hebrewNER

1. רקע

HebrewNER היא חבילת תכנה לזיהוי שמות פרטיים בטקסט עברי. משימת זיהוי שמות פרטיים בטקסט היא אחת ממשימות חילוץ מידע (information extraction) בתחום עיבוד השפה הטבעית. במשימה זו אנו רוצים לתת לכל מילה בטקסט תיוג "שם פרטי" או "לא שם פרטי". במקרה של מתן תיוג "שם פרטי" יש לציין את סוג שם.

זיהוי השמות הפרטיים במערכת זו מתבצע ע"י שילוב מספר מודלים: מודל מרקוב חבוי (HMM), מודל אנטרופיה מקסימלית ומודל המשלב ביטויים רגולאריים עם לקסיקון הנבנה בעזרת קורפוס אימון. מודל האנטרופיה המקסימלית הוא המודל המוצלח ביותר ומצליח להתמודד עם כמות גדולה של מאפיינים. אולם המערכת אשר הציגה את התוצאות הטובות ביותר היא המערכת המשולבת.

החבילה כתובה בשפת java ומשתמשת בכלים חיצוניים אשר כתובים ב-java גם הם. לקבלת הניתוח המורפולוגי של המילים נעשה שימוש במנתח חלקי דיבר הכולל מפיג עמימות מורפולוגית של מנחם אדלר מאוניברסיטת בן גוריון. ליצירת מודל אנטרופיה מקסימלית נעשה שימוש חבילת opennlp. חבילה זו מבוססת על עקרון הקוד הפתוח וכתובה ב-java. היא משמשת לצרכי אימון ויצירת מודל הסתברותי ע"פ עקרונות האנטרופיה המקסימלית בהינתן קבוצת מאפיינים, קבוצת תגים וקורפוס אימון מתויג מחולק לטוקנים.

התיוג מתבצע על טקסטים בשפה העברית בפורמט UTF8. כעיבוד מקדים למשימת התיוג מתבצעים תהליכים של זיהוי משפטים וטוקניזציה שלהם (טוקן- קבוצת תווים המתוחמת ע"י רווחים או סימנים מיוחדים כגון נקודה, סימן שאלה ועוד).

2. סוגי התגים

רשימת התגים למשימת זיהוי ותיוג שמות פרטיים:

PERS - ביטוי שם אדם

LOC - ביטוי שם מקום

ORG - ביטוי שם ארגון

TIME - ביטוי זמן

DATE - ביטוי תאריך

MONEY - ביטוי כסף

PERCENT - ביטוי אחוזים

O – "אחר", לא שם פרטי

3. שיטת התיוג

שיטת התיוג היא כפי שהוגדרה במשימת CoNLL2003. לפי שיטה זו מילה בתוך ביטוי X תקבל תג I_X . עבור שני ביטויים מאותו הסוג (X) המופיעים ברצף, המילה הראשונה בביטוי השני תקבל את התג B_X. תיוג ייתן לטוקן כולו ואין המילה מפורקת לחלקיה השונים. כלומר, תחיליות וסופיות יכללו בתיוג.

4. מה בחבילה?

החבילה מכילה את הספריות הבאות:

- **bin** – מכילה שני קבצי פקודות להרצה וקימפול המערכת בסביבת Windows (ראה 5,6).
- **src** - קבצי המקור בשפת java. מכילה ספריות נפרדות עבור כל אחד מהמודלים, ספריה עבור המערכת לזיהוי משפטים (המשתמשת גם היא במודל אנטרופיה מקסימלית) וספריות עזר המשמשות את כל המודלים.
- **lists** – רשימות מילים המשמשות את המודלים: מילון ורשימות מילים שונות.
- **models** – ספריה המכילה את המודלים השונים כפי שנתקבלו לאחר אימון על קורפוס האימון.
- **jars** – מכילה קבצי jar של המערכות החיצוניות.
- **data** – קבצי מידע המשמשים את המנתח המורפולוגי.
- **corpus** - קובץ המכיל את קורפוס האימון ששימש את שלושת המודלים.
- **input_samples** - קבצי טקסט עברי בפורמט UTF8 .
- **javadocs** – קבצי תיעוד של קבצי המקור.

5. התקנה

החבילה אינה דורשת התקנה מיוחדת. דורשת java בגרסת 1.7 ומעלה.

כדי להריץ את קבצי הפקודות יש לשנות את משתנה JAVA_HOME בקבצים

```
bin\build.bat
bin\hebNER
```

להיות הספריה בה מותקנת java על המחשב.

לאחר פעולת הקימפול תיווצר בספריה HebrewNER ספריה חדשה hebrewNER המכילה את הקבצים המקומפלים.

6. הרצה

ניתן להריץ טסטים בעזרת JUnit. טסטים אלה בודקים ששלושת קבצי הדוגמה מספקים את הפלט הדרוש.

לאחר הקימפול יש להריץ את המערכת ע״י הקובץ bin\hebNER מהספריה HebrewNER:

Usage: bin\hebNER <source file> <output file>

כאשר source file הוא קובץ טקסט בעברית בפורמט UTF8. למשל:

```
> bin\hebNER input_samples\500.txt output_samples\500.ner
```

פורמט הטקסט המתויג: כל שורה בקובץ מייצגת טוקן. כאשר העמודה הראשונה בשורה היא הטוקן עצמו והשנייה היא התג שאר ניתן ע״י המערכת. בין העמודות מפריד טאב, ומשפטים מופרדים ע״י שורה ריקה.

בנוסף לשיטות התיוג, לכל מודל יש מחלקות המשמשות לאימון על בסיס קורפוס אימון. מחלקות אלה נמצאות בספריות של כל מודל.

7. יצירת קשר

המערכת נבנתה במסגרת עבודת המחקר לתואר שני ע״י נעמה בן מרדכי מאוניברסיטת בן גוריון benmorde@cs.bgu.ac.il, בהנחיית ד״ר מיכאל אלחדד elhadad@cs.bgu.ac.il.

מתייג חלקי דיבר ומפיג עמימות מורפולוגית נכתב ע״י מנחם אדלר מאוניברסיטת בן גוריון: adlerm@cs.bgu.ac.il.

עדכון (07/08/13):

הפרויקט עודכן לג'אווה 7, גרסה מעודכנת של opennlp (בעזרת maven).
על ידי אלעזר גרשוני elazarg@gmail.com

8. קישורים

1. פרויקט opennlp: http://opennlp.sourceforge.net/
2. פרויקט MaxEnt: http://maxent.sourceforge.net/
3. משימת CoNLL2003: http://www.cnts.ua.ac.be/conll2003/ner/