, ונאמן בעזרתה classifier. כעת נניח שכל ערך ב-Wikipedia שבו שני אנשים מוזכרים כנשואים, הם גם מוזכרים ככאלה שיצאו למסעדה יחד. באלו מקרים אנו מצפים שה-classifier המאומן יחזה ששני אנשים נשואים, למרות שלא אוזכרו ככאלה בטקסט? הסבירו את תשובתכם.

ד. (10 נק׳) הסבירו מהם Path Features וכיצד Semantic Role Labeling יכול לפתור את בעיית הדלילות (sparsity) בשימוש בהם.

<u>שאלה 2 (40 נקודות)</u> :

א. (10 נק׳) הגדירו פורמלית את שיטת Kneser-Ney עבור החלקה של מודל שפה מרקובי מסדר ראשון (Bigram Language Model). הגדירו את כל המושגים בהם השתמשתם.

ב. (15 נק׳) נתונות מילים w_1,w_2,w_3,w_4 ונתון כי כל המילים הופיעו בקורפוס C, אבל שאף זוג מילים מתוכן לא הופיעו צמודות. כעת נניח שבשימוש בהחלקת Kneser-Ney עבור מודל שפה מרקובי מסדר ראשון שאומן על הקורפוס C, ההסתברות המשוערכת של הופעת w_2 מיד אחרי w_1 (כלומר $P(w_2|w_1)$) היא 0.001 וההסתברות המשוערכת של הופעת w_3 מיד אחרי w_1 (כלומר $P(w_3|w_1)$) היא 0.01.
האם ניתן לקבוע אם ההסתברות המשוערכת עבור הופעת w_2 מיד אחרי w_4 גדולה יותר מההסתברות

המשוערכת עבור הופעת w_3 מיד אחרי w_4? הוכיחו את תשובתכם.

ג. (10 נק') הגדירו באופן פורמלי מהי רשת נוירונים רקורנטית (Recurrent Neural Network, RNN), והסבירו כיצד ניתן להגדיר באמצעות רשת כזאת מודל שפה.

ד. (5 נק') הסבירו כיצד מתמודד מודל שפה המבוסס על RNN עם הבעיה אותה החלקת Kneser-Ney מנסה לפתור.

<u>שאלה 3 (40 נקודות):</u>

א. (10 נק') הגדירו באופן פורמלי מהו מודל Probabilistic Context Free Grammar (PCFG).

ב. (10 נק') איזה הנחת אי-תלות מותנה מניח מודל PCFG? האם הנחה זו מתקיימת עבור השפה האנגלית? לוו את תשובתכם בדוגמא.

ג. (10 נק') נניח כעת שנתון לנו מודל PCFG לניתוח תחבירי של השפה האנגלית, ושערוך (כלומר ערך מספרי) עבור כל אחד מהפרמטרים של המודל. כמו כן, נניח שהמודל נתון בצורה של Chomsky Normal Form. כתבו פסאודו-קוד לאלגוריתם המקבל משפט כסדרה של מילים $w_1,\ldots,w_n$, ומחזיר את ההסתברות המקסימלית (ע"פ המודל) של עץ תחבירי עבור המשפט (כלומר של עץ תחבירי שעליו הם $w_1,\ldots,w_n$).

ד. (10 נק') שנו את האלגוריתם שהצעתם בסעיף ג' כך שבמקום להחזיר את ההסתברות המקסימלית לעץ תחבירי, הוא יחזיר את **סכום** ההסתברויות עבור כל העצים התחביריים עבור המשפט.

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

<u>שאלה 4 (20 נקודות):</u>

א. (10 נק') הסבירו כיצד מחשבים Precision, Recall ו-F1 בין שני עצים תחביריים מסוג constituency.

ב. (10 נק') מדוע לא ניתן להסתפק רק ב-Precision או רק ב-Recall כמדדי אבלואציה (evaluation)?

<u>שאלה 5 (20 נקודות):</u>

בשאלה זאת נעסוק ב-transition-based dependency parsing (ראו נספח). הוכיחו שבשימוש במערכת המעברים arc standard עבור יצירת עץ נתון, מספר המעברים עבור משפט בעל n מילים הוא בדיוק $2n$.

בהצלחה!

Appendix: The arc-standard transition system

Transition set $\mathcal{T}$:

SHIFT	move one item from the buffer to the stack: $(\Sigma,\ i\|B,\ A) \Rightarrow (\Sigma\|i,\ B,\ A)$
LEFT-ARC	create arc $j \rightarrow i$ and remove i: $(\Sigma\|i\|j,\ B,\ A) \Rightarrow (\Sigma\|j,\ B,\ A \cup \{(j,i)\})$ Condition: $i \neq 0$
RIGHT-ARC	create arc $i \rightarrow j$ and remove j: $(\Sigma\|i\|j,\ B,\ A) \Rightarrow (\Sigma\|i,\ B,\ A \cup \{(i,j)\})$

Initial configuration:

$$c_s(w_1, w_2, w_3, \ldots,) = ([ROOT], [1, 2, 3, \ldots], \emptyset)$$

Terminal configuration:

$$c_t = ([ROOT], [], A)$$

Legend:

Σ	stack
B	buffer
A	set of arcs constructed so far
i, j	two items at the top of the stack (j is the top)
$ROOT$	the root node of the tree

A configuration is written as (Σ, B, A).