תוכנית עבודה

זיהוי שמות בטקסט בעברית - NER (Name Entity Recognition)

– קבוצה 120

צוות הפרוייקט:

1. שחר אוסובסקי, 300579786, shahar.osovsky@mail.huji.ac.il
2. דוד גיל גבירץ, 200117679, gilgvirts@gmail.com

מנחה:

אפי לוי, efle@cs.huji.ac.il

תקציר:

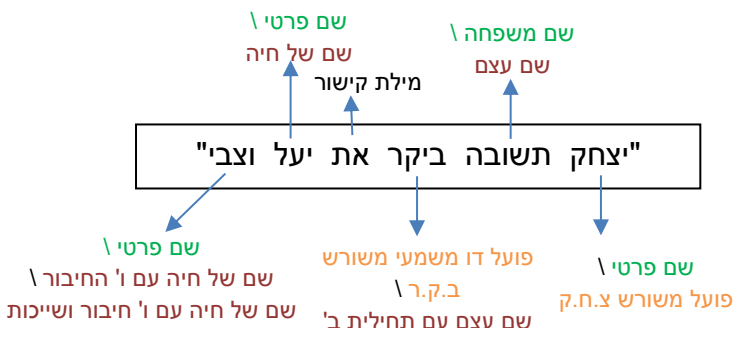
ככל שעוברות השנים, פחות ופחות טקסטים כתובים על נייר וכמות הטקסטים הדיגיטליים עולה משמעותית. כאשר מידע מאוחסן בצורה דיגיטלית ניתן לבצע עליו עיבודים וניתוחים שונים. כיום תחומי ניתוח המידע הדיגיטלי מתפתחים, ובתוכם תחום עיבוד שפה טבעית.

עיבוד שפה טבעית פירושו לפתח אלגוריתם, במגוון דרכים, שיספק ללקוח (משתמש או פלטפורמה אחרת), מידע על הטקסט. מידע זה יכול להיות לדוגמא חלקי הדיבר בטקסט (נושא, נשוא וכו'), תיוג של הטקסט. עיבוד מתקדם הוא למצוא סמנטיקה בטקסט, איסוף מידע רלוונטי ועוד.

אנחנו בחרנו להתמקד ב**תיוג טקסט**. בתיוג טקסט אופייני יש את התגים הבאים (עבור כל מילה בנפרד): *שם פרטי*, *מקום*, *ארגון*, *תאריך*, *זמן*, *אחוזים* וביטוי *כסף*. תיוג זה נקרא זיהוי ישויות או באנגלית (**NER**) Named Entity Recognition. ניתן למצוא בכלי זה יישומים רבים, אנחנו הגענו לנושא כיוון שרצינו לזהות קשרים בין ישויות בטקסט, והצעד הראשון הוא זיהוי הישויות. ניתן לחשוב על יישומים נוספים, כגון כריית מידע לצרכים מסחריים או ביטחוניים, ועוד מספר אפשרויות פחות נוראיות.

המטרה הראשית של הפרויקט היא פיתוח מערכת לומדת מתאימה לזיהוי שמות פרטיים בעלת ביצועים גבוהים מהמערכות הקיימות. הצעד הראשון הוא להכיר ולהבין את העבודה שכבר נעשתה בנושא. לאחר מכן ניצור מערכת בסיסית משלנו, ונשפר את המאפיינים אותה היא מקבלת בתקווה להגיע לביצועים מוצלחים יותר. במידה ונצליח לממש מערכת NER יעילה, נרצה להראות יישום אפליקטיבי שלה. כשלב אפשרי, נבנה תוסף לדפדפן או יישומון המזהה ישויות בכתבות ומציג את הקשרים ביניהן.

זיהוי ישויות בעברית היא משימה משמעותית יותר מסובכת מאשר באנגלית, בין היתר בגלל המאפיין של אות גדולה בתחילת שם (Capital letters) אשר לא קיים בעברית. לכן, כלי NER בעל יכולות גבוהות כבר קיים באנגלית, ובשפות נוספות. בעברית קיימים קשיים נוספים כגון כפל משמעות:



דוגמא נוספת למילה בעלת 4 משמעויות שונות באותו המשפט:
"אִשָּׁה נַעֲלָה ,נַעֲלָה נַעֲלָה ,נַעֲלָה אֶת הַדֶּלֶת בִּפְנֵי בַּעֲלָהּ."

המדד הכמותי לבדיקת הצלחת מערכת תיוג נקרא F-measure מדד זה הוא שילוב של מדד הזיהוי החסר ($Missing$ - התוכנה תייגה נכון חלק מהביטוי), למדד הזיהוי השגוי ($Incorrect$ - התוכנה תייגה ביטוי שלם בתג שגוי), כאשר $Spurious$ מתייחס לתיוג של ביטוי שלא אמור להיות מתויג.

$$, F_{measure} = \frac{2*Percision*Recall}{Percision+Recall}$$

כאשר $Percision = \frac{Correct}{Correct+Incorrect+Spurious}$ ו- $Recall = \frac{Correct}{Correct+Incorrect+Missing}$

ה- F-measure של המערכת הקודמת בעברית שנכתבה (ופורסמה בתזה אקדמית) הוא 79.1. ציון זה הושג ע"י נעמה בן מרדכי בהדרכת פרופסור מיכאל אלחדד באוניברסיטת בן גוריון, בשנת 2006. עבודתה משלבת שני מערכות לומדות, שרשראות מרקוב ואנטרופיה מקסימלית, בנוסף למילונים וכללי שפה ותחביר.

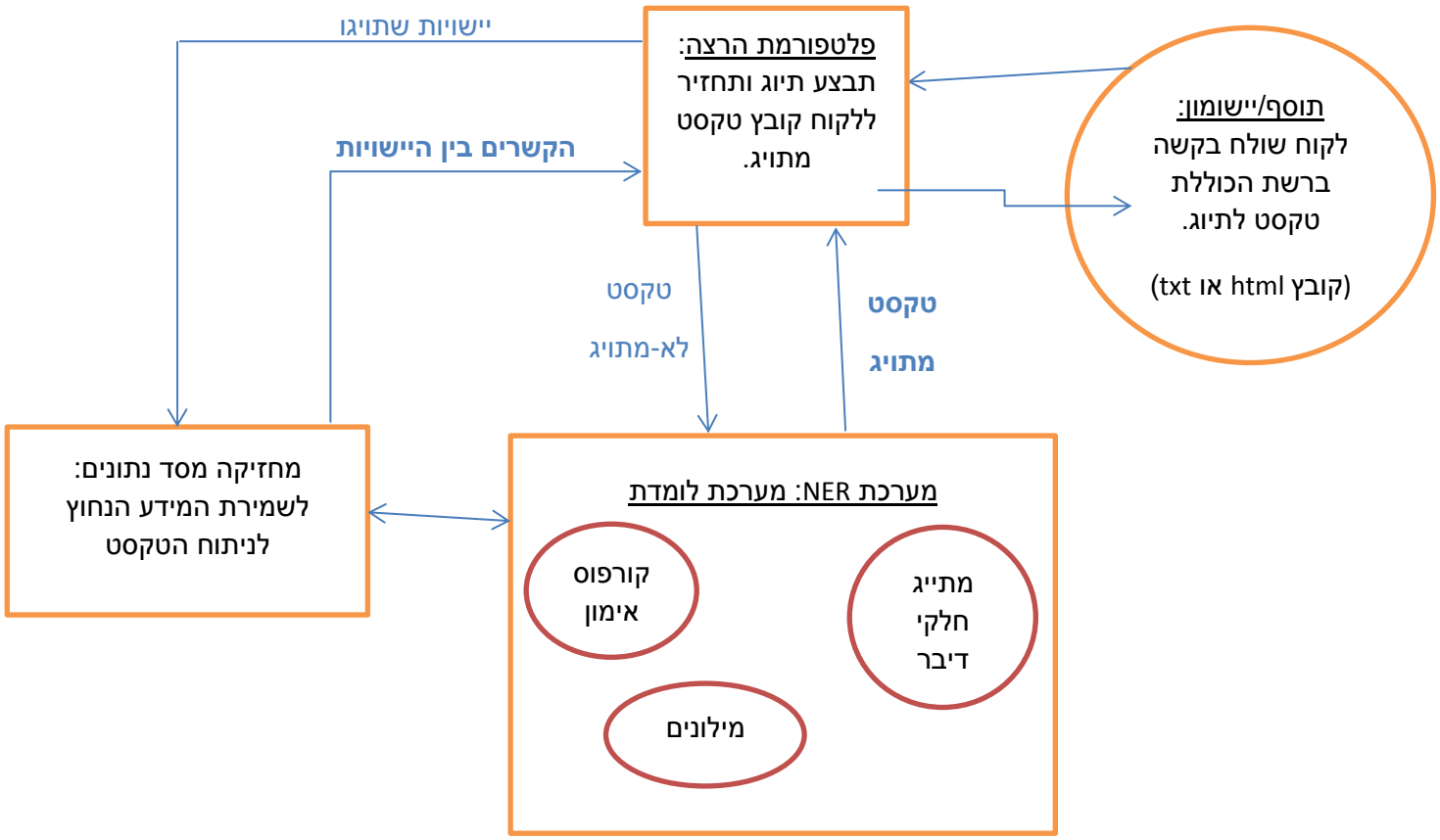
בשנים האחרונות חלו שיפורים משמעותיים ביכולות חישוב, אלגוריתמים מתקדמים יותר (– CRF Conditional Random Field, Structure Prediction for Linear Chain, עליהם לא נרחיב כאן) וכמו כן ישנם מסדי מידע גדולים מאוד ברשת, ולכן אנו מאמינים כי יש מקום להשגת תוצאות טובות יותר.

באנגלית כאמור, הושגו תוצאות משמעותית יותר טובות, כאשר הגיעו לF-measure הגדול מ- 95%, כאשר המינימום שהוגש היה 86.72% בכנס CoNLL ו-92.28% בכנס CMU.*

הגישה הנאיבית היא כי ניתן לפתח תוכנה לזיהוי שמות על ידי פיתוח כללים על פיהם נחליט האם מילה היא שם או לא, בהתאם לכללי שפה והתחביר. שיטה זאת לבדה אינה מגיעה לאחוזי הצלחה גבוהים, כיוון שאינה פשוטה לעדכון עם התפתחות השפה, ומתקשה בטקסט לא תקני. דרך נוספת, אלגנטית ומעשית יותר, היא שילוב כללים יחד עם תוכנה לומדת. תוכנה לומדת מקבלת טקסט מתויג בצורה נכונה, ומפיקה ממנו על ידי כימות מאפיינים מסוימים, כגון מהי המילה ואיזה תג קיבלו מילים בסביבתה, מידע כיצד לתייג טקסט חדש.

בבניית מערכת NER ישנם מספר מרכיבים הדורשים זמן ממושך. בחירת מאפיינים ומדידת השפעתם על ביצועי המערכת. בניית מילונים, גם אם בעזרת שיטות אוטומטיות. בניית קורפוס (אסופת טקסטים מתויגים) דורשת זמן רב, ולרוב נעשית על ידי מתייגים רבים.

*מתוך: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling
Jenny Rose Finkel, Trond Grenager, and Christopher Manning Computer Science Department
Stanford University



מצבנו הנוכחי: יש לנו דרך להריץ בדיקות NER על סטים של אימון מתוך הקורפוס שברשותנו. כתבנו טסט הערכת ביצועים בסיסי שנותן לנו F-Measure להערכת איכות התיוג שביצענו.
כרגע מכשולים טכניים שעומדים לפנינו הם השגת גישה לקלאסטר של האוניברסיטה כדי להריץ את ריצת ה"למידה" של הכלי. כמובן שניהול זמן הוא מכשול בסיסי שקיים גם הוא.

התוכנית להמשך:

תאריך	שחר	דוד גיל
12/17	כתיבת טסט הערכת ביצועים ספציפי לפי סוגיי תגים (**P0**).	הרצת CRF עם מגוון פיצ'רים על גבי הקלאסטר (**P0**).
1/18	כתיבת שיטת ניהול מסמכים\קבצים עבור המערכת הלומדת (**P0**).	החזרת פידבק למערכת של התוצאות כדי להפוך אותה למערכת לומדת (**P0**).
2/18	הגעה ליעד- מערכת NER משופרת	
3/18	המשך מחודש קודם תיקון באגים אחרונים (**P0**).	המשך מחודש קודם. בדיקות אחרונות (**P0**).
4/18	יישום מעשי של מערכת ה- NER: בניית תוסף כרום/יישומון רשת (**P1**).	יישום מעשי של מערכת ה- NER: בניית מאגר נתונים (**P1**).
5/18	כתיבת מצגת	כתיבת פוסטר
6/18	כתיבת ספר	