קורס 22933 | מבוא לעיבוד שפה טבעית

תיוג חלקי דיבר בעברית | מטלה 12

מועד הגשה : 15 לאפריל 2018

הנחיות כלליות

- במטלה זו 5 שאלות עליהן יש לענות, ועוד שאלת בונוס שאינה חובה.
- את המטלה יש להגיש בזוגות, בפורמט ZIP באתר הקורס.
- שאלות הבהרה ושאלות לדיון ייענו רק דרך הפורום.
- את המטלה יש להגיש עד 15 באפריל, 23:59.

מבוא

במטלה זו נממש אלגוריתם סטטיסטי לתיוג חלקי דיבר עבור טקסטים בעברית. עברית הינה שפה שמית הידועה במורפולוגיה העשירה שלה, ובסדר מילים פחות קשיח מאשר בשפה האנגלית. זאת ועוד, לשפה העברית כמות קטנה לאין שיעור של מידע מתוייג מאשר זה הקיים לשפה האנגלית. כיצד ישפיעו אספקטים אלה על דיוק המתייג? מה הדיוק המיטבי אליו ניתן להגיע באמצעות מתודות קיימות? במטלה זו נחקור באופן אמפירי שאלות אלה.

קבצי המידע

המתייג שנממש במטלה זו מקבל אוסף משפטים בעברית כקלט ומחזיר כפלט את אותו אוסף משפטים מתויג בחלקי דיבר. בניגוד לשפה האנגלית, בה המתייג מקבל משפטים המוגדרים כרצפים של מילים כאשר כל מילה תואמת לחלק דיבר אחד, בעברית מודרנית סטנדרטית הקלט למתייג יהיה רצף של מילים היכולות להכיל יותר מחלק דיבר אחר, כגון שכשנבוא / ובבית / אהבתיה . הקבצים שתקבלו עברו כבר סגמנטציה כך שכל סגמנט בקלט תואם לחלק דיבר אחד: ש כש נבוא / ו ב ה בית / אהבתי את היא. הטקסטים העבריים כתובים באותיות לטיניות משמאל לימין, על מנת לפשט את העבודה איתם, ומניחים את המיפוי המוגדר במאמר treebank הזמין באתר המטלה, טבלא **11**. תאור התאגים ומשמעותם לעברית נמצא גם הוא בגוף המאמר ומומלץ לעיין בו. עבור מימוש המתייג נדרשים הקבצים הבאים, הזמינים להורדה באתר הקורס.

- heb-pos.train
- heb-pos.test
- heb-pos.gold

קבצי האימון והגולד זמינים בפורמט הבא: כל משפט מוגדר כרצף של סגמנטים מורפולוגים מתוייגים. הסגמנטים מופיעים בסדרם הכרונולוגי, סגמנט אחד לשורה, ולאחר כל סגמנט מופיע התג הנכון לו בהקשר, כאשר הוא מופרד בטאב. כל משפט מסתיים בשורה ריקה. שורות המכילות הערות מתחילות בסולמית. פורמט קבצי הבדיקה זהה לפורמט קבצי האימון, בהשמטת העמודה של התג (ובהשמטת הטאב). ראו דוגמאות באיורים בנספח.

מבנה הארכיטקטורה

על הארכיטקטורה המיועדת לכלול שלושה רכיבים המצייתים לפורמט ההרצה הבא

- ./train $< model >< heb\text{-}pos.train >< smoothing(y/n) >$
  Output: trained parameter files (including the estimated probabilities)

  ./decode $< model >< heb\text{-}pos.test >< param\text{-}file1 > [< param\text{-}file2 >]$
  Output: $< *.tagged >$

  ./evaluate $< *.tagged >< heb\text{-}pos.gold >< model >< smoothing(y/n) >$
  Output: $< *.eval >$

הפרמטר model יכול לקבל את הערכים הבאים

- 1 indicate the baseline (majority vote) tagger
  2 indicates a bi-gram tagger
  3 indicates a tri-gram tagger
  4 indicates ...

כל ניסוי מסתיים בהפקת שני הקבצים הבאים

- *.tagged
- *.eval

מדדי הערכה

הערכת איכות התוצאות תעשה באופן הבא
הניחו כי קובץ הבדיקה מכיל N משפטים
נסמן את המשפטים באופן הבא: $x = x_{1_j}...x_{n_j}$ כאשר $j = 1...N$
נסמן את התיוגים הנתונים בקובץ הגולד עבור משפט j באופן הבא: $(x_{1j}, t_{1j})...(x_{n_j}, t_{n_j})$
נסמן את התיוגים הנתונים בקובץ הפלט עבור משפט j באופן הבא: $(x_{1j}, t'_{1j})...(x_{n_j}, t'_{n_j})$
אורך כל משפט יסומן n_j עבור משפט j

- Calculating word accuracy for sentence j
  (proportion of correct tags in a sentence)

$$A_j = \frac{1}{n_j} \times \sum_{\{t'_{ij} | t_{ij} = t'_{ij}, i=1..n_j\}} 1$$

- Calculating sentence accuracy for sentence j
  (whether or not all tags are correct in a sentence)

$$All_j = 1 \ \textit{iff} \ (A_j == 1), \ \textit{otherwise} \ All_j = 0$$

- Calculating word accuracy for the test corpus $j = 1...N$
  (proportion of correct tags in the corpus)

$$A = \frac{\sum_{j=1}^{N} A_j \times n_j}{\sum_{j=1}^{N} n_j}$$

- Calculating sentence accuracy for the test corpus $j = 1..N$
  (proportion of correctly tagged sentences)

$$All = \frac{\sum_{j=1}^{N} All_j}{N}$$

שאלות

1. סטטיסטיקה תיאורית

הכירו את הקורפוס! ממשו תוכנית פשוטה המייצרת סטטיסטיקה תיאורית, והריצו אותה על המידע שבקובץ האימון ובקובץ הגולד שברשותכם. עבור כל קובץ ענו על השאלות הבאות:

1. כמה מופעי יוניגרם של סגמנטים יש בקורפוס ?
2. כמה סוגי יוניגררם של סגמנטים יש בקורפוס ?
3. כמה מופעי סגמנט־תג יש בקורפוס ?
4. כמה סוגי סגמנט־תג יש בקורפוס ?
5. מהו מדד העמימות עבור תיוג חלקי הדיבר בקורפוס ? כלומר, מהו הערך הממוצע של תגים שהופיעו פר סגמנט?
6. דונו בהבדלים בסטטיסטיקות בין קובץ האימון לקובץ הגולד, ונסו לשער ולהסביר כיצד הבדלים אלה עלולים או עשויים להשפיע על ביצועי המתייג

2. המתייג הבסיסי

ממשו מתייג בסיסי על פי המפרט הבא:

- כיתבו רכיב אימון המוצא עבור כל סגמנט את התג השכיח ביותר עבורו בקורפוס האימון
- כיתבו רכיב תיוג שבהנתן קובץ משפטים לא מתוייגים וקובץ הפרמטרים שנוצר באימון, מתייג כל סגמנט בתג הכי שכיח עבורו. מילים לא ידועות יתויגו כ NNP
- כיתבו רכיב הערכה המקבל קובץ מתויג וקובץ גולד ונותן כפלט את קובץ המדדים המתואר בהקדמה

ענו על השאלות הבאות:

1. הגדירו במדוייק את סכמת הפרמטרים במודל
2. הגדירו במדוייק את נוסחאות המשערכים של הפרמטרים במודל
3. :הגדירו במדוייק את סיבוכיות זמן הריצה של המודל
   - מהי סיבוכיות האימון?
   - מהי סיבוכיות התיוג?

   על הסיבוכיות להיות מוגדרת במונחי אורך המשפט וגודל סט הקטגוריות
4. דווח מהו הדיוק הכולל עבור תיוג קובץ הבדיקה (macro-avg)

3. מתייג מסדר ראשון

ממשו מתייג מבוסס שרשראות מרקוביות נסתרות HMM.

- ממשו תכנית train המקבלת קובץ אימון כמוגדר בהקדמה ומייצרת שני קבצי פרמטרים. קובץ פרמטרים לקסיקליים lex.* וקובץ פרמטרים תחביריים gram.* .

  - קובץ הפרמטרים הלקסקלי מכיל סגמנט יחיד בשורה. עבור כל סגמנט מצורפת רשימה של הקטגוריות האפשריות עבורו בקובץ האימון וערכי ההסתברות של כל אחת מהן (emission probabilities) מופרדים בטאב. פורמט הקובץ מודגם בנספח. מילים לא ידועות יתויגו כNNP בהסתברות 1. זכרו להמיר את ערכי ההסתברויות לביטויים לוגריתמים.
  - קובץ הפרמטרים התחבירי מכיל את הסתברויות המעבר ממצב לממצב (transition probabilities) בפורמט הנקרא DARPA/SRILM ומוגדר בנספח. על המשערכים להסתמך על ספירות בקובץ האימון בלבד. זכרו להמיר את ערכי ההסתברויות לביטויים לוגריתמים.

- ממשו תכנית decode המקבלת כקלט קובץ משפטים לא מתוייגים ואת שני קבצי הפרמטרים שהגדרנו ומחזירה קובץ משפטים מתויג כפלט. התוכנית אמורה למצוא בזמן פולינומיאלי את רצף המצבים)או התגים(הממקסם את הסתברות פונקציית המטרה. הקובץ המתויג הינו בפורמט הזהה לקובץ הגולד שהוגדר בהקדמה.

ענו על השאלות הבאות

1. מהי פונקציית המטרה של המודל?
2. מהן נוסחאות הפרמטרים של המודל?
3. מהן נוסחאות המשערכים של הפרמטרים במודל?
4. ?מהי סיבוכיות זמן הריצה של אלגוריתם האימון
5. מהי סיבוכיות זמן הריצה של אלגוריתם התיוג?
6. מהו הדיוק macro-avg עבור קובץ הבדיקה
7. במקום לתייגן כ NNP, ממשו טכניקת החלקה למילים לא ידועות.
   האם תוצאותיכם השתפרו?
8. ממשו בנוסף טכניקת החלקה למעברים לא ידועים.
   האם תוצאותיכם השתפרו ביחס לסעיף הקודם??

4. ניתוח התוצאות

1. עבור המודל הטוב ביותר שלך חשב את מטריצה הבילבול (confusion matrix). מהן שלוש השגיאות הנפוצות ביותר לפי המטריצה? תזכורת: מטריצת הבלבול היא מטריצה המכילה עבור כל כניסה m_{ij} כמה פעמים מילה האמורה לקבל תג t_j קיבלה בפועל את התיוג t_i.

   ב t_{ii} אם כן כותבים את מספר המקרים בהם המתייג צדק עבור תג t_i.

2. תייגו את המשפט הבא ידנית ולאחר מכן תייגו אותו באמצעות המתייג שכתבתם
אישה נעלה נעלה נעלה, נעלה את הדלת לפני בעלה.

השוו את שני הרצפים, שלכם ושל המתייג האוטומטי. היכן צדק המתייג? היכן הוא טעה? מהם אחוזי הדיוק ביחס לתיוג הידני? איזו אינפורמציה חסרה לדעתכם במודל על מנת לדייק יותר?

3. עבור המודל הטוב ביותר שלכם חלקו את קובץ ה ל10 חלקים שווים, ממוספרים מ1 עד 10. הריצו את הניסויים הבאים:

- train on 1, decode pos.test
- train on 1+2, decode pos.test
- ...
- train on 1+2+3+ ... + 10, decode pos.test

ציירו את עקומת הלימוד של המודל. על ציר הx האיטרציות 1 עד 10, ולכל איטרציה סמנו את התוצאה בגרף. מה יקרה להערכתכם אם יתווספו עוד טקסטים מתוייגים לאימון ?

5. מתייג מסדר אקראי - שאלת בונוס

הרחיבו את מימוש המאמן והמתייג שלכם עבור שרשראות מעברים מרקוביות מסדר N.
)רמז: הגדירו מחדש מהו מצב בשרשרת מרקובית(

1. האם הגדרות הפרמטרים השתנו? אם כן כיצד
2. האם אלגוריתם התיוג השתנה? אם כן כיצד
3. האם סיבוכיות האימון השתנתה? אם כן כיצד
4. האם סיבוכיות התיוג השתנתה? אם כן כיצד
5. מהן התוצאות עבור N=2,3,4 ומה לדעתכם תהיה מגמת התוצאה עבור N=5 ?

6. סיכום ומסקנות

תנו תאור סכמטי כולל של הקוד שמימשתם וכן תארו ממעוף הציפור את התוצאות והמסקנות מהתרגיל. בפרט, נסו לענות על השאלות המחקריות לגבי המתייג בעברית שהוצגו בהקדמה לתרגיל והציעו בכתב דרכים אפשריות לטיוב ושיפור עתידי של המתייג שלכם.

מה להגיש

הגישו ספריה מכווצת המכילה את הספריות והקבצים הבאים:

- Your **train** execution script/wrapper
- Your **decode** execution script/wrapper
- Your **evaluate** execution script/wrapper
- **src/** Your documented code.
- **exps/** All input, output and parameter files, from all experiments.
- **results/** All your result files.
- **doc/** A typed report (addressing the above questions) in *.pdf format.

בהצלחה!

```
#————————————
# Train and Gold File Format
#————————————

# sentence 1
SEG_tab_POS
SEG_tab_POS
...
...
SEG_tab_POS

# sentence 2
SEG_tab_POS
...
...

#————————————
```

איור 1: פורמט קבצי האימון

```
#————————————
# Test File Format
#————————————

# sentence 1
SEG
SEG
...
...
SEG

# sentence 2
SEG
...
...

#————————————
```

איור 2: פורמט קבצי הבדיקה

#————————————————————————————————
Part-of-Speech Tagging Evaluation
#————————————————————————————————
#
Model: < *specify* >
Smoothing: < $y|n$ >
Test File: : < *specify* >
Gold File: : < *specify* >
#
#————————————————————————————————
sent-num word-accuracy sent-accuracy
#————————————————————————————————
1 < *seg-accuracy* > < *sent-accuracy* >
2 < *seg-accuracy* > < *sent-accuracy* >
....
N < *seg-accuracy* > < *sent-accuracy* >
#————————————————————————————————
macro-avg < *seg-accuracy-all* > < *sent-accuracy-all* >

איור 3: פורמט קבצי ההערכה

```
SEG  POS1 logprob1 POS2 logprob2 ...
SEG  POS1 logprob1 POS2 logprob2 ...
....
SEG  POS1 logprob1 POS2 logprob2 ...
```

איור 4: פורמט קובץ הפרמטרים הלקסיקלי

```
\data\
ngram 1 = <number of unigrams>
ngram 2 = <number of bigrams>
...
<line-break>
\1-grams\
logprob  POS
logprob  POS
...
<line-break>
\2-grams\
logprob  POS POS
logprob  POS POS
...
```

איור 5: SRILM/DARPA **פורמט קובץ הפרמטרים התחבירי**