אוניברסיטת בן - גוריון בנגב

הפקולטה למדעי הטבע

המחלקה למדעי המחשב

זיהוי שמות פרטיים בעברית

חיבור לשם קבלת תואר "מגיסטר" בפקולטה למדעי הטבע

מאת נעמה בן מרדכי

אלול תשס"ה ספטמבר 2005

זיהוי שמות פרטיים בעברית

חיבור לשם קבלת תואר "מגיסטר" בפקולטה למדעי הטבע

מאת נעמה בן מרדכי

שם המנחה ד"ר מיכאל אלחדד

המחלקה מדעי המחשב

הפקולטה למדעי הטבע

אוניברסיטת בן – גוריון בנגב

חתימת המחבר________________________________ תאריך________________

אישור המנחה_______________________________ תאריך________________

אישור יו"ר ועדה מחלקתית______________________________ תאריך________________

תקציר

משימת זיהוי שמות פרטיים בטקסט היא אחת ממשימות הוצאת מידע 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 השם. במחקר הנוכחי נתרכז בתיוג שמות אדם, מקום, ארגון, תאריך, זמן, כסף ואחוזים.

בשנים האחרונות ישנה התעסקות במשימה זו בשל היותה שלב בסיסי בניתוח אוטומטי של טקסט. עבודה רבה נעשתה עבור טקסט בשפה האנגלית ואף נערכו תחרויות בהן חוקרים שונים אשר הציגו את עבודתם והשוו ביניהן. בשפה העברית החלה התעסקות במשימה זו, אולם עד כה לא הוצגו תוצאות מספקות.

בשפה האנגלית יש שימוש באותיות גדולות בתחילת כל שם פרטי. תכונה זו של השפה מפשטת במידה רבה את משימת התיוג. השפה העברית מורכבת הרבה יותר. בעברית יש שימוש באותיות אחידות לאורך כל המשפט, אולם בעיה זו אינה הבעיה היחידה בהתמודדות עם ניתוח אוטומטי של השפה. העברית עשירה מורפולוגית ובעלת רב משמעיות גדולה. שורש של מילה יכול לקבל צורות שונות בהתאם למאפיינים כגון: גוף, מין, זמן, כמות, סמיכות, שייכות ועוד. מאחר והעברית כתובה כיום ללא ניקוד, מילה מסוימת הכתובה הטקסט יכולה להתפרש במספר אופנים. תכונה נוספת של העברית היא סדר חופשי של מילים לעומת אנגלית בה סדר המילים מובנה יותר.

בשפה האנגלית קיימות מערכות שונות המשתמשות בשיטות שונות. לשיטות אוטומטיות (למשל מודל מרקוב חבוי, מודל אנטרופיה מקסימלית) יש יתרון גדול בשל הדינאמיות שלהן. שפה טבעית היא בד"כ דבר משתנה, השפה מתפתחת ונוספים שמות חדשים. מערכת אוטומטית גמישה יותר לשינויים כאלה. יתרון נוסף של המערכות האוטומטיות הוא שפיתוחן אינו דורש ידע לשוני מקצועי.

מודל האנטרופיה המקסימלית הוא המודל אשר הציג את התוצאות הטובות ביותר עבור השפה האנגלית מבין כל השיטות האוטומטיות. שיטה זו מתמודדת היטב עם מספר גבוה של מאפיינים וגם עם כמות דלילה של מידע. עקרון האנטרופיה המקסימלית משמש לבניית מודל סטטיסטי הסתברותי. ע"פ עקרון זה על המודל לשקף את כל המידע הנתון, אך להימנע מלהניח דבר על מה שלא ידוע. המודל הסטטיסטי נבנה על בסיס קורפוס אימון מתויג. לאחר בניית המודל ההסתברותי, מתבצע חיפוש למציאת רצף התגים הסביר ביותר בטקסט חדש.

השלב הראשון במחקר זה היה שלב הגדרת המשימה. מכיוון שלא נעשתה עבודה רבה בתחום זה בעברית, היה צורך לנסח הגדרת משימה חד משמעית. נוסחו הנחיות לתיוג ונבדקה רמת ההסכמה בקרב מתייגים אנושיים על הנחיות אלו.

במהלך המחקר נבנו שלושה מודלים לתיוג, שניים מהם אוטומטיים. המודלים הם: מודל מרקוב חבוי, מודל אנטרופיה מקסימלית ומודל קו ההתחלה המבוסס על ביטויים רגולריים ולקסיקון הנבנה מקורפוס האימון.

המחקר התמקד במציאת מאפיינים עבור משימת תיוג השמות הפרטיים מיוחדים לשפה העברית. המאפיינים היו דומים עבור שלושת המודלים. מצאנו כי המאפיינים החשובים ביותר הם המאפיינים המקומיים. המאפיינים הקובעים את תפקיד המילה במשפט הם המאפיינים של המילה עצמה ושל 1-2 מילים סביבה. ראינו כי המאפיינים הדומיננטיים ביותר הם מאפייני המילון ומאפיינים תחביריים - מורפולוגיים. עם זאת, התוצאות הטובות ביותר מתקבלות משילוב קבוצה גדולה ומגוונת של מאפיינים. בנוסף למאפיינים,ראינו כי לקורפוס האימון השפעה רבה על תוצאות התיוג.

מודל האנטרופיה המקסימלית היה המודל המוצלח ביותר, אשר הציג תוצאות גבוהות בהרבה מהשניים האחרים. עם זאת, המערכת אשר מציגה את התוצאות הטובות ביותר היא מערכת אשר משלבת את שלושת המודלים.

המערכת המשולבת הציגה תוצאות טובות ואחוזי דיוק גבוהים.

תוכן העניינים

1 רקע ומבוא 10

1.1 משימת זיהוי השמות הפרטיים 10

1.2 שימושים של מערכת לזיהוי שמות פרטיים 11

1.3 ייחודה של השפה העברית 12

1.4 הערכת מערכות לזיהוי שמות פרטיים 14

2 עבודות קודמות 16

2.1 הגישה הידנית 16

2.2 גישות אוטומטיות: מודלי מרקוב 18

2.2.1 מודל מרקוב סמוי (Hidden Markov Model) 18

2.2.2 מערכת לזיהוי שמות פרטיים המבוססת על HMM 19

2.2.3 HMM מבוסס תוים 21

2.3 גישות אוטומטיות: שיטת אנטרופיה מקסימלית 23

2.3.1 אנטרופיה מקסימלית 23

2.3.2 אילוצים על המודל 24

2.3.3 אלגוריתם GIS לבניית מודל אנטרופיה מקסימלית 25

2.3.4 שימוש במודל 26

2.3.5 מערכות זיהוי שמות פרטיים המבוססות על מודל אנטרופיה מקסימלית 26

2.4 גישות אוטומטיות: ROBUST RISK MINIMIZATION 30

2.5 שילוב מערכות 33

3 הסכמה 35

3.1 הגדרת משימת תיוג שמות פרטיים בטקסט 35

3.2 מבחני הסכמה 35

3.2.1 מבחן גבולות (TEXT) 35

3.2.2 מבחן "הסוג" (TYPE) 36

3.2.3 המבחן המשולב (TEXT&TYPE) 36

3.3 חישוב הסכמה 37

3.3.1 הסכמה חלקית 37

3.3.2 סטטיסטיקת קאפה 37

3.4 מהלך הניסוי 39
3.5 תוצאות ומסקנות 40
3.5.1 מבחן t 40
3.5.2 תוצאות שני השלבים הראשונים של הניסוי 40
3.5.3 תוצאות השלב השלישי של הניסוי 43

4 הקורפוס 45

5 קו ההתחלה 47

6 זיהוי שמות פרטיים באמצעות מודל אנטרופיה מקסימלית 49

6.1 חבילת MaxEnt 49
6.2 אלגוריתם החיפוש - Beam Search 52
6.3 המאפיינים 53
6.3.1 מאפייני מבנה 53
6.3.2 מאפייני לקסיקון 53
6.3.3 תיוגים קודמים 53
6.3.4 מילון 53
6.3.5 ביטויים רגולריים מקומיים 55
6.3.6 ביטויים רגולריים ארוכים 56
6.3.7 שימוש במנתח חלקי דיבר ומפיג עמימות 57
6.3.8 עיבוד מוקדם 58
6.3.9 מכפלות מאפיינים 58
6.4 ניתוח תוצאות 59
6.4.1 חלקי דיבר 61
6.4.2 מילון 62
6.4.3 למה 63
6.4.4 עיבוד מוקדם 65
6.4.5 השפעת וסוג גודל הקורפוס 65
6.4.6 סיכום התוצאות 68

7 זיהוי שמות פרטיים באמצעות מודל מרקוב חבוי (HMM) 69

7.1 המודל הבסיסי 69
7.2 חלקי דיבר כמצבים 70
7.3 HMM המשלב מאפיינים 71
7.4 השוואה למודל האנטרופיה המקסימלית 74

8 שילוב מערכות 76

9 השוואה למערכות קיימות 78

9.1 השוואה למערכת אנטרופיה מקסימלית של גנדי למברסקי 78

9.2 השוואה למערכות עבור השפה האנגלית.. 80

9.3 השוואה למערכות עבור שפות נוספות 81

10 סיכום ומסקנות 83

11 הצעות למחקר עתידי 84

12 נספח 1: קוים מנחים לתיוג שמות פרטיים בעברית 81

13 נספח 2: דוגמא לטקסט מתויג.. 96

14 סימוכין 100

רשימת טבלאות ואיורים

איור 2-1: מבנה ה- "IDENTIFINDER" 20

איור 2-2: HMM מבוסס תוים 21

טבלה 2-1: מאפייני מערכת ה-RRM 31

טבלה 3-1: תוצאות מבחן TEST של שלבים 1+2 בניסוי ההסכמה 40

טבלה 3-2: תוצאות מבחן TYPE של שלבים 1+2 בניסוי ההסכמה 41

טבלה 3-3: תוצאות מבחן TEXT&TYPE של שלבים 1+2 בניסוי ההסכמה 41

טבלה 3-4: בלבול של שלבים 1+2 בניסוי ההסכמה 42

טבלה 3-5: תוצאות מבחן TEXT של שלב 3 בניסוי ההסכמה 43

טבלה 3-6: תוצאות מבחן TYPE של שלב 3 בניסוי ההסכמה 44

טבלה 3-7: תוצאות מבחן TEXT&TYPE של שלב 3 בניסוי ההסכמה 44

טבלה 4-1: חלוקת הקורפוס למקורות 45

טבלה 4-2: חלוקת הקורפוס לביטויים 46

טבלה 4-3: אורך ביטוי ממוצע בקורפוס 46

טבלה 5-1: קו ההתחלה 48

איור 6-1: ארכיטקטורת MAXENT 51

טבלה 6-1: כרכי המילון 55

טבלה 6-2: סוגי ביטויים רגולריים 56

טבלה 6-3: תוצאות מודל האנטרופיה המקסימלית 59

טבלה 6-4: תוצאות מערכת אנטרופיה מקסימלית מבוססת חלקי דיבר 61

טבלה 6-5: תוצאות מערכת אנטרופיה מקסימלית ללא חלקי דיבר 62

טבלה 6-6: תוצאות מערכת אנטרופיה מקסימלית מבוססת מילון 62

טבלה 6-7: תוצאות מערכת אנטרופיה מקסימלית ללא מילון 63

טבלה 6-8: תוצאות מערכת אנטרופיה מקסימלית מבוססת למה 64

טבלה 6-9: תוצאות מערכת אנטרופיה מקסימלית ללא למה 65

טבלה 6-10: תוצאות מערכת אנטרופיה מקסימלית ללא רשימות עיבוד מוקדם................................ 65

טבלה 6-11: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "מעריב"..................................... 66

טבלה 6-12: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "הארץ"..................................... 66

טבלה 6-13: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "ערוץ 7" 67

טבלה 6-14: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס כלכלה...................................... 67

טבלה 7-1: תוצאות מערכת HMM הבסיסי 70

טבלה 7-2: תוצאות מערכת HMM – חלקי דיבר כמצבים 71

טבלה 7-3: תוצאות מערכת HMM – שילוב מאפיינים.. 72

טבלה 8-1: תוצאות המערכת המשולבת.. 77

טבלה 9-1: השוואת מאפיינים עם גנדי למברסקי 79

טבלה 9-2: תוצאות המערכת של גנדי למברסקי 80

1 רקע ומבוא

1.1 משימת זיהוי השמות הפרטיים

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

בשנים האחרונות ישנה התעסקות במשימה זו בשל היותה שלב בסיסי בניתוח אוטומטי של טקסט. עבודה רבה נעשתה עבור טקסט בשפה האנגלית ואף נערכו תחרויות [23] בהן חוקרים שונים אשר הציגו את עבודתם והשוו ביניהן. בשפה העברית החלה התעסקות במשימה זו [8] אולם עד כה לא הוצגו תוצאות מספקות.

ההגדרה למשימה עבור השפה האנגלית ניתנה ב-1999 ע"י Chinchor, Brown, Ferro, Robinson [18] . לפי הגדרה זו, המשימה מתחלקת לשלוש תת משימות: שמות ישויות (אדם, ארגון, מקום) ביטויי זמן (תאריך, שעה, משך זמן) וביטויים מספריים (כסף, אחוז, מידה, כמות). במשימת CoNLL2003 ההגדרה הייתה שונה. התרכזו בשמות אדם, מקום וארגון (ישויות) והגדירו את כל השאר כ"שונות". בקטגורית ה"שונות" נכללו גם שמות דגמים, ספרים, שירים אירועים ועוד. עד כה לא הייתה קיימת הגדרת משימה עבור השפה העברית. חלק ממטרת מחקר זה היה לנסח הגדרות חד משמעיות לתיוג.

במחקר הנוכחי נתרכז בתיוג שמות אדם, מקום, ארגון, תאריך, זמן, כסף ואחוזים.

דוגמאות:

- "ממשלת קלינטון"

  ממשלת <PERSON> קלינטון </PERSON >

- "יו"ר הליכוד, אריאל שרון, ..."

  יו"ר <ORGANIZATION> הליכוד </ORGANIZATION >, <PERSON > אריאל שרון </PERSON >, ...

- "בישראל, ביום שלישי בשעה שמונה..."

  <LOCATION> בישראל </LOCATION >, <DATE > ביום שלישי </DATE > <TIME > בשעה שמונה </TIME> ...

- "עשרים מיליון שקלים חדשים"

  <MONEY> עשרים מיליון שקלים חדשים </MONEY>

נשתמש בשיטת תיוג 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 (טוקן- קבוצת תווים המתוחמת ע"י רווחים) כולו ואין המילה מפורקת לחלקיה השונים. כלומר, תחיליות וסופיות יכללו בתיוג.

מתקבל אוסף של 15 תגים אפשריים.

לדוגמא עבור המשפט:

"קולין פארל ואנג'לינה ג'ולי משחקים בסרט החדש"

רצף התגים הנכון הוא

O O O O I_PERSON B_PERSON I_PERSON I_PERSON

בתיוג הטוקן "ואנג'לינה" ו' החיבור נכללת בתיוג ואינה מקבלת תיוג נפרד.

במחקר זה ננסה ליישם שיטות לזיהוי שמות פרטיים שהוכיחו את עצמן בשפות אחרות. קיימות מערכות שונות המשתמשות בשיטות שונות. מערכות המבוססות על חוקים ידניים הוכיחו את עצמן להיות הטובות ביותר. בניית מערכת ידנית דורשות ידע לשוני מקצועי וזמן פיתוח רב. כמו כן מערכות אלה רגישות מאוד לשינויים בשפה ובאופי הטקסט. לשיטות אוטומטיות יש יתרון גדול בשל הדינאמיות שלהן. שפה טבעית היא בד"כ דבר משתנה, השפה מתפתחת (סלנג) ונוספים שמות חדשים. מערכת אוטומטית גמישה יותר לשינויים כאלה. למידה אוטומטית מתבססת על למידה מקבוצה של טקסטים. במקרים של שינוי הגדרות, אופי הטקסטים או אפילו השפה לעיתים השינוי הנדרש ממערכת אוטומטית הוא פשוט אימון מחדש של המערכת.

במחקר זה ננסה לבנות מערכת לזיהוי שמות פרטיים אוטומטית עבור השפה העברית. ננסה לבחון אילו מאפיינים משפיעים ביותר על תיוג שמות בשפה העברית וכיצד משפיע גודל וסוג הקורפוס על אופי הלמידה של המערכת.

1.2 שימושים של מערכת לזיהוי שמות פרטיים

כאמור, מערכת לזיהוי שמות פרטיים הינה כלי בסיס במשימות עיבוד שפה טבעית ויכולה לסייע במשימות שונות:

- מנועי חיפוש מדויקים יותר – למשל נוכל לבצע חיפוש על השם "שרון" כשם אדם ולא לקבל בחזרה מסמכים העוסקים באזור השרון. משתמש יוכל לבקש לראות מסמכים המכילים מידע על אדם, מקום או ארגון מסוים ולקבל תוצאות מדויקות יותר.
- הכנה אוטומטית של אינדקס לספרים ופרסומים.
- הבלטה אוטומטית של שמות במסמך בהתאם לעניין הקורא (למשל בפרסומים כלכליים יתכן ונרצה להבליט שמות של חברות).
- בסיס למערכות תרגום אוטומטי.

- עבור השפה העברית, זיהוי שמות פרטיים בטקסט יכול לסייע בניתוח המשפט במקרים של עמימות הטקסט.
- בסיס למערכות המתעסקות במשימות שונות של הוצאת מידע מטקסט.

1.3 ייחודה של השפה העברית

בשפה האנגלית המשימה פשוטה יחסית בשל השימוש באותיות גדולות בתחילת כל שם פרטי. לכן מצפים לאחוזי דיוק גבוהים במערכת אנגלית. הבעיות הנותרות באנגלית הן קביעת סוג התג וגבולות הביטוי. דוגמא לבעיה בקביעת סוג התג: המילה "Washington" משמשת הן כשם אדם והן כשם מקום. גם משימת קביעת גבולות הביטוי אינה תמיד פשוטה. למשל: האם במשפט " California's Silicon Valley" יש שני ביטויי שם מקום ("California" ו-"Silicon Valley") או שהמשפט כולו הינו ביטוי מקום אחד?
בנוסף קיימת בעיה באנגלית להבחין בשמות פרטיים אשר מופיעים בתחילת משפט (שכן באנגלית אות ראשונה במשפט תמיד תהייה אות גדולה).

בעברית המשימה מורכבת הרבה יותר. נתבונן בפסקה הבאה:

" האלוף ישראל זיו, חתם אחר הצהריים (ה') בקהיר עם עמיתיו המצרים על ההבנות החדשות, במסגרתן ייפרסו 750 שוטרי עלית של משמר הגבול המצרי, בתגבור מסוקים וכלי שיט, לאורך ציר פילדלפי. אתמול קיבל ההסכם, שמשנה מספר סעיפים בנספח הצבאי לחוזה השלום, את אישור הכנסת . גוף מיוחד שהוקם בצה"ל אמור לסייע למצרים בנושא ולהוות את מנגנון שיתוף הפעולה בין שתי המדינות."

מספר קשיים העולים בפסקה זו:

- "ישראל" מופיע כשם אדם ולא שם מקום.
- המילה "זיו" יכולה להתפרש גם כשם עצם.
- "ה'" מציין תאריך – יום חמישי.
- המילה "מצרים" מופיעה פעם כביטוי השתייכות לקבוצה (כמו "ישראלים") ופעם כשם מקום.
- המילה "עלית" יכולה להתפרש גם כשם של חברה.
- האם "משמר הגבול המצרי" הוא ביטוי אחד או רק "משמר הגבול"?
- המילה "הכנסת" יכולה להתפרש במספר מובנים, למשל במובן של "להכניס"
- המילים "בקהיר" "בצה"ל" "למצרים" מופיעות עם תחיליות.

יש לציין כי לעיתים החלטת התיוג אינה חד משמעית גם למתייגים אנושיים (למשל בביטוי "משמר הגבול המצרי") . בעיה זו מתעוררת גם עבור השפה האנגלית. לשם כך יש להגדיר כללי תיוג חד משמעיים ולוודא כי רמת הסכמה על כללים אלה גבוהה בקרב מתייגים אנושיים. באנגלית ההגדרה קיימת ב-[18]. עבור עברית הוגדרו במחקר זה קוים מנחים לתיוג (ראה נספח 1) ונבחנה רמת ההסכמה עליהם (ראה פרק 3: הסכמה).

בעברית יש שימוש באותיות אחידות לאורך כל המשפט. קשה יותר להבחין בשמות הפרטיים ולכן יש לאסוף מידע על תכונות המילה והקשרה במשפט.

שמות פרטיים בעברית, בעיקר שמות אדם, מגיעים ממקורות שונים. ישנם פעלים ("אסף", "רן") , שמות תואר ("יפה", "לבנה") ושמות עצם ("אילן","מודיעין", "תקוה") רבים המשמשים גם כשמות פרטיים. מקרים אלו הם רבים והופכים את משימת התיוג לקשה ומורכבת.

השפה העברית מורכבת יותר מסיבות נוספות. העברית עשירה מורפולוגית ובעלת רב משמעיות גדולה. שורש של מילה יכול לקבל צורות שונות בהתאם למאפיינים כגון: גוף, מין, זמן, כמות, סמיכות, שייכות ועוד. מאחר והעברית כתובה כיום ללא ניקוד ולעיתים נעשה שימוש בכתיב חסר, מילה מסוימת הכתובה הטקסט יכולה להתפרש במספר אופנים.

דוגמאות:

- במשפט: "דוד שמש כאחראי חדר אוכל" שתי המילים הרב משמעיות הן "דוד" ו-"שמש". למילה "דוד" שתי משמעויות כשם עצם (boiler ,uncle) ומשמעות אחת כשם עצם פרטי – שם אדם. למילה "שמש" משמעות כשם עצם (sun), פועל מהשורש ש.מ.ש וכן היא גם משמשת כשם משפחה. בנוסף לביטוי "דוד שמש" משמעות בפני עצמו. עם זאת הפירוש הנכון של המילה "דוד" הוא שם אדם ושל המילה "שמש" הוא פועל, זאת ניתן להבין רק כאשר הקשרן במשפט מובן.
- "המפלגה הנוצרית דמוקרטית בראשותה של אנג'לה מרקל ניצחה הלילה בבחירות ברוב דחוק". המילים "דמוקרטית" ו-"נוצרית" משמשות כחלק מביטוי שם פרטי "המפלגה הנוצרית דמוקרטית" ולא כשמות תואר. בשפה האנגלית הבחנה זו תהייה קלה יותר שכן כל המילים בביטוי יקבלו אות תחילית גדולה.
- "אביב הגיע, פסח בא". גם למילה "אביב" וגם למילה "פסח" משמעות הן כתאריך והן כשם אדם. המילה "פסח" יכולה לשמש גם כפועל מהשורש פ.ס.ח. שוב, לא ניתן לקבוע את המשמעות הנכונה מהתבוננות במילה בלבד ובמקרה זה גם לא ניתן להגיע להחלטה חד משמעית בבחירת סוג השם.

תכונה נוספת של העברית היא סדר חופשי של מילים לעומת אנגלית בה סדר המילים מובנה יותר. קשה יותר לקבוע את תפקיד כל מילה במשפט.

רב משמעיות זו מקשה על כל המשימות של עיבוד וניתוח אוטומטי של טקסט. מערכת לזיהוי שמות פרטיים מתבססת לרוב על מנתח מורפולוגי ומנתח חלקי דיבר. מנתחים אלו עלולים לטעות ולבחור את הפירוש השגוי של המילה.

בעברית קיימת תופעה של הצמדת מילות יחס לשמות ופועלים, תופעת ההצמדה (agglutination). למשל: "ביתי" – הבית שלי, "ראיתיו" – ראיתי אותו. תופעה זו מציבה בפנינו בעיה חדשה: בעיית פירוק המילה. חלוקת המשפט למילים אינה מספיקה. יש לחלק כל מילה לחלקים ממנה היא מורכבת כגון תחיליות, סופיות וסמיכות. גם בעיה זו אינה ניתנת לפתרון חד משמעי. היא מוסיפה אי סדר לשפה ופותחת פתח נוסף לטעויות.

שימוש במילים לועזיות נפוץ בשמות ארגונים וחברות. מילים לועזיות מופיעות לעיתים באותיות לטיניות ולעיתים באותיות עבריות. במעבר לשפה העברית יתכנו צורות שונות של כתיבה. קשה להכניס למילונים את כל האפשרויות לכתיבה.

שימוש בלוח שנה עברי הוא תכונה ייחודית נוספת. במשימת זיהוי התאריכים בעברית יש לקחת בחשבון תאריכים עבריים כגון: "ה' באייר תש"ח". בנוסף, שמות ימים ("שלישי", "ג'") מופיעים במספר צורות ולעיתים ללא המילה "יום" לפניהם בניגוד לשפה האנגלית בה המילה יום היא חלק מהשם (Sunday).

1.4 הערכת מערכות לזיהוי שמות פרטיים

הערכת ביצועיה של מערכת מתבצעת בהשוואה לתיוג ידני של טקסט. במשימת התיוג הידני ייתכנו תיוגים שונים לטקסט זהה ולכן חשוב לתייג ע"פ כללים חד משמעיים המוגדרים מראש. כמובן שלא ניתן לצפות ממערכת אוטומטית לקבל החלטה חד משמעית במקומות בהם יש בלבול גם אצל מתייגים אנושיים.

נהוג להשתמש בשלושה מבחנים להערכת מערכת אוטומטית בהשוואה לתיוג ידני (אשר נחשב כ"אמת") : [5]

- מבחן הגבולות (TEXT): בדיקה האם גבולות הביטוי נכונים ללא התייחסות לסוג התג שניתן. לדוגמא נתבונן בביטוי "אוניברסיטת תל אביב". יתכן והמערכת תתייג באופן הבא: "אוניברסיטת" – שם ארגון, "תל אביב"- שם מקום. התיוג הנכון הוא "אוניברסיטת תל אביב" – שם ארגון. בתיוג המתואר המערכת נכשלה במבחן הגבולות. אם הביטוי כולו יקבל תג של שם מקום התיוג ייחשב כנכון במבחן הגבולות למרות שסוג התג שגוי.
- מבחן ה"סוג" (TYPE): בדיקה האם התג שקיבלה מילה מסוימת נכון ללא התייחסות לגבולות הביטוי. לדוגמא: אם נתייג את הביטוי "אוניברסיטת בן גוריון" כשני ביטויי ארגון נפרדים התיוג ייחשב כנכון במבחן הסוג.
- המבחן המשולב (TEXT & TYPE): בדיקת נכונות גבולות הביטוי וסוג התג שהוא קיבל.

במבחן הסוג ובמבחן המשולב ישנן ארבע תוצאות אפשריות:

- נכון (correct) – תג ידני זהה לתג שהתקבל ע"י המערכת.
- לא נכון (incorrect) תג ידני שונה מהתג שהתקבל ע"י המערכת.
- חסר (missing)– תג ידני קיים אולם המערכת תייגה את הביטוי/מילה כ"לא שם פרטי"
- מזויף (spurious) – המערכת תייגה את הביטוי אשר לא קיבל תיוג ידני.

במחקר זה נשתמש במבחן המחמיר – המבחן המשולב. מבחן זה אינו לוקח בחשבון ביטויים אשר תויגו חלקית.

המערכת נמדדת במושגים של *precision*, *recall* ו-*f-measure* אשר מוגדרים באופן הבא :

$$\boldsymbol{RECALL} = \frac{correct}{correct + incorrect + missing}$$

מדד ה-*recall* מציג את היחס בין מספר השמות הפרטיים אשר זוהו נכונה לבין סה"כ שמות פרטיים בטקסט. מדד זה מסמן את אחוז השמות שהמערכת הצליחה לזהות.

$$\boldsymbol{PRECISION} = \frac{correct}{correct + incorrect + spurious}$$

מדד ה-*precision* מסמן את הדיוק של המערכת – כמה מהשמות הפרטיים שזוהו הם אכן שמות פרטיים. קיימת תופעה של חילופין בין *recall* ל-*precision* כאשר אחד עולה, השני יורד וההפך. לכן מוגדר מדד המשולב ביניהם:

$$\boldsymbol{F} = \frac{2 * PRECISION * RECALL}{PRECISION + RECALL}$$

2 עבודות קודמות

כאמור, עבודה רבה נעשתה בבנית מערכות לזיהוי שמות פרטיים בשפה האנגלית ובשפות נוספות. השוואה בין ביצועים של מערכות התאפשרה במסגרת תחרויות שנעשו בין מערכות אשר אומנו על אותו קורפוס אימון ונבדקו על אותו קורפוס בדיקה. תחריות כאלה היו MUC-7, CoNLL2002 ו- CoNLL2003[23]. השוואה בין מערכות שלא השתתפו באותה התחרות קשה. אין מדדים מוחלטים לקביעת ביצועי המערכת, תוצאות הבדיקה תלויות במידה רבה בקורפוס האימון ובקורפוס הבדיקה אשר עמדו לרשות המערכת.

בהמשך מוצגת סקירה של העבודות המרכזיות אשר נעשו בתחום ואשר הציגו תוצאות טובות בתחרויות אשר בהן השתתפו או אשר הציגו גישה מעניינת לבעיה.

2.1 הגישה הידנית

מערכות שנבנו בשיטה הידנית מבוססות על האינטואיציה של המפתחים האנושיים שלהן. מערכות אלה מורכבות ממספר רב של חוקי רדוקציה רגישים להקשר.

גנדי למברסקי סוקר בעבודת התיזה שלו [8] את המערכות הידניות הקיימות עבור משימת זיהוי השמות הפרטיים באנגלית:

מערכות לזיהוי שמות פרטיים שנבנו בשיטה הידנית והציגו תוצאות טובות הן: **IsoQuest** (השיגה ב -MUC-7 f-measure=91.6%), מערכת **Proteus** של אוניברסיטת ניו יורק (השיגה ב MUC-7 f-measure=88.19%), מערכת FOCILE (השיגה ב MUC-7 f-measure=81.9%).

מערכת **Proteus** [1] נכתבה בשפת lisp עבור השפה האנגלית. המערכת מורכבת ממספר גדול של כללים רגישים להקשר. כללים אלה ברובם אינטואיטיביים, אולם לכל כלל במערכת ישנם יוצאים מן הכלל רבים. בגלל מורכבות השפה, אי אפשר ליצור כללים הנכונים תמיד.

להלן דוגמאות לכללים מתוך Proteus ולמקרים בהם הכלל נכון או לא נכון:

- Title Capitalized_Word -> Title Person_name
  - **Correct:** Mr. Jones, Gen. Schwarzkopf
  - **Incorrect:** Mrs. Field's Cookies (A corporation), Mr. Ten-Percent (nick-name for a corrupt third-world official)
- from Date to Date -> Date
  - **Correct:** from August 3 to August 9, 1999

• **Incorrect:** We moved the conference from April to June.

מהדוגמאות ניתן לראות שעבור כל כלל ניתן למצוא מיקרים בהם הוא לא תקף. אי אפשר לנסח כללים עבור כל יוצא דופן ולכן לכל מערכת ידנית ישנה קבוצת חוקים המיישבים מחלוקות. כלומר, במידה וכמה חוקים מתאימים לאותו המשפט, חוק מיישב מחלוקת יבחר את הפירוש המתאים למשפט. תכונה זו של המערכת הידנית מכבידה על פיתוחה, שכן עם כל הוספה של חוק חדש למערכת יש לבדוק היכן הוא מתנגש עם חוקים אחרים ולנסח חוקים מיישבי מחלוקת בהתאם. מערכת ידנית אינה גמישה לשינויים, יתכן ששינוי חוק אחד ידרוש עבודה רבה בעדכון שאר מערכת החוקים.

חסרון בולט של המערכות הידנית הוא זמן הפיתוח. פיתוח מערכת ידנית טובה דורש כוח אדם מקצועי וזמן רב לניסוח הכללים. למשל, לבניית מערכת IsoQuest נדרשו שנתיים. ביצועי המערכת יהיו תלויים במידה רבה ביכולתו של כוח האדם ובכמות העבודה (והכסף) שהושקעו במערכת.

מערכות ידניות נבנות לתמיכה בשפה ספציפית ובמבנה מסוים של מאמרים. מערכות אלה מציגות ביצועים טובים על פורמט קבוע. לדוגמא, המבנה של מאמרים מעיתון "הארץ": תאריך עברי, תאריך לועזי, כותרת המאמר, מחברי המאמר ובסוגריים שם הצלם. כלל המזהה את תבנית המאמר ישפר ללא ספק את ביצועי המערכת על מאמרי "הארץ". אולם במאמרים אחרים הכלל חסר משמעות ואף יכול להזיק.

מעבר לשפה אחרת ידרוש מערכת שונה לחלוטין. מעבר לפורמט אחר יהיה ידני וידרוש שיכתוב של הכללים. כל מעבר כזה יהיה יקר מכיוון שיהיה תלוי בעבודה רבה של בלשן חישובי.

עם זאת, לא ניתן להתעלם מיכולותיה של המערכת הידנית. ניתן לבנות מערכות ידניות חזקות מאוד. בנוסף, ישנן תבניות הניתנות לזיהוי בקלות ע"י כללים תחביריים ולכן שילוב כללים ידניים (למשל תבניות של ביטויים רגולריים) במערכות שאינן ידניות יכול להיות מאוד יעיל ולשפר ביצועים.

2.2 גישות אוטומטיות: מודלי מרקוב

2.2.1 מודל מרקוב סמוי (Hidden Markov Model)

מודל מרקוב מתאר תהליכים סטוכסטיים, בהם אירוע (משתנה מקרי) תלוי במספר סופי וקבוע של אירועים קודמים. סדר המודל קובע את מספר האירועים הקודמים בהם תלוי כל אירוע. להלן נציג מודל מסדר 1. במודל כזה, אירוע בזמן t תלוי רק באירוע הקודם לו (זמן t-1). קל להכליל מודל זה למודלי מרקוב מסדרים גבוהים יותר. האירועים מתוארים במודל ע"י קבוצת מצבים. במודל מרקוב סמוי (HMM) כל מצב פולט בהסתברות מסוימת סמל. המודל נקרא "סמוי" שכן בדרך כלל סדרת הסמלים הנפלטים ידועה, בעוד סדרת המצבים שפלטה אותם אינה ידועה.

מבחינה פורמאלית, מודל מרקוב סמוי M מוגדר כשלישיה: $M=(\Sigma,Q,\Theta)$ כאשר:

- Σ - א"ב הסמלים.
- Q - קבוצה סופית של מצבים המסוגלים לפלוט סמלים מא"ב Σ.
- Θ - קבוצת ההסתברויות המורכבת מ:
  - o הסתברויות מעבר בין מצבים. נסמן a_{kl} עבור כל $k,l\in Q$.
  - o הסתברויות פליטה. נסמן $e_k(b)$ עבור כל $k\in Q$ ו- $b\in\Sigma$.

בהינתן מודל מרקוב, תהליך מרקובי הוא תהליך בו נוצרת שרשרת סמלים במהלך מעבר בין מצבים באופן הסתברותי על פי המודל.

על מנת לבנות מודל יש לחשב את שלושת מרכיביו. המודל נבנה על בסיס דוגמאות לרצפי סמלים שנפלטו ע"י המודל. אם רצף המצבים שייצר דוגמאות אלו אינו ידוע המודל נבנה ע"י אלגוריתם באום-וולש. אנו נתרכז במקרים בהם רצף המצבים ידוע. במקרים אלה ניתן להעריך את מרכיבי המערכת באמצעות אומדני נראות מקסימלית (maximum likelihood estimators) . המודל מחושב ע"י שימוש בקבוצת אימון בה ידועות הן סדרות המצבים והן סדרות הפלטים. נוכל לסרוק את סדרת המצבים וסדרת הפלטים ולחשב:

- A_{kl} - מספר המעברים ממצב k למצב l.
- $E_k(b)$ - מספר הפעמים בהם מצב k פלט סמל b.

נחשב אומדני נראות מקסימלית:

$$a_{kl} = \frac{A_{kl}}{\sum_{q\in Q} A_{kq}}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{\sigma\in\Sigma} E_k(\sigma)}$$

בהינתן HMM ורצף פלטים, נרצה לחשב את רצף המצבים הסביר ביותר שהביא ליצירת רצף הפלטים. ניתן להתייחס למודל כאל מרחב חיפוש, שבו יש תת מרחב של כל המסלולים, המתאימים לרצף פלטים נתון, מהם יש למצוא את רצף המצבים בעל ההסתברות הגבוהה ביותר. האלגוריתם המתאים ביותר לבעיה זו הוא אלגוריתם תכנון דינאמי ויטרבי (Viterbi).

2.2.2 מערכת לזיהוי שמות פרטיים המבוססת על HMM

את בעיית זיהוי שמות פרטיים ניתן להציג באופן הבא: בהינתן רצף מילים $(w_1, w_2, ..., w_n)$, נרצה למצוא את רצף התגים $(t_1, t_2, ..., t_n)$ בעל ההסתברות הגבוהה ביותר. נניח שרצף התגים ממודל ע"י תהליך מרקובי ושההסתברות לתיוג מילה מסוימת תלויה רק בחלון קבוע של הקשר. למשל, במודל מרקוב מסדר 1 תיוג המילה הנוכחית יהיה תלוי רק במילה ובתג הקודמים. אם נחשב הסתברויות בהתחשב במילה הנוכחית בלבד נקבל שההסתברות לתג t_i במילה w_i תלויה בתדירות המילה המתויגת בתג זה בקורפוס האימון.

במקרה של זיהוי שמות פרטיים, המצבים החבויים הם התגים, והפלטים הם המילים בקורפוס. המודל נבנה ע"י שימוש בקורפוס אימון מתויג. בהינתן מודל ומשפט לתיוג, יש למצוא את סדרת התגים הסבירה ביותר שפלטה את המשפט.

עבודות רבות נעשו על בעיית זיהוי שמות פרטיים בשימוש ב-HMM בשפות שונות. למשל: אנגלית [12], גרמנית [2], ספרדית והולנדית [4] וכן יפנית וסינית [3].

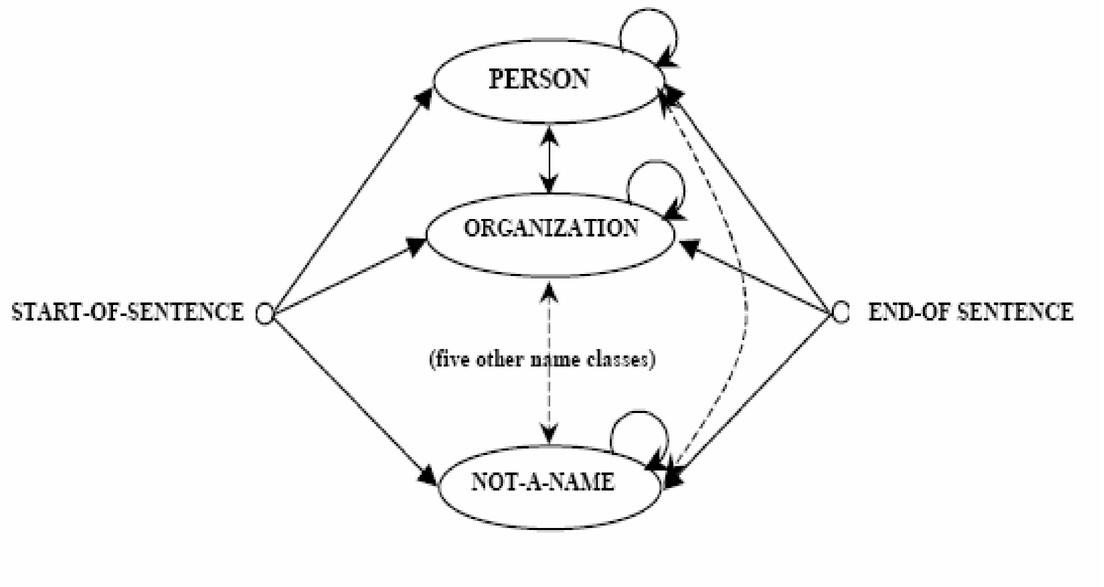
עבודה שהוצגה ב-CoNLL2002 ע"י **Malouf** [4] הציגה מערכת לזיהוי שמות פרטיים בהולנדית, שנבחנה גם על קורפוס ספרדי. המודל נבנה ע"י שימוש במודל מרקוב סמוי פשוט כפי שהוסבר לעיל. ההנחה הבסיסית היא שהסתברויות המעבר והסתברויות הפליטה הן בלתי תלויות ולכן:

$$P(w_i \mid t_i, t_{i-1}) = P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

המערכת הראתה תוצאות גבוהות מקו ההתחלה בהולנדית וספרדית אך נמוכות ממערכות שהשתמשו בשיטת האנטרופיה המקסימלית (שתפורט בהמשך) בשתי השפות. המגבלה העיקרית של שימוש במודל מרקוב פשוט הוא הנחת האי-תלות, הנחה שככל הנראה לא מתקיימת במקרה של זיהוי שמות פרטיים.

הנחה זו לא קיימת במערכות האנטרופיה המקסימלית. חסרון נוסף של המודל הפשוט הוא אי היכולת לשלב תכונות שונות של המילה וההקשר בחישוב הסתברויות המעבר והפליטה.

עבודה מוקדמת, אשר הוצגה ב-MUC-7 היא מערכת "**IdentiFinder**" [12], המציגה מודל מרקוב מורכב יותר. המערכת מומשה לשפה האנגלית והשיגה תוצאה של F=90.44% במשימת MUC-7.
המערכת משתמשת ב-HMM הירכי.
המצבים במודל מחולקים לאזורים, כאשר קיים איזור עבור כל תג וקיים איזור עבור התג "לא שם פרטי".
בנוסף קיימים שני מצבים עבור תחילת משפט וסיום משפט. המודל מתואר באיור הבא:



איור 1-2: מבנה ה- "IdentiFinder"

בתוך כל איזור קיים מודל סטטיסטי אשר פולט מילה אחת עבור כל מצב. כלומר, עבור כל תג קיים מודל סטטיסטי של המילים הנפלטות על ידו. לכן מספר המצבים במודל המילים הוא כמספר המילים בקורפוס.
בנוסף קיים מצב "סוף" עבור כל תג המתאר את סוף הביטוי.
ההסתברות למעבר למצב הבא תלויה במצב הנוכחי, במילה הנוכחית ובמאפיינים של המילה. המאפיינים מתארים תכונות של המילה כגון: ביטוי מספרי, מילה המתחילה באות גדולה, מילה שכולה אותיות גדולות, מילה ראשונה במשפט, מילה המופיעה בלקסיקון מילים חשובות, ועוד.
למודל ההירכי יתרון בכך שהוא מאפשר הגדרת מודל מילים שונה עבור כל תג. עבור כל תג מוגדרות הסתברויות שונות לפליטת מילים וכן ישנה התייחסות למאפיינים שונים עבור כל תג.
המודל נבנה ע"י שימוש בקורפוס אימון מתויג המכיל כ-790,000 מילים ו-65,500 שמות פרטיים.
בהינתן מודל ומשפט לחיזוי יש לבצע חיפוש ויטרבי למציאת רצף התגים הסביר ביותר כמו שתואר במודל הקודם. החיבור בין שני המודלים נעשה באופן הבא: מתחילים במודל הראשי במצב "תחילת משפט", כשנכנסים לאזור-תג מסוים עוברים למודל המילים המתאים, כשמגיעים למצב "סוף" במודל המילים חוזרים למודל הראשי וממשיכים הלאה לאיזור-תג הבא.

כאמור, מערכת "IdentiFinder" השיגה תוצאות טובות ב-MUC-7. מתכנתי המערכת טענו שע"י הגדלת קורפוס האימון ניתן להעלות את ביצועי המערכת ב-2%-4%.

מעבר ליתרון של יצירת מודל שונה עבור כל תג, מערכת זו משלבת התייחסות למאפיינים של המילה ולא רק למילה עצמה כפי שמתואר במודל הפשוט.

חסרון ה-"IdentiFinder " הוא שהמערכת משתמשת במאפיין אחד בלבד בו זמנית. כתוצאה מכך, המערכת מתקשה למדל מצב שבו למילה מספר מאפיינים. לדוגמא: המילה "MUC-7" כתובה באותיות גדולה ומכילה מספר ומקף. המערכת אינה מסוגלת להחליט איזה מהמאפיינים חשוב יותר לחיזוי וכן אינה מסוגלת להשתמש בשניהם.

2.2.3 HMM מבוסס תוים

גישה מעניינת לבעיית זיהוי שמות פרטיים הוצגה במשימת CoNLL2003 [23] ע"י **Klein** , **Smarr**, **Nguyen** ו-**Manning** מאוניברסיטת סטנפורד [2]. המודל שהוצג עושה שימוש בתווים במקום במילים.

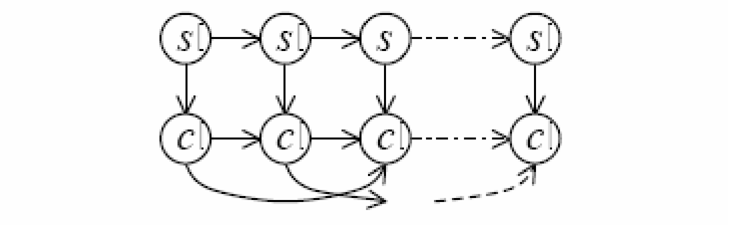
במהלך המחקר פותחה גישה של התייחסות לתווים במקום למילים שלמות. גישה זו באה להתמודד עם בעיית ההתמודדות עם כמות גדולה של מידע . המידע הנתון מפוזר ויש עדויות מעטות לכל מאורע. לכן נדרשים מודלים מתוחכמים וגדולים המתייחסים לכל מילה בקורפוס. רבים מהמאפיינים המשמשים מערכות קיימות מתייחסים לחלקים של המילה: אותיות גדולות, ספרות, תחיליות, סופיות, ניקוד וכו'.

גישת התווים מאפשרת הקטנה של המודל (מספר קטן של תווים לעומת מספר עצום של מילים בקורפוס ללא איבוד של מידע ומאפיינים.

ב- HMM מבוסס תווים, כל מצב פולט תו בודד. המצב הנוכחי תלוי במצב הקודם וב-n-1 התווים הקודמים.

הצגה גרפית של המודל: s – מצב תג

c – תו נפלט



איור 2-2: HMM מבוסס תוים

קביעת המצבים והסתברויות המעבר: יש לדאוג שכל התווים במילה יקבלו את אותו התג. ניתן להבטיח זאת ע"י פונקצית המעבר בין המצבים. כל מצב הוא זוג (t,k) כאשר t סוג התג ו-k משך הזמן שהמערכת נמצאת בתג t. למשל, $(PERS,2)$ מייצג את המצב שהמערכת נמצאת באות השנייה בתוך ביטוי שם אדם (PERS). לסיום ביטוי יש מצב מיוחד. מצב סיום ביטוי t יהיה (t,F) k חסום (אמפירית נקבע להיות 6),

ונשאר קבוע במידה ואורך הביטוי גדול מ-k תוים. כמו כן יש לקבוע את פונקצית המעבר כך שנמנע אי עקביות. לדוגמא, נדאג שמצב *(PERS,2)* יוכל לעבור רק למצב *(PERS,3)* או למצב *(PERS,F)* ולא לאף מצב אחר. נוכל לעשות זאת ע"י הצבת הסתברות 0 למעבר לכל שאר המצבים. מצבים סופיים יוכלו לעבור רק למצבי התחלה.

קביעת הסתברויות פליטה: יש לקבוע הסתברות לפליטת תו c_0 בהינתן n התווים הקודמים והמצב s הנוכחי. כלומר יש למצוא את $P\left(c_0 \mid c_{-1,\ldots}, c_{n-1}, s\right)$.

ההסתברויות חושבו בהתבסס על קורפוס האימון. לאחר יצירת המודל, זיהוי שמות במשפט חדש נעשה ע"י חיפוש ויטרבי.

ניסיון לשלב מילונים ומידע חיצוני הוריד את ביצועי המערכת.

המערכת השיגה ממוצע של F=83.2% ב-CoNLL2003 על טקסטים בשפה האנגלית. תוצאות אלה גבוהות בכ-30% מתוצאות של HMM פשוט מבוסס מילים שהופעל על אותו הטקסט. תוצאות אלה טובות מאוד בהתחשב בעובדה שמעבר לחלון של 6 תווים אין שימוש במאפיינים תחביריים מורכבים יותר. מחקר זה הדגיש את ההנחה שמידע מקומי הוא המידע החשוב ביותר בכל הקשור לזיהוי שמות פרטיים במשפט. התוצאות הטובות הביאו לבניית מערכת אנטרופיה מקסימלית המבוססת אך ורק על מאפיינים מקומיים.

2.3 גישות אוטומטיות: שיטת אנטרופיה מקסימלית

2.3.1 אנטרופיה מקסימלית

בשנים האחרונות נעשה שימוש נרחב בשיטות של הסתברות מותנית במודלים של NLP. שיטות אלה נותנות תוצאות מדויקות ומאפשרות שילוב של מאפיינים בלשניים רבים. שיטות אלה מאפשרות בנייה אוטומטית של מודלים ללא תלות בשפה.

עקרון האנטרופיה המקסימלית [10] [13] משמש לבניית מודל סטטיסטי הסתברותי. ע"פ עקרון זה על המודל לשקף את כל המידע הנתון, אך להימנע מלהניח דבר על מה שלא ידוע. במילים אחרות, אם לא נמצא מידע המבחין בין שני אירועים, יש לשער כי שניהם סבירים באותה מידה.

בבניית המודל נשתמש במידע רציף (משפטים בטקסט), נוציא ממנו מאפיינים מקומיים ונבנה מודל הסתברותי.

נגדיר מספר מושגים:

- **אירוע** - זוג סדור (x,t) כאשר x תוכן מסוים ו- t התג המתאים לו. למשל, במשימת תיוג השמות הפרטיים ניתן להציג קורפוס מתויג כאוסף של אירועים: כל מילה בקורפוס עם התג המתאים לה.
- **מאפיין**- הפונקציה $f(x,t)$ היא עדות בסיסית המקשרת בין התוכן הנצפה x לבין התג t שאותו אנו רוצים לחזות. בד"כ מאפיין הוא פונקציה בינארית המוגדרת על כל תחום האירועים האפשריים והבוחרת תת קבוצה מהמרחב. במשימת זיהוי שמות פרטיים מאפיינים יכולים להיות תכונות תחביריות (למשל חלקי דיבר), או תכונות הקשורות למקורות חיצוניים כגון מילונים.

  דוגמאות למאפיינים (x היא מילה בקורפוס):

$$f(x,t)\begin{cases} 1\ x \text{ is capitalized and } t = \text{LOCATION} \\ 0 \text{ otherwise} \end{cases}$$

$$f(x,t)\begin{cases} 1\ X \text{ is all capital letters and } t = \text{ORGANIZATION} \\ 0 \text{ otherwise} \end{cases}$$

על מנת לבנות מודל הסתברותי המבוסס על עקרון האנטרופיה המקסימלית, יש צורך בקורפוס מתויג, אוסף מאפיינים ושאר המידע הנחוץ לחישוב המאפיינים עבור כל מאורע בקורפוס.

בחירת המאפיינים היא השלב המרכזי והחשוב היותר בבניית מודל מוצלח. זהו תהליך של ניסוי וטעייה.

נניח שיש בידינו קורפוס מתויג ואוסף מאפיינים מוצלח. נרצה ליצור מודל הסתברותי $P(t \mid x)$ אשר בהינתן תוכן x יודע לחשב את ההסתברות שלו לקבל תג t. החיזוי בנקודה מסוימת ברצף תלוי רק במאפיינים הפעילים באותה הנקודה (כלומר המאפיינים שערכם שווה ל-1). במודל המבוסס על מאפיינים יתכן ולכל מאפיין תהייה השפעה שונה על חיזוי התג ולכן נרצה נצמיד לכל מאפיין משקל, כלומר לקבוע עד כמה המאפיין משפיע על החיזוי. המשקולות הם פרמטרים של המודל ונקבעים ע"פ עקרונות האנטרופיה המקסימלית ע"פ קורפוס אימון כפי שיתואר בהמשך.

2.3.2 אילוצים על המודל

- לכל מאפיין f_i נדרוש שהתוחלת ביחס שלו ביחס לקורפוס $\widetilde{E}(f_i)$ תהיה שווה לתוחלת שלו ביחס למודל $E(f_i)$. כלומר:

$$\forall i : \widetilde{E}(f_i) = E(f_i)$$

התוחלת של f_i ביחס לקורפוס נתונה ע"י:

$$\widetilde{E}(f_i) \equiv \sum_x \sum_t \widetilde{p}(x,t) f_i(x,t)$$

כאשר $\widetilde{p}(x,t)$ היא ההסתברות האמפירית של אירוע (x,t) המוגדר כמספר הפעמים בהן האירוע מופיע בקורפוס ביחס לסך כל האירועים בקורפוס.
התוחלת של f_i ביחס למודל נתונה ע"י:

$$E(f_i) \equiv \sum_x \widetilde{p}(x) \sum_t p(t \mid x) f_i(x,t)$$

כאשר $\widetilde{p}(x)$ היא ההסתברות האמפירית של x (תדירות המילה בקורפוס).
משילוב המשוואות נקבל:

$$\forall i : \sum_x \sum_t \widetilde{p}(x,t) f_i(x,t) = \sum_x \widetilde{p}(x) \sum_t p(t \mid x) f_i(x,t)$$

- כאמור, דרישה נוספת על המודל היא להימנע מלהניח דבר על מה שלא ידוע. נדרוש מודל שיש בו הכי פחות ודאות. לשם כך נשתמש במושג אנטרופיה.

אנטרופיה מוגדרת כרמת אי הוודאות של ההתפלגות. או במילים אחרות רמת ה"הפתעה" של המודל. אי ודאות של הסתברות מותנית נמדדת ע"י אנטרופיה מותנית *H(p)* הנתונה ע"י:

$$H(p) \equiv -\sum_{x} \tilde{p}(x) \sum_{t} p(x \mid t) \log p(t \mid x)$$

נרצה לבחור במודל בעל האנטרופיה המקסימלית מבין כל המודלים המספקים את הדרישה לגבי תוחלת המודל.

קיבלנו בעיית אופטימיזציה עם אילוצים. בכדי לפתור את הבעיה נעשה שימוש בכופלי לגרנז'. ניתן להסתכל על ערכי כופלי לגרנז' במערכות אנטרופיה מקסימלית כמשקולות של המאפיינים. כלומר החשיבות של המאפיין בקביעת ההתאמה של תג מסוים למילה.

2.3.3 אלגוריתם GIS לבניית מודל אנטרופיה מקסימלית

האלגוריתם המשמש לבניית מודל על בסיס קורפוס אימון וקבוצת מאפיינים הוא Generalized Iterative Scaling (GIS) של **Darroch** ו-**Ratcliff** [14].

בהינתן אוסף מאפיינים $f_i(x,t)$ והתוחלת שלהם ביחס לקורפוס אימון $\tilde{E}(f_i)$, כל איטרציה של האלגוריתם מחשבת את הערכים החדשים λ_i (משקלות המאפיינים) עד להתכנסות.

אתחול ההסתברויות נעשה ע"י (Z הוא קבוע הנרמול):

$$\lambda_i^{(0)} = 0$$

$$p^{(0)}(t \mid x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,t)\right)}{Z_\lambda(x)}$$

$$Z_\lambda(x) = \sum_t \exp\left(\sum_i \lambda_i f_i(x,t)\right)$$

בכל איטרציה j מתבצעים השלבים הבאים:

1. חישוב התוחלת של $f_i(x,t)$ ע"פ המודל הנוכחי:

$$E^{(j)}(f_i)=\sum_x \widetilde{p}(x)\sum_t p_\lambda^{(j)}(t\mid x)f_i(x,t)$$

2. עדכון הפרמטרים:

$$\lambda_i^{(j+1)}=\lambda_i^{(j)}+\log\frac{\widetilde{E}(f_i)}{E^{(j)}(f_i)}$$

3. חשב את המודל הבא $p^{(j+1)}(t\mid x)$ על סמך הפרמטרים החדשים ע"פ אותה נוסחא.
4. המשך עד להתכנסות.

2.3.4 שימוש במודל

לשם שימוש במודל האנטרופיה המקסימלית במשימת תיוג השמות הפרטיים יש להשתמש בקורפוס מתויג וקבוצת מאפיינים המתאימים למשימה זו. כשיש בידינו מודל $P(t\mid x)$ נוכל להשתמש בו לזיהוי שמות פרטיים בטקסט חדש.

בשלב הראשון יש לאסוף מאפיינים עבור כל מילה בטקסט. בשלב השני נחשב הסתברות מותנית עבור כל אחד מהתגים האפשריים. לבסוף נחפש את רצף התגים בעל ההסתברות הגבוהה ביותר עבור רצף המילים הנתון ע"י שימוש באלגוריתם חיפוש רצף כגון ויטרבי או Beam Search.

2.3.5 מערכות זיהוי שמות פרטיים המבוססות על מודל אנטרופיה מקסימלית

עבודות רבות נעשו על בעיית זיהוי שמות פרטיים בשימוש בשיטת האנטרופיה המקסימלית בשפות שונות. מודלים אלה מתמודדים טוב עם מספר גבוה של מאפיינים וכן עם מאפיינים מתנגשים כפי שהוכח ניסיונית [5]. קיימים מספר כלי-מדף [15] [17] לבניית מודלי אנטרופיה מקסימלית, ולכן העבודה העיקרית של המפתחים היא בחירת המאפיינים.

אחת העבודות הראשונות שנעשו לפתרון בעיית שמות פרטיים בשימוש במודל האנטרופיה המקסימלית היא מערכת **MENE** שפותחה ע"י **Andrew Borthwick** מאוניברסיטת ניו יורק [5].

MENE כתובה בשפת ++C ופרל ומשתמשת ב-MEMT לחישוב ההתפלגות על סמך קורפוס אימון.

MEMT הוא כלי מדף שיודע לחשב את ערכי המודל ע"פ עקרונות אנטרופיה מקסימלית בהינתן קבוצת מאפיינים, קבוצת תגים וקורפוס אימון מחולק לפי מילים [15].

הקורפוס של Borthwick הכיל כ-300,000 מילים. הוא בחר מאפיינים באופן אמפירי.

להלן חלק מהמאפיינים שנבחרו:

- ביטויים רגולריים כגון: "מילה המתחילה באות גדולה", "מילה המכילה מספר".

- מאפיינים לקסיקליים - מאפיינים הנוגעים למילים שמסביב למילה הנוכחית. נבחנו המילים בחלון של 2 מילים לפני ואחרי המילה הנוכחית. דוגמא:

$$f\left(w,t\right)\begin{cases}1 \text{ the word before } w \text{ is "Mr" and } t = \text{PER} \\ 0 \text{ otherwise}\end{cases}$$

- מאפייני section - מבנה הטקסט היה ידוע מראש. לכן הטקסט חולק לאזורים (תאריך, כותרת,טקסט) והוספו מאפיינים המשלבים את האזור בו הופיע האירוע.
- מילונים – מילונים הכילו ביטויי שמות פרטיים עבור ביטויים באורך מילה או יותר. לדוגמא השתמשו במילוני עבור: שמות אדם פרטיים, שמות חברות, שמות אוניברסיטאות ועוד.
- מערכות חיצוניות – מאפיינים של תיוגים של מערכות ידניות לזיהוי שמות פרטיים שולבו במודל בחלק מהבדיקות.

מערכת MENE הוצגה במשימת MUC-7 וזכתה במקום הרביעי עם F=84.22% (אחרי מערכות שניבנו בגישה הידנית וה-"IdentiFinder " שהוצגה בסעיף של HMM). בשילוב עם מערכות ידניות MENE הציגה תוצאות של F=92%. המערכת הציגה גם תוצאה של 77.98% על טקסט הכתוב כולו באותיות גדולות – התוצאה הגבוהה ביותר במשימת MUC-7 עבור טקסט כזה. החיסרון העיקרי של המערכת הוא דלות המאפיינים. מערכות אנטרופיה מקסימלית עם אוסף מאפיינים רחב יותר הציגו תוצאות טובות יותר כפי שנראה בעבודה הבאה.

Chieu ו- **Ng** מאוניברסיטת סינגפור הציגו במשימת CoNLL2003 מערכת לזיהוי שמות פרטיים המבוססת על עקרון האנטרופיה המקסימלית [7]. המערכת זכתה זכתה במקום השני עבור טקסט בשפה האנגלית.

המודל הוא מודל זהה למודל שהוצג לעיל ולמודל ששימש לבניית MENE. קורפוס האימון נלקח מתוך ידיעות של סוכנות רויטרס. המאפיינים המשמשים את המערכת הם אלה שהביאו לתוצאות הטובות.

להלן דוגמאות למאפיינים המשמשים את המערכת:

- ביטויים רגולריים ושילוב שלהם (כלומר מאפיינים המכילים שילוב של 2 או יותר ביטויים רגולריים נפרדים).
- מילים נדירות – בעיבוד מוקדם על הקורפוס נאספו כל המילים המופיעות ביותר מ-5 מאמרים. מילים שלא נמצאו ברשימה זו הוגדרו כנדירות.
- תחיליות/סופיות - בעיבוד המוקדם נאספו תחיליות וסופיות מיוחדות לכל מחלקה ע"י שימוש במטריצת קורלציה.
- בעיבוד המוקדם נאספו גם מילים המופיעות בסמוך לקטגוריה מסוימת בחלון של 2 מילים לפני ואחרי הקטגורי ע"י שימוש במטריצת קורלציה.

- מאפיינים בחלון של 2 מילים לפני ואחרי המילה הנוכחית
- מילונים
- תיוג קודם של המילה
- רצפים של מילים המתחילות האות גדולה
- עבור ביטויים המתחילים באות גדולה נבדק אם גם שאר ההופעות של הביטוי בשאר המסמך מתחילים באות גדולה.
- חלוקת הטקסט לאזורים ע"פ תבנית ידועה מראש של המסמכים בקורפוס.

המערכת נבחנה במשימת CoNLL2003 על קורפוס של סוכנות הידיעות רויטרס והשיגה ציון של F=88.31% עבור השפה האנגלית. היא זיהתה טוב שמות מקומות ושמות אנשים ברמה הגבוהה ביותר. המערכת התאימה מאוד לשפה האנגלית, אך המעבר לשפה חדשה לא היה מוצלח. עבור טקסט בשפה הגרמנית לא נצפו תוצאות טובות כמו בשפה האנגלית והמערכת הגיעה למקום שביעי ב-CoNLL2003 (F=65.67%). תכונות השפה הגרמנית שונות מאוד מתכונות השפה האנגלית ולכן נדרשים מאפיינים מיוחדים לשפה זו.

גנדי למברסקי מאוניברסיטת בן גוריון בנה במסגרת עבודת התזה שלו מערכת לזיהוי שמות פרטיים הטקסט בשפה העברית [8].

לבניית המודל נעשה שימוש בחבילת MaxEnt [17] שנכתבה במסגרת הפרויקט OpenNLP [16] ומבוססת על עקרון הקוד הפתוח. החבילה משמשת לצורכי אימון ויצירת מודל הסתברותי ע"פ עקרונות האנטרופיה המקסימלית ואלגוריתם GIS.

המאפיינים ששימשו את המערכת מותאמים לתכונות השפה העברית. להלן רשימת המאפיינים:

- שימוש במנתח מורפולוגי: ניסיון לפתור את בעיית הרב משמעיות של השפה העברית. המנתח משתמש בשיטות הסתברותיות ובלשניות כדי לבחור בפירוש הסביר ביותר עבור מילה (אחוזי הצלחה של המנתח – 85%).
- מילונים מעודכנים לעברית. למשל: לוח שנה עברי, שמות חברות וארגונים וכו'.
- מאפיין "לועזי" – מילים עם אותיות לועזיות.
- שימוש בלמה – הצורה המילונית של המילה ללא תוספת מילות יחס וסיומות.
- חלקי הדיבר של המילה
- תג קודם של הלמה
- ביטויים רגולריים לזיהוי תאריכים, ביטויי כסף וכו'.

לצורך הניסויים נבחרו ותויגו כ-50 מאמרים של מאמרי עיתון "הארץ" (קורפוס קטן מאוד ששימש בחלקו לאימון ובחלקו לבדיקה).

קו הבסיס לעבודה היו תוצאות שהתקבלו ע"י שימוש במילונים בלבד. קו הבסיס שהתקבל היה Fe=58.6%, והתוצאות נעו סביב F=76%-81% . המערכת הציגה אחוז דיוק גבוה (92%) ו-recall

נמוך (66%-70%). כלומר המערכת לא הצליחה לזהות אחוז גבוה של ביטויים אולם הביטויים שזוהו, זוהו נכון. ביטויים שלא זוהו הם ביטויי זמן ותאריך. הקושי נובע מהיעדר תבנית שתוכל לתאר ביטויים אלה. מכיוון שהקורפוס ששימש לצורכי אימון ובדיקה היה קטן מאוד קשה להסיק משהו על ביצועי המערכת. יחד עם זאת התוצאות מספקות בשל ההוכחה ששיטת האנטרופיה המקסימלית עובדת גם בעברית והדיוק גבוה.

2.4 גישות אוטומטיות: Robust Risk Minimization

Johnson ו-**Zhang** ממעבדות המחקר של IBM בני-יורק הציגו ב-CoNLL2003 מערכת זיהוי שמות פרטיים המבוססת על שיטת ה-Robust Risk Minimization [11]. זוהי שיטת אוטומטית הסתברותית מבוססת מאפיינים בדומה לשיטת האנטרופיה המקסימלית. המערכת הגיעה למקום הרביעי עבור טקסט בשפה האנגלית וכן הצליחה לעשות את המעבר לשפה הגרמנית בהצלחה יחסית והגיעה למקום השלישי בגרמנית.

עבודות מוקדמות עם שיטת זו על בעיית ה-text chunking הראו אחוזי הצלחה גבוהים. היתרון של השיטה הוא היכולת לשלב מספר רב של מאפיינים. בבניית המערכת הייתה התמקדות במאפיינים מקומיים המשפיעים בצורה החזקה ביותר על החיזוי.

המערכת מחשבת הסתברות מותנית לתג c בהינתן וקטור מאפיינים הקשורים למילה במקום ה-i x_i כלומר $P(t_i = c \mid x_i)$. המודל מניח שהתיוג הנוכחי תלוי במילה הנוכחית ובתיוגים הקודמים כלומר: $P(t_i = c \mid x_i) = P(t_i = c \mid \{w_i\}, \{t_j\}_{j \le i})$. בהינתן מודל שכזה ניתן למצוא את רצף התגים הסביר ביותר ע"י שימוש באלגוריתם תכנון דינמי.

במערכת זו ההסתברות המותנית היא מהצורה:

$$P(t_i = c \mid x_i, \{t_{i-l}, \ldots, t_{i-1}\}) = T(w_c^T x_i + b_c)$$

כאשר:

$T(y) = \min(1, \max(0, y))$ - העברה של הערך y אל האינטרואל $[0,1]$.

w_c - משקל הווקטור x_i. זהו פרמטר של המערכת ומוערך ע"י שימוש בקורפוס אימון.

b_c - קבוע. זהו פרמטר של המערכת ומוערך ע"י שימוש בקורפוס אימון.

בהסתמך על עבודה מוקדמת של Zhang , בהינתן קורפוס אימון מתויג ניתן לחשב את המודל ע"י פתרון בעיית האופטימיזציה עבור כל c:

$$\inf_{w,b} \frac{1}{n} \sum_{i=1}^{n} f(w_c^T x_i + b_c, y_c^i)$$

כאשר:

$$y_c^i = \begin{cases} 1 & t_i = c \\ -1 & \text{otherwise} \end{cases}$$

$$f(p,y) = \begin{cases} -2py & py < -1 \\ \frac{1}{2}(py-1)^2 & py \in [-1,1] \\ 0 & \text{py} > 1 \end{cases}$$

הפונקציה המתוארת נקראת פונקצית הסיכון (risk function) ובעיית האופטימיזציה המביאה את הפונקציה למינימום נקראת Robust Risk Minimization.

כאמור, אוסף המאפיינים היה ברובו מורכב ממאפיינים מקומיים. מאפיין מיוצג ע"י פונקציה בינארית . נלקחו מאפיינים הקשורים למילה הנוכחית וכן למילים בחלון סביבה. רשימת המאפיינים מתוארת בטבלה הבאה:

Feature ID	Feature description
A	Tokens that are turned into all upper-case, in a window of ±2.
B C	Tokens themselves, in a window of ±2. The previous two predicted tags, and the conjunction of the previous tag and the current token
D	Initial capitalization of tokens in a window of ±2.
E	More elaborated word type information: initial capitalization, all capitalization, all digitals, or digitals containing punctuations.
F	Token prefix (length three and four), and token suffix (length from one to four).
G H	POS tagged information provided in shared the task. chunking information provided in the shared task: we use a bag-of-word representation of the chunk at the current token
I	The four dictionaries provided in the shared task: PER, ORG, LOC, and MISC.
J	A number of additional dictionaries from different sources: some trigger words for ORG, PER, LOC; lists of location, person, and organizations.

טבלה 2-1: מאפייניי מערכת ה-RRM

המערכת קיבלה ציון של F=85.5% במשימת CoNLL2003 והגיעה למקום רביעי. התקבלו תוצאות טובות גם עבור טקסט שכולו כתוב באותיות גדולות (עליה של 6%-10% באחוזי השגיאה). במסגרת בניית המערכת נעשו בדיקות לבחינת השפעת המאפיינים המקומיים. ביצועי המערכת נבדקו ללא מאפיינים חיצוניים, כלומר ללא מילונים וגם במקרה זה נראו תוצאות טובות (עליה של 2% באחוזי השגיאה).

המעבר לשפה הגרמנית היה מוצלח יחסית למערכות אחרות והמערכת הגיעה למקום שלישי בשפה הגרמנית במשימת CoNLL2003 (F=71.27%).

עבודה זו חיזקה את המסקנה שהמאפיינים המקומיים הם המשמעותיים בחיזוי שמות פרטיים בטקסט ויש לבחור אותם בקפידה תוך תהליך ארוך של ניסוי וטעייה. כמו כן, בשל מורכבות השפה והמשימה של

זיהוי שמות פרטיים בטקסט מערכות הסתברותיות המסוגלות להתמודד עם מספר רב של מאפיינים, ומעריכות את משקל כל מאפיין ביחס לקטגוריה, הן המוצלחות ביותר. מערכות אלה ניתנות לבניית הצורה אוטומטית וקלה יחסית ומראות תוצאות טובות.

2.5 שילוב מערכות

התוצאות הטובות ביותר שנצפו במשימת CoNLL2003 הן בשפה האנגלית והן בשפה הגרמנית היו של מערכת שהוצגה ע"י **Florian** , **Ittaycheriah**, **Jing**, **Zhang** ממעבדות IBM בני-יורק [6]. עבודה קודמת של Florian [9] המציגה שילוב של מערכות קיימות הוצגה במשימת CoNLL2002 והגיעה למקום שני בספרדית ומקום שלישי בהולנדית.

המערכת הינה שילוב של ארבע מערכות המבוססות על שיטות סטטיסטיות:

1. מערכת Robust Risk Minimization (RRM) שתוארה בסעיף הקודם [11]
2. מערכת אנטרופיה מקסימלית (ME)
3. HMM היררכי בדומה לזה שהוצג לעיל
4. מערכת Transformation-Based Learning (TBL) - אלגוריתם מונחה טעויות בעל שני שלבים עיקריים: מתחיל בהשמת תגים כלשהם על קורפוס האימון וממשיך אוטומטית בבחירת התיוג שמוריד את מספר הטעויות בצורה המשמעותית ביותר.

כשבוחנים כל מערכת בנפרד, התוצאות הטובות ביותר התקבלו ע"י מערכת האנטרופיה המקסימלית וה-RRM .

שילוב המערכות מאפשר ניצול ההבדלים בין הגישות:

- TBL היא מערכת המשתמשת רק במאפיינים בולטים בעוד ME ו-RRM משתמשים במספר רב של מאפיינים.
- כל מערכת משתמשת בשיטת חיפוש שונה למציאת הרצף הסביר ביותר.
- הפלט של כל מערכת שונה. TBL ו-RRM נותנים תשובה אחת בעוד ME ו-HMM מחזירים הסתברויות עבור כל קטגוריה.

את ארבע המערכות שימשו מאפיינים מקומיים וכלליים כגון: חלקי דיבר, תחיליות/סופיות, מילונים וכו'.
ארבע המערכות אומנו על אותו קורפוס מתויג שסופק במשימת CoNLL2003 .

השלב המשמעותי ביותר בבניית מערכת שהיא שילוב של מערכות קיימות הוא בחירת שיטת ה"הצבעה", כלומר בחירת תיוג אחד מבין תוצאות תיוגי ארבע המערכות. נרצה להתחשב ביתרונות ובחסרונות של כל מערכת ליצירת מערכת אחרת חזקה יותר.

אפשרויות לשילוב מערכות:

- מתן זכות "הצבעה" שווה לכל מערכת.
- מתן משקל להחלטת כל מערכת. כל מערכת "מצביעה" לקטגוריה אחת בלבד.

- מתן משקולות שונים לכל מערכת בכל קטגוריה. כלומר, עבור מערכת מסוימת נקבע משקולות שונים עבור קטגוריות שונות. שוב, כל מערכת "מצביעה" לקטגוריה אחת בלבד.
- המערכות RRM ו-ME לא מחזירות החלטה אחת אלא סדרה של הסתברויות. נרצה להשתמש בהסתברויות אלו ולא להתייחס רק לקטגוריה עם ההסתברות הגבוהה ביותר. במקרים אלה נרצה להציב משקולות לכל הסתברות בנפרד.

לאחר ניסיונות בכל שיטות השילוב, האפשרות האחרונה הראתה את התוצאות הטובות ביותר. משקולות נקבעו ע"י שימוש בקורפוס אימון וע"י שימוש בפונקצית הסיכון שראינו בשיטת ה-RRM.
כאמור השילוב הניב תוצאות טובות מאוד. עבור טקסט בשפה האנגלית התקבלה תוצאה של F=88.79%(מקום ראשון ב-CoNLL2003). גם המעבר לשפה הגרמנית היה מוצלח עם תוצאות של F 72.41% (מקום ראשון).

3 הסכמה

3.1 הגדרת משימת תיוג שמות פרטיים בטקסט

מערכת סטטיסטית זקוקה לקורפוס אימון מתויג על מנת לחשב את התפלגות ההסתברויות. על מנת לבנות קורפוס עברי מתויג תחילה יש להגדיר את משימת תיוג שמות פרטיים בשפה העברית באופן חד משמעי. לשם כך נוסחו הנחיות המתבססות על תכונות השפה העברית ועל הגדרת המשימה עבור השפה האנגלית [18]. לאחר ניסוח ההנחיות, יש לבחון עד כמה הן בהירות וחד משמעיות. בחינה זו נעשתה באמצעות מדידת רמת ההסכמה בניסוי בו קבוצת אנשים תייגה טקסטים שונים לאחר קריאת ההנחיות (וללא ידע מוקדם).

3.2 מבחני הסכמה

מבחני ההסכמה הוגדרו על בסיס מבחני הערכת מערכת לזיהוי שמות פרטיים שהוגדרו ע"י Borthwick [5].

3.2.1 מבחן גבולות (TEXT)

בדיקה האם יש הסכמה על גבולות הביטוי. במבחן זה לא נסתכל על סוג התג שקיבל הביטוי אלא רק על תחילתו וסופו.

דוגמא 1:

נתבונן בשני תיוגים שונים של המשפט :

"בית המשפט המחוזי בירושלים"

- מתייג א':

<ORG> בית המשפט המחוזי בירושלים </ORG>

- מתייג ב':

<ORG> בית המשפט המחוזי </ORG> <LOC> בירושלים </LOC>

ניתן לראות כי מתייג א' תייג את המשפט כולו כביטוי ארגון ואילו מתייג ב' התייחס למשפט כשני ביטויים שונים: הראשון של ארגון והשני של מקום. אין הסכמה על גבולות הביטוי למרות ששלושת המילים הראשונות של המשפט קיבלו תיוג זהה ע"י שני המתייגים.

דוגמא 2:

המשפט:

"מדינת ישראל"

- מתייג א':

<LOC> מדינת ישראל </LOC>

- מתייג ב':

<ORG> מדינת ישראל </ORG>

שני המתייגים נתנו תיוגים שונים לביטוי אולם שניהם הסכימו על הגבולות ולכן במבחן הגבולות נתייחס לדוגמא זו כאל הסכמה.

3.2.2 מבחן "הסוג" (TYPE)

במבחן זה נתייחס לתג שאותו קיבלה המילה ולא לגבולות הביטוי כולו. כלומר, נשווה בין התיוגים השונים שקיבלה אותה מילה ע"י המתייגים השונים.

דוגמא 3:

המשפט:

"חודש יולי 2005"

נתבונן בשלושה תיוגים שונים:

- מתייג א':

<DATE> חודש יולי 2005 </DATE>

- מתייג ב':

חודש <DATE> יולי 2005 </DATE>

- מתייג ג':

חודש <DATE> יולי </DATE> <DATE> 2005 </DATE>

עבור המילה "חודש" יש הסכמה בין מתייגים ב' ו-ג'.

עבור המילים "יולי" ו-"2005" יש הסכמה בין שלושת המתייגים למרות שגבולות הביטויים שונים.

3.2.3 המבחן המשולב (TEXT&TYPE)

מבחן המשלב את שני המבחנים הקודמים. נאמר שישנה הסכמה בין שני מתייגים רק אם גבולות הביטוי זהים בשני התיוגים וכן הביטוי קיבל את אותו התג בשניהם. זוהי הבדיקה המחמירה ביותר. לדוגמא, עבור שלושת המשפטים שהוצגו לעיל אין הסכמה כלל בין המתייגים במבחן זה.

3.3 חישוב הסכמה

ההסכמה בניסוי חושבה בשני שלבים: תחילה חושבה ההסכמה בין המתייגים בשיטת ההסכמה החלקית (partial agreement) ולאחר מכן חושבה ההסכמה בשיטת הקאפה (kappa statistics).

3.3.1 הסכמה חלקית

קיימות מספר שיטות לחישוב הסכמה בקבוצה בהנחה שכל חברי הקבוצה נחשבים כשווים. שתי שיטות עיקריות הן "הרוב קובע" ("majority rules") ושיטת ההסכמה החלקית. בשיטת "הרוב קובע" דעת הרוב תחשב כ"אמת" ונבדוק מהי רמת ההסכמה עם דעת הרוב. בשיטת ההסכמה החלקית כל דעה נחשבת ומקבלת ציון על רמת ההסכמה.
הציון עבור כל דעה בשיטת ההסכמה החלקית מחושב ע"י: מספר הקולות התומכים בדעה זו/ מספר הקולות עבור סך כל הדעות.

דוגמא:

נתבונן בתיוגים שהוצגו בסעיף הקודם בדוגמא 3 עבור המשפט:

"חודש יולי 2005"

נחשב הסכמה חלקית על תיוג המילים במשפט ע"פ מבחן ה-TYPE:
המילה "חודש" קיבלה בתיוג הראשון תג "תאריך" ובשני התיוגים באחרים תג "אחר". כלומר הציון עבור ההסכמה על התג "תאריך" יהיה 1/3 והציון עבור ההסכמה על התג "אחר" יהיה 2/3.
ע"י חישוב דומה נקבל שההסכמה עבור המילים "יולי" ו-"2005" היא 1.

אם גודל הקבוצה הוא d, וקטגוריה מסוימת קיבלה ציון הסכמה של 1/d הרי שאין הסכמה כלל על קטגוריה זו. הסכמה מלאה בתוך הקבוצה תקבל ציון 1.
ניתן לראות מהדוגמא שיתכן ומילה מסוימת (במבחן ה-TYPE, שכן במבחנים האחרים נתייחס לביטויים ולא למילים) תקבל ציון על הסכמה בקטגוריות שונות. זאת בניגוד לשיטת "הרוב קובע" בה כל מילה תקבל ציון אחד בקטגורית הרוב.

3.3.2 סטטיסטיקת קאפה

סטטיסטיקת קאפה לוקחת בחשבון את ההסתברות להסכמה מקרית בתוך הקבוצה. שכן אם קבוצת מתייגים תתייג טקסט באופן אקראי, נצפה לרמה מסוימת של הסכמה. בהנחה שכל המתייגים בלתי תלויים וברמה שווה (כלומר אף אחד מהם לא נחשב למומחה), מקדם קאפה (the kappa coefficient) יחושב ע"י [20]:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

כאשר: $P(A)$ - רמת ההסכמה בקבוצה

$P(E)$ - ההסתברות להסכמה מקרית בתוך הקבוצה.

ככל שמספר הקטגוריות וגודל הקבוצה עולים ההשפעה של ההסכמה המקרית יורדת.

נשים לב שכאשר ההסכמה בקבוצה מלאה (כלומר שווה ל-1) מקדם קאפה שווה ל-1. כאשר ההסכמה בקבוצה שווה להסכמה המקרית מקדם קאפה שווה ל-0.

בניתוח התוצאות של הניסויים שביצענו שילבנו את מקדם קאפה בתוצאות מבחן ה"סוג". שילוב מדד קאפה עם מבחן הגבולות מורכב הרבה יותר שכן ההסתברות המקרית תלויה באורך הביטויים השונים ולכן אינה קבועה.

לקבלת תוצאות מבחן ה"סוג" חושבה תחילה ההסכמה החלקית כפי שתואר לעיל. לאחר מכן חושבה התוחלת של ההסכמה המקרית עבור גודל הקבוצה ואוסף הקטגוריות שהוגדר לניסוי. ציון ההסכמה הסופי חושב ע"פ הנוסחה למציאת מקדם קאפה.

3.4 מהלך הניסוי

בניסוי השתתפו שישה מתייגים. הטקסטים חולקו ביניהם באופן כזה שכל טקסט תויג ע"י שלושה אנשים, אולם לא הייתה חלוקה לקבוצות קבועות. הטקסטים נלקחו מתוך כתבות של אתרי אינטרנט: "הארץ" (http://www.haaretz.co.il), "NRG מעריב" (http://www.nrg.co.il) וערוץ-7 (http://www.a7.org). נלקחו כתבות בנושאים שונים כגון חדשות, כלכלה, אופנה, רכילות וכו'.

התיוג נעשה באמצעות כלי עזר לתיוג ידני – **wordfreak** [19]. wordfreak הוא כלי המאפשר מתן תג לכל מילה בטקסט ע"פ קבוצת תגים מוגדרת מראש. תוצאות התיוג נשמרות בקובץ xml. לשם הניסוי (ומשימת תיוג הקורפוס בכלל) הותאם wordfreak למשימת תיוג שמות פרטיים ע"י הזנת אוסף התגים וכן בוצע שלב של חלוקת הטקסט למילים (parsing, על בסיס זיהוי ביטויים רגולריים). כלי זה הפך את משימת התיוג הידני למשימה פשוטה ומהירה יחסית.

בשלב הראשון של הניסוי ניתנו למתייגים הנחיות מינימליות לתיוג. הנחיות אלה כללו את סוגי התגים ומספר דוגמאות מצומצם. כל משתתף קיבל כ-10 טקסטים, ובסה"כ תויגו 25 טקסטים (כ-350 ביטויים בסך הכל). מטרת שלב זה הייתה לבדוק שההנחיות אינן מוסיפות בלבול למשימה וכן לבדוק האם ההנחיות המפורטות אכן תורמות להבנת המשימה באופן חד משמעי (כלומר ציפינו שרמת ההסכמה תעלה בעקבות קריאת ההנחיות) .

בשלב השני ניתנו למתייגים הנחיות מפורטות וקבוצה חדשה של טקסטים לתיוג. כל משתתף תייג כ-30 טקסטים ובסה"כ תויגו כ- 70 טקסטים (כ-1000 ביטויים).

לאחר השלב השני של הניסוי בוצעה הערכה של תוצאות שני השלבים וחושבה מידת ההסכמה ע"פ המבחנים שתוארו לעיל. כמו כן חושבה רמת הבלבול בין כל שתי קטגוריות. התוצאות אפשרו לנו לאתר מקרים בעיתיים של השפה והנחיות לא ברורות. במקומות בהם רמת הבלבול הייתה גבוהה נוסחו כללים חדשים או הורחבו כללים קיימים. כתוצאה מתהליך זה נוסחה הגרסה הסופית והמורחבת של ההנחיות (ראה נספח 1).

בשלב השלישי ניתנו לקבוצה ההנחיות הסופיות וקבוצה חדשה של טקסטים לתיוג. כל משתתף תייג כ-20 טקסטים ובסך הכל תויגו כ- 60 טקסטים (כ-800 ביטויים).

3.5 תוצאות ומסקנות

3.5.1 מבחן t

כדי להעריך עד כמה השינוי בהסכמה בין השלבים השונים של הניסוי משמעותי השתמשנו במבחן t (t-test). מבחן t הוא מבחן סטטיסטי אשר מאפשר לקבוע האם ההבדל בין ממוצעים הוא משמעותי או מקרי [21]. ניתן להשתמש במבחן t אפילו עבור מדגמים קטנים.

מבחן t מחושב ע"פ הנוסחא :

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

כאשר: $\overline{X_i}$ - ממוצע המדגם ה-i

s_i^2 - השונות של המדגם ה-i

n_i - גודל המדגם ה-i

לאחר שחושב ערך מבחן ה-t יש לבדוק בטבלת ערכי מבחן ה-t אם היחס שקיבלנו גבוה מספיק כדי לומר שההבדל בין המדגמים משמעותי. בכדי לבדוק זאת יש לקבוע את רמת הסיכון (risk level, נקראת גם alpha level) . נהוג להשתמש ברמת סיכון 0.05. המשמעות של רמת סיכון זו היא שהקביעה נכונה ב-95%. בנוסף יש לקבוע את רמת החופש של המדגם. אנחנו השתמשנו ברמת חופש ∞ שכן במרבית המקרים מספר המילים/הביטויים שתויגו היה גבוה. עבור רמת סיכון 0.05 ודרגת חופש ∞ נדרוש שערך מבחן ה-t יהיה גבוה מ-1.96 על מנת שנוכל להגיד שהשינוי בהסכמה משמעותי. בטבלאות להלן מוצג ההפרש בין ערך מבחן ה-t לערך סף זה (כלומר כאשר ההפרש גדול מ-0 נוכל לומר שקיים שינוי משמעותי בין המדגמים).

3.5.2 תוצאות שני השלבים הראשונים של הניסוי

TEXT				
Minimal Guidelines		Guidelines Version 1		t-test
% Agreement	num of exp	% Agreement	num of exp	
79	336	87	972	**0.02**

טבלה 3-1: תוצאות מבחן TEST של שלבים 1+2 בניסוי ההסכמה

	TYPE				
	Minimal Guidelines		Guidelines Version 1		t-test
	% Agreement	num of words	% Agreement	num of words	
PERS	97.4	104	99.3	371	-0.29
LOC	86.7	153	90.6	333	**0.01**
ORG	80.9	133	83.4	443	-0.98
DATE	80.3	12	85.1	82	-1.28
MONEY		0	100	23	
TIME	42.8	13	88.7	27	**3.64**
PERCENT		0	98.2	37	
MISC EVENT	66.2	4	71.4	39	-0.95
MISC MODEL	37.2	7	63	32	**1.75**
MISC AFF	76.6	42	75.8	105	-2.12
MISC ENT	32.3	2	45.9	10	**0.5**

טבלה 3-2: תוצאות מבחן TYPE של שלבים 1+2 בניסוי ההסכמה

	TEXT & TYPE				
	Minimal Guidelines		Guidelines Version 1		t-test
	% Agreement	num of exp	% Agreement	num of exp	
PERS	91.2	76	98	220	**0.79**
LOC	76.2	119	86.1	263	**1.34**
ORG	75.2	86	83.3	255	**0.37**
DATE	63	9	72.1	61	-0.98
MONEY	0	0	100	7	
TIME	37.5	8	75	16	**2.58**
PERCENT	0	0	95	20	
MISC EVENT	66.7	2	70.2	19	-1.47
MISC MODEL	33.3	7	65.1	21	**3.46**
MISC AFF	72.9	43	73.5	108	-1.84
MISC ENT	33.3	1	42.9	7	**0.02**

טבלה 3-3: תוצאות מבחן TEXT&TYPE של שלבים 1+2 בניסוי ההסכמה

הטבלה הבאה מציגה את הבלבול בין הקטגוריות השונות. הבלבול חושב ע"פ היחס בין מספר המילים אשר הייתה עליהן מחלוקת בין שני תגים למספר הכולל של מילים אשר קיבלו את התג (באחוזים).

	PERS	LOC	ORG	DATE	MISC EVENT	MISC MODEL	MISC AFF	MISC ENT	OTHER
PERS	0	0	0	0	0	0.3	0	0	1
LOC	0	0	6.6	0	0	0.9	1.2	0	7.2
ORG	0	5	0	0	0.2	0	2.5	0.5	24.2
DATE	0	0	0	0	2.4	0	0	0	25.6
TIME	0	0	0	0	0	0	0	0	22.2
PERCENT	0	0	0	0	0	0	0	0	5.4
MISC EVENT	0	0	2.6	5.1	0	10.3	0	0	30.8
MISC MODEL	3.1	9.4	0	0	12.5	0	0	12.5	43.8
MISC AFF	0	3.8	10.5	0	0	0	0	0	49.5
MISC ENT	0	0	20	0	0	40	0	0	40

טבלה 3-4: בלבול של שלבים 1+2 בניסוי ההסכמה

תחילה נתבונן בתוצאות התיוג בשלב הראשון של הניסוי. בקטגורית "שם אדם" ישנה הסכמה גבוהה בין המתייגים. ההסכמה הנמוכה ביותר נצפתה בקטגורית הזמן ובקטגוריות ה"שונות" למיניהן. ניתן לראות את הפרשי ההסכמה בין המבחנים השונים באותן קטגוריות. תוצאות המבחן המשולב נמוכות מתוצאות מבחן "הסוג" זאת משום שישנה הסכמה נמוכה יחסית על גבולות הביטויים (הסכמה של 79%). ניתן להסיק מכך שקיימים ביטויים עבורם ישנה הסכמה על תיוג חלק מהביטוי אולם אין הסכמה על תחילתו וסופו.

בהינתן הנחיות מפורטות (הגרסה הראשונה) ההסכמה עלתה בכל הקטגוריות בכל המבחנים. ע"י בחינת ערכי מבחן t ניתן לומר שנצפה שינוי משמעותי בהסכמה במבחן ה"סוג" עבור הקטגוריות "שם מקום", "זמן", ובשתי קטגוריות של "שונות". כמו כן ניתן לראות שינוי משמעותי במבחן הגבולות (שינוי של 8%). במבחן המשולב ניתן לראות את השינוי הגבוה ביותר עבור רוב הקטגוריות שכן המבחן הזה משלב את השיפור בשני המבחנים הנפרדים.

בשלב השני של הניסוי, ההסכמה הגבוהה ביותר נצפתה, שוב, בקטגוריית "שם אדם" וכן עבור ביטויי אחוזים. ההסכמה בקטגוריית הזמן עלתה במידה רבה (עליה של כ-40% בהסכמה).

אם נסתכל בטבלת הבלבול, אפשר לראות באופן בולט כי הבלבול הגדול ביותר הוא בין התיוג "לא שם פרטי" (OTHER) לתיוג שם פרטי כלשהו. ישנם שמות עצם רבים אשר מופיעים לפני שם פרטי ואשר אין הסכמה בין המתייגים על השאלה האם שמות עצם אלה הם חלק מביטוי השם. בלבול זה גורם לחוסר הסכמה על גבולות ולהסכמה חלקית בלבד על תיוג הביטוי כולו.

דוגמאות לשמות עצם אשר לא ברור אם לתייגם כחלק מביטוי השם:

- "**קיבוץ** עין גדי"

- "**יום** שלישי"
- "**שעה** 8:00"
- "**כביש** ירושלים תל-אביב"
- "**מפלגת** הליכוד"
- "**צומת** גילת"

בקטגוריות ה"שונות" ההסכמה עלתה, אך ממשיכה להיות נמוכה. ההסכמה בקטגוריות אלו נעה בין 45% ל-75%. בקטגוריות אלו הנטייה הייתה לא לתייג את הביטויים כשם פרטי כלל (כלומר לתת לביטוי כולו את התיוג "אחר"). הבעייתיות ב"שונות" היא שהגדרתם כוללנית מידי. למשל במשפטים:

- "**מוסלמים** הפגינו בהר הבית"
- "יורש העצר **האוסטרי**"

המילים "האוסטרי" ו-"מוסלמים" יקבלו שתיהן את התג "MISC AFF" כלומר תיוג של השתייכות לארגון או לקבוצה. זאת למרות שהמילה "מוסלמים" היא שם עצם והמילה "אוסטרי" היא בכלל תואר השם. לכן במקרים של ספק הנטייה הייתה לא לתייג את הביטוי כלל. בעיה נוספת נתגלתה כאשר במהלך התיוג הופיעו ביטויים שלא הוגדר כלל כיצד לתייגם כגון שמות שירים, ספרים וחגים. במקרים אלו כל מתייג תייג לפי שיקול דעתו או לא תייג כלל. מיקרים אלו הורידו את רמת ההסכמה עבור קטגוריות רבות.

לאחר הסקת המסקנות מהשלב השני של הניסוי נוסחו ההנחיות שוב עם כללים חדשים עבור הנקודות הבעייתיות.

3.5.3 תוצאות השלב השלישי של הניסוי

TEXT			
Guidelines Version 2		t-test vs. minimal	t-test vs. version 1
% Agreement	num of exp		
96.9	815	**3.44**	**0.31**

טבלה 3-5: תוצאות מבחן TEXT של שלב 3 בניסוי ההסכמה

	TYPE			
	Guidelines Version 2		t-test vs. minimal	t-test vs. version 1
	% Agreement	num of words		
PERS	99.6	361	**0.03**	-1.19
LOC	95.4	256	**2.55**	**1.06**
ORG	95	312	**4.01**	**6.1**
DATE	100	69	**1.1**	**3.4**
MONEY	100	11		
TIME	100	3	**6.27**	**0.63**

PERCENT	100	23		-0.53
MISC EVENT	45.9	5	-4.44	-4.6
MISC MODEL	94.1	23	**7.74**	**3.15**
MISC AFF	95.5	112	**1.9**	**4.19**
MISC ENT	100	9	**66.83**	**7.97**

טבלה 3-6: תוצאות מבחן TYPE של שלב 3 בניסוי ההסכמה

	TEXT & TYPE			
	Guidelines Version 2		t-test vs. minimal	t-test vs. version 1
	% Agreement	num of exp		
PERS	99.4	235	**1.46**	-0.25
LOC	96.6	213	**5.41**	**3.85**
ORG	95	175	**4.1**	**4.1**
DATE	100	39	**2.3**	**5.81**
MONEY	100	4		
TIME	100	2	**13.04**	**1.54**
PERCENT	100	12		-0.59
MISC EVENT	50	2	-2.96	-3.07
MISC MODEL	94.4	18	**13.16**	**2.17**
MISC AFF	92.9	112	**2.35**	**4.86**
MISC ENT	100	4		**9.83**

טבלה 3-7: תוצאות מבחן TEXT&TYPE של שלב 3 בניסוי ההסכמה

התוצאות של השלב האחרון של הניסוי מצביעות על רמת הסכמה גבוהה בכל הקטגוריות. בקטגוריה MISC EVENT (שונות – שמות אירועים) נצפתה רמת הסכמה נמוכה, אולם כמות העדויות הקטנה בניסוי השני (שני ביטויים בלבד) אינה מאפשרת קביעה גורפת.
בקטגוריות תאריך, זמן, אחוז, כסף ו- MISC ENT ישנה הסכמה של 100%. בשאר הקטגוריות (פרט ל-MISC EVENT) רמת ההסכמה היא מעל 90%.
ע"פ תוצאות מבחן ה-t ניתן לראות שינוי משמעותי ברוב הקטגוריות מול תוצאות שני השלבים הקודמים.
הגורם העיקרי לחוסר ההסכמה שנותר הוא, שוב, שמות העצם. בתיוג שמות מקומות ושמות ארגונים ההנחיות לגבי שמות עצם מורות לתייג שמות עצם כחלק מביטוי השם אם שם העצם הוא חלק מהשם הרשמי או צורת דיבור נפוצה. כפי שניתן לראות בתוצאות הניסוי האחרון, ישנן פעמים בהן הקביעה הזאת אינה חד משמעית וכל מתייג קיבל החלטה שונה. למרות זאת ניתן לומר שרמת האי – הסכמה ירדה באופן משמעותי עם מתן ההנחיות ולגבי רב המקרים ההנחיות הינן חד משמעיות.

4 הקורפוס

בכדי לבנות מערכת סטטיסטית לומדת יש צורך בקורפוס אימון מתויג. הקורפוס בו השתמשנו מכיל כתבות מאתרי האינטרנט "הארץ" (http://www.haaretz.co.il), מעריב "NRG" (http://www.nrg.co.il) ואתר האינטרנט של "ערוץ-7" (http://www.a7.org). נלקחו כתבות בנושאים שונים כגון חדשות, אופנה, ורכילות. בנוסף נבחר אוסף מיוחד של כתבות משלושת האתרים העוסק בענייני כלכלה. כל הכתבות הן בפורמט UTF-8 . הכתבות מאתר האינטרנט של "ערוץ-7" בעלות אופי שונה מהכתבות שנלקחו מ"הארץ" ו"מעריב". אופי הניסוח שונה וכן אוסף השמות הפרטיים הנפוצים שונה.

הקורפוס כולו מכיל כ-57,000 מילים וכ-4700 ביטויי שם פרטי.

חלוקת הקורפוס לפי מקורות:

המקור	מס' מילים
הארץ	כ- 25,000
מעריב	כ- 5,000
ערוץ-7	כ- 22,000
כלכלה	כ- 5,000

טבלה 4-1: חלוקת הקורפוס למקורות

התיוג נעשה באמצעות כלי עזר לתיוג ידני – **wordfreak** [19] כפי שתואר בפרק הסכמה. התיוג התבצע על בסיס ההנחיות לתיוג כפי שנתקבלו לאחר שהושגה הסכמה מקסימלית בניסויי ההסכמה (ראה נספח 1: הנחיות לתיוג).

חלוקת הקורפוס לביטויים:

התג	מס' ביטויים
אדם	כ- 1400
מקום	כ- 1250
ארגון	כ- 1150
תאריך	כ- 400

זמן	כ- 100
כסף	כ- 200
אחוזים	כ- 200

טבלה 4-2: חלוקת הקורפוס לביטויים

אורך ביטוי ממוצע:

התג	אורך ממוצע (מס' מילים)
אדם	1.557
מקום	1.314
ארגון	1.776
תאריך	1.69
זמן	1.89
כסף	2.649
אחוזים	1.243

טבלה 4-3: אורך ביטוי ממוצע בקורפוס

הקורפוס שהתקבל הינו קורפוס צנוע בהשוואה לכאלה ששימשו מחקרים דומים בתחום. לדוגמא, **Borthwick** [5] השתמש בקורפוס של 300,000 אלף מילים ובמשימת CoNLL2003 [23] השתמשו בקורפוס של 250,000 מילים. עבודת התיוג הינה עבודה ארוכה ומייגעת ולכן הקורפוס שעבדנו איתו קטן יחסית. למרות זאת, כפי שנראה בהמשך גודל זה מספיק על מנת לאסוף עדויות עבור המאפיינים השונים ואימון המערכות.

עבור הניסויים השונים חולק הקורפוס לקבוצת אימון וקבוצת בדיקה. התבצעו שלוש חלוקות אקראיות של הקורפוס כאשר בכל חלוקה גודל קבוצת האימון היה 75% מהקורפוס כולו וגודל קבוצת הבדיקה 25%. כלומר בכל קבוצת בדיקה היו כ-1100 ביטויי שם פרטי ובכל קבוצת אימון כ-1600 ביטויים. כל התוצאות המוצגות בהמשך הושגו ע"י ניסויים בשלוש קבוצות אלה.

5 קו ההתחלה

בכדי להגדיר את קו ההתחלה עבור משימת זיהוי השמות הפרטיים נבנתה מערכת מבוססת קורפוס אימון וביטויים רגולריים.

בשלב הראשון ישנו מעבר על קורפוס האימון המתויג ובניית לקסיקון עבור כל ביטויי השמות הפרטיים המופיעים בו לפחות פעם אחת. הלקסיקון מחולק לפי קטגוריות ומחזיק ביטויי שמות פרטיים באורכים שונים.

כדי לזהות ביטויי כסף, אחוזים, זמן ותאריך הוגדרו ביטויים רגולריים. הוגדרו מילונים לשימוש הביטויים הרגולריים: מילון מספרים – ביטויי מספר במילים, סוגי מטבעות שונים (שקל, דולר, יורו), מילון ימי השבוע, ומילון חודשים עבריים ולועזיים. הביטויים מזוהים גם בצמוד לתחיליות שונות. הביטויים הם:

- ביטוי כסף: ביטוי כסף הוא ביטוי של מטבע או ביטוי מספרי (ספרות או מספר במילים) הצמוד לביטוי מטבע. דוגמאות לביטויים שזוהו: "שקל אחד", "400 מיליון דולר"," יורו".
- ביטוי אחוז: זוהה ביטוי המכיל את המילה "אחוז" על צורותיה השונות או הסימן % בצמוד לביטוי מספרי (ספרה או במילים). לדוגמא:" 87.5%", "כשלושה אחוזים" , "אחוז אחד".
- ביטוי תאריך: יום בשבוע, מספר בין 1-31 בצמוד לשם חודש, אות עברית צמודה לשם חודש עברי או ביטוי המורכב ממספר/אות, שם חודש וביטוי שנה (מספר 4 ספרתי או שנה עברית). לדוגמא: "יום ראשון" , "יום ג'", "4 ביוני 2005" , "אחד באפריל".
- ביטוי זמן: שני מספרים המופרדים ביניהם בנקודותיים או נקודה. הראשון בין 0-59 השני בין 0-24. המילה "שעה" יכולה להופיע לפני המספר. לדוגמא: "השעה 16:50".

בהינתן טקסט חדש לתיוג, עבור כל מילה מתבצעת בדיקה של ההתאמה שלה ושל המילים סביבה לביטויים הרגולריים. אם רצף המילים מתאים לביטוי לאחד הביטויים המילה מקבלת את התג המתאים. אם לא נמצאה התאמה לביטוי רגולרי מתבצע חיפוש שלה עם המילים סביבה (חלון של ± 2) בלקסיקון. אם הביטוי נמצא בלקסיקון המילה מקבלת את התג המתאים. אחרת, המילה מקבלת תיוג "אחר".

תוצאות קו ההתחלה:

	% Precision	% Recall	% F-measure
TOTAL	71.54	48.46	57.78
PERS	59.03	23.91	34.03
LOC	93.22	52.49	67.16
ORG	61.22	44.5	51.54
DATE	46.66	77.59	58.28
MONEY	81.09	83.93	82.49
TIME	77.27	45.83	57.54
PERCENT	76.11	91.67	83.17

טבלה 5-1: קו ההתחלה

אחוז השגיאה של קו ההתחלה גבוה ב-2% מאחוזי השגיאה של קו ההתחלה שהושג עבור השפה אנגלית ב-CoNLL2003 באמצעים דומים.

הביטויים הרגולריים מצליחים לזהות ביטויי שמות ברמה סבירה אך לא מזהים את כולם. כמו כן ניתן לראות שביטויי מקומות הם ביטויים החוזרים על עצמן ברמה גבוהה יחסית ומכאן התוצאות הגבוהות של קו ההתחלה (אחוז שגיאה של כ-34%).

6 זיהוי שמות פרטיים באמצעות מודל אנטרופיה מקסימלית

השיטה אשר הציגה את התוצאות הטובות ביותר באנגלית עד כה הינה שיטת האנטרופיה המקסימלית. שיטה זו מתמודדת היטב עם מספר גבוה של מאפיינים וגם עם כמות דלילה של מידע (קורפוס קטן). קיימים מספר כלי מדף לבניית מודל אנטרופיה מקסימלית, במחקר זה נעשה שימוש בחבילת MaxEnt . העבודה העיקרית בבניית מודל אנטרופיה מקסימלית היא עבודת בחירת המאפיינים. זוהי עבודת ניסוי וטעייה אשר בסופה מתקבל אוסף המאפיינים המוצלח ביותר עבור המשימה. בנוסף יש לבחון את השפעת גודל וסוג הקורפוס על המערכת המתקבלת.

6.1 חבילת MaxEnt

הפרויקט OpenNLP [16] מרכז כלים לעבודה בתחום עיבוד שפה טבעית. במסגרתו נכתבה החבילה MaxEnt [17]. 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. בהמשך נתאר בקצרה את ארכיטקטורת המערכת ואופן השימוש בה.

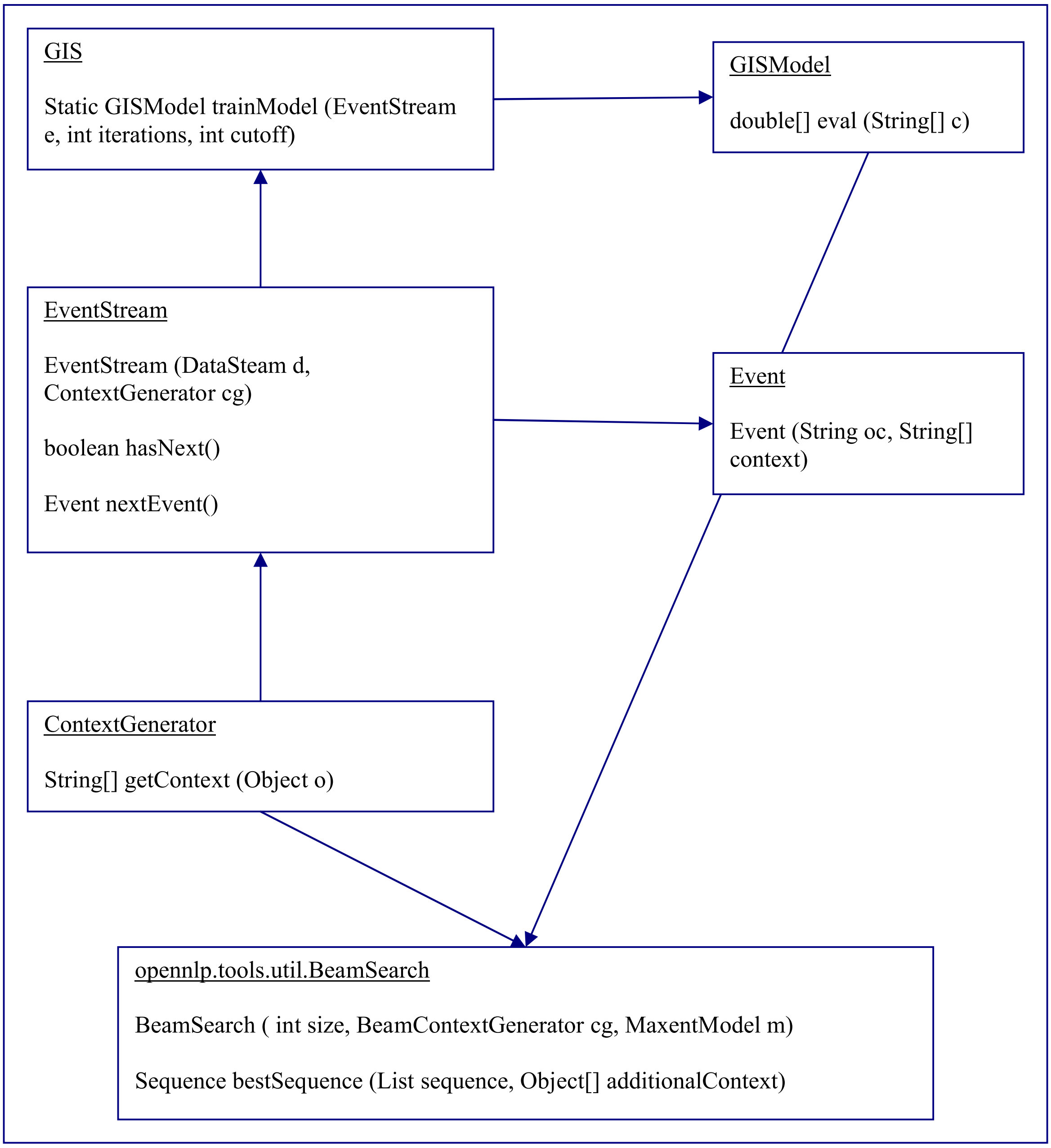
בגרעין המערכת נמצאת המחלקה GIS . מחלקה זו בונה מודל (GISModel) מרצף של אירועים (Event). רצף האירועים מתקבל מהמחלקה EventStream. אירוע הוא עצם המייצג טוקן בקורפוס האימון. הוא מורכב מהתג אותה קיבלה הטוקן ומההקשר שלו. את ההקשר המחלקה EventStream מייצרת בעזרת המחלקה ContextGenerator. המחלקה ContextGenerator מכילה שיטה getContext (Object o) המקבלת מערך של טוקנים ואינדקס ומחזירה את ההקשר של טוקן המבוקש. הקשר של אירוע הוא מערך של מחרוזות המורכב מהמאפיינים הפעילים באותו אירוע. מאפיין יכול לקבל ערך השייך לקבוצה סגורה ובדידה של ערכים (לא בהכרח ערכים בינאריים). לאימון המודל מוגדר ערך סף,זהו המספר המינימאלי של עדויות הדרוש להופעת המאפיין באימון על מנת שיהיה חלק מהמודל הסופי. בסופו של דבר, המערכת הופכת כל מאפיין לקבוצה של מאפיינים בינאריים.

בכדי לבנות מערכת אנטרופיה מקסימלית על המשתמש לספק גרסאות ל-EventStream ול-ContextGenerator. מחלקות אלו הופכות את קורפוס האימון לרצף אירועים ומאפשרת אימון המודל על סמך הקורפוס וקבוצת המאפיינים הרצויה. מחלקת ה-GIS בונה את המודל ההסתברותי על פי אלגוריתם GIS ומייצרת GISModel.

השיטה העיקרית במחלקה GISModel היא השיטה eval(String[] c) אשר מקבלת מערך המייצג הקשר של טוקן מסוים ומחזירה מערך של הסתברויות עבור כל תג אפשרי. למחלקה GISModel יש שיטות שמירה וטעינה מקובץ.

המחלקה opennlp.tools.util.BeamSearch משמשת למציאת רצף התגים הסביר ביותר לשם תיוג טקסט חדש המחולק לטוקנים. לשם כך יש לספק לה את אותו ContextGenerator אשר שימש לאימון. השיטה bestSequence מבצעת חיפוש על בסיס האלגוריתם beam search (יפורט בהמשך) ומחזירה את הרצף הסביר ביותר ואת רצף ההסתברויות שיצרו אותו. ניתן להגדיר ל- BeamSearch מהו רצף חוקי כדי למנוע שגיאות בחוקיות רצף התגים.

ארכיטקטורת MaxEnt מתוארת בתרשים הבא:



איור 6-1: ארכיטקטורת MaxEnt

בחבילה opennlp.tools ניתן למצוא מערכות בסיסיות למשימות שונות בעיבוד השפה האנגלית המשתמשות במודל האנטרופיה המקסימלית. ניתן למצוא מערכת לזיהוי שמות פרטיים, זיהוי משפטים, chunker, מתייג חלקי דיבור ועוד.

השימוש בחבילה הוא נח ופשוט. החבילה מאפשרת גמישות ושימוש בגורמים חיצוניים שונים. לדוגמא ניתן לשלב מתייג חלקי דיבור המספק תחבירי מידע ל-ContextGenerator . המשימה היחידה הנותרת היא בחירת קבוצת המאפיינים.

6.2 אלגוריתם החיפוש - Beam Search

אלגוריתם החיפוש אשר שימש למציאת רצף התגים הסביר ביותר עבור משפט נתון הוא אלגוריתם Beam Search כפי שמתואר ע"י Ratnaparkhi ב-[25].

האלגוריתם משתמש בהסתברות המותנית $p(t \mid h)$ כפי שמחושבת ע"י מודל האנטרופיה המקסימלית.

זוהי ההסתברות למתן תיוג מסוים בהינתן ההיסטוריה של המילה (או במילים אחרות, ההקשר שלה).

בהינתן משפט $\{w_1,....,w_n\}$ ההסתברות המותנית לרצף התגים $\{t_1,....,t_n\}$ מחושב ע"י:

$$P(t_1...t_n \mid w_1...w_n) = \prod_{i=1}^{n} p(t \mid h)$$

האיטרציה ה-i של האלגוריתם מתבוננת במילה ה-i ומחשבת הסתברויות מותנית עבור רצפי תגים המסתיימים במילה ה-i . בכל איטרציה נבחרים K (מספר קבוע – מוגדר ע"י המשתמש) הרצפים הסבירים ביותר והם ממשיכים לאיטרציה הבאה.

נגדיר $W = \{w_1,...,w_n\}$ להיות רצף המילים , s_{ij} להיות רצף התגים בעל ההסתברות ה-j בגודלה עד המקום ה-i.

האלגוריתם:

1. ייצר תגים אפשריים עבור w_1 , בחר K תגים בעלי ההסתברות הגבוהה ביותר וקבע את s_{1j} $1 \le j \le K$ בהתאם.
2. עבור $i = 2$ עד n בצע:
   א. עבור $j = 1$ עד K בצע:
      i. ייצר תגים עבור w_i בהינתן הרצף $s_{(i-1)j}$ וצור רצף חדש ע"י שרשור כל תג לרצף הקיים.
   ב. מצא את K הרצפים בעלי ההסתברות הגבוהה ביותר מבין הרצפים שהתקבלו בשלב א. קבע את s_{ij} $1 \le j \le K$ בהתאם.
3. החזר את הרצף בעל ההסתברות הגבוהה ביותר s_{n1}

בשלב ייצור התגים יש שימוש בכל התגים החוקיים בהינתן הרצף הקודם. במידה ולא הוגדרו כללי חוקיות לרצף, כל התגים האפשריים נלקחים בחשבון. למשל במקרה של המערכת לזיהוי שמות פרטיים מתן תג B_X עבור קטגוריה מסוימת X אינו חוקי אחרי התיוג "אחר".

6.3 המאפיינים

במהלך המחקר נבדקה השפעת מאפיינים רבים על ביצועיה של המערכת. חלק מהמאפיינים שנבחנו נוסו בהצלחה בשפות אחרות וחלק מהמאפיינים ייחודיים לעברית. קבוצת המאפיינים המוצגת בהמשך היא זו שהביאה לתוצאות המוצלחות ביותר. כל המאפיינים מוגדרים עבור כל מילה בנפרד ע"פ תכונותיה, מקומה במשפט והמילים סביבה.

6.3.1 מאפייני מבנה

נקבעו מאפיינים בינאריים המציינים מילה בתחילת מאמר, מילה בתחילת משפט ומילה בסוף משפט. הכתבות בהן השתמשנו לא היו בנויות במבנה מוגדר של כותרת, שם מחבר ותוכן. הן היו בנויות מפסקאות של תוכן בלבד. במקרים שמבנה הכתבה שונה וידוע רצוי לשלב מאפיינים של מילה בכותרת, מילה בכותרת משנה וכו'.

6.3.2 מאפייני לקסיקון

המילה עצמה משמשת כמאפיין. קבוצת הערכים האפשריים של מאפיין זה היא בעצם כל המילים בקורפוס האימון. הוגדרו מאפייני לקסיקון עבור המילה הנוכחית ומילים בחלון של ±2 סביבה. מאפיין זה מאפשר לזהות מילים המופיעות בתדירות גבוהה בתוך שמות פרטיים מסוימים, או לחילופין מילות חיבור ויחס אשר מופיעות בתדירות גבוהה ולרוב אינן משמשות כשמות פרטיים.

6.3.3 תיוגים קודמים

- מאפיין המציין את סוג התג שההופעה האחרונה של המילה באותה הכתבה קיבלה. סוג התג יהיה זהה במקרים של אזכור שם מספר פעמים באותה הכתבה. פעמים רבות המילה לא תקבל את אותו התג לאורך הכתבה כולה. לדוגמא במשפט: "שרה שרה שיר שמח" רק המופע הראשון של המילה "שרה" צריך לקבל תיוג של שם פרטי.
- כדי לאפשר חיפוש רצף התגים הסביר ביותר שולבו מאפיינים המקבלים את התגים שניתנו ע"י המערכת למילה הקודמת לנוכחית ולמילה לפניה (אם קיימות כאלה).

6.3.4 מילון

מילון הינו מאפיין משמעותי. בשל הבעיות שמתעוררות בשפה העברית (ראה פרק 1: רקע ומבוא) לא ניתן להסתמך על מילון בלבד ויש לשלבו עם מאפיינים נוספים על מנת לקבל החלטה.

לצורך בניית המילון נאספו ביטויים ממקורות שונים באינטרנט. המילון מחולק לכרכים ומכיל ביטויים באורכים של 1-3 מילים. בסה"כ יש כ-7000 ביטויים במילון.

פירוט כרכי המילון:

שם הכרך	דוגמאות	מס' ביטויים
תאריכים	אוגוסט, חורף, תשרי, חול המועד פסח, יום ראשון	כ- 150
שמות אדם פרטיים	אמיר, דוד, שירלי, בתיה	כ- 2500
שמות משפחה	אברמוביץ, בוזגלו, בן ארצי	כ-1500
מדינות	ארה"ב, מדגסקר, מקסיקו	כ-500
ערים	אום אל פאחם, תל אביב, ז'נבה	כ-1000
מקומות	כנרת, נגב, קליפורניה	כ-200
מספרים	אחד, מיליון, חצי	כ- 50
ארגונים	חמאס, משרד הבריאות, ש"ס	כ- 400
חברות	דייהטסו, טבע	כ- 600
כסף	דולר, ש"ח	כ- 30

טבלה 6-1: כרכי המילון

ישנן מילים המופיעות במספר מילונים. לדוגמא: המילה "ירדן" מופיעה במילון שמות אדם ובמילון שמות מקומות, המילה "אביב" מופיעה במילון תאריכים ובמילון שמות אדם.

ערך המאפיין הוא שם הכרך בו הביטוי נמצא. אם הביטוי נמצא במספר כרכים, ערך המאפיין הוא איחוד שלהם.

מאפייני מילון הוגדרו על המילה הנוכחית ועל מילים בחלון של ±1 סביבה (מאפיין עבור כל מילה בנפרד).

מאפיין נוסף נקבע ע"י הכניסה המילונית של ביטויים באורך 2-3 אשר מכילים את המילה הנוכחית ומילים סביבה. ערך המאפיין נקבע ע"פ הביטוי הארוך ביותר הנמצא במילון.

6.3.5 ביטויים רגולריים מקומיים

זיהוי ביטויים רגולריים שונים בגודל של טוקן אחד.

סוג הביטויים השונים:

סוג הביטוי	דוגמאות
סיפרה בודדת	1
מספר בעל 2 ספרות (יכול לעזור לזיהוי שעה, תאריך)	14 , 89
מספר בעל ארבע ספרות (שנה לועזית)	1999
שעה	14:50
מספר טבעי	89999
מספר רציונאלי	15.9999
מספר עם אחוז אחוז עם מספר (בעקבות תקלות בפורמט הטקסט)	19% , 15.5% %5
אחוז בודד	%
מספר עם קו נטוי (תאריכים)	14/6/05
טוקן המכיל מספר	כ-30
טוקן המכיל נקודה	איי.בי.אם
טוקן המכיל מקף	באר-שבע
אותיות לועזיות	IBM

טבלה 6-2: סוגי ביטויים רגולריים

ערך המאפיין הוא סוג הביטוי הרגולרי אליו הוא שייך או "לא שייך לביטוי רגולרי".
הוגדרו מאפיינים כאלה עבור המילה הנוכחית וחלון של ±2 מילים סביבה.

6.3.6 ביטויים רגולריים ארוכים

בנוסף לביטויים הרגולריים המוגדרים על כל טוקן בנפרד הוגדרו ביטויים רגולריים המזהים ביטויים באורכים שונים (2-5 טוקנים). מתקבלים שבעה מאפיינים בינאריים המקבלים ערך "1" אם המילה הנוכחית היא מילה בתוך ביטוי רגולרי וערך "0" אחרת.

פירוט הביטויים:

- ביטויי תאריך, זמן, אחוזים וכסף כפי שהוגדרו בקו ההתחלה.
- ביטוי שם מקום: ביטוי המורכב משם עצם מתוך רשימה סגורה של שמות עצם הנפוצים בביטויי מקומות כשלאחריו שם אדם המופיע במילון שמות אדם.

דוגמאות לשמות עצם כאלה: דרום, מחוז, הר, כפר, רחוב, שדרות, שכונת/שכונה, צומת, קיבוץ, כביש.

דוגמא לביטויי מקום שזוהו בדרך זו: "קיבוץ עינת", "הר הרצל", "שדרות רגר".

- ביטוי שם ארגון: שם עצם מתוך רשימה סגורה של שמות עצם הנפוצים בביטויי ארגון כשלאחריו שם אדם או שם מקום המופיע במילון.

דוגמאות לשמות עצם כאלה: מכבי, הפועל, ממשלת, משטרת, עיריית, קרן, אוניברסיטת, מדינת.

דוגמא לביטויי מקום שזוהו בדרך זו: "אוניברסיטת בן גוריון", "עירית תל אביב", "ממשלת ישראל".

- ביטוי במרכאות – רצף של מילים בתוך מרכאות. ביטוי בתוך מרכאות יכול לציין ציטוט או שם פרטי.

6.3.7 שימוש במנתח חלקי דיבר ומפיג עמימות

מנתח חלקי דיבר הוא כלי חיוני ביותר למשימת התיוג שכן הוא עוזר במציאת הניתוח הנכון של המילה. נעשה שימוש במנתח צורני לעברית הכולל הפגת עמימות מורפולוגית של מני אדלר [22] אשר מספק תכונות מורפולוגיות רבות של המילה. בשימוש במנתח הצורני יש לקחת בחשבון את אחוזי השגיאה שלו הנעים בין 5% ל-8%.

בעזרת המנתח הצורני הוגדרו מספר מאפיינים:

- מאפייני חלקי דיבר: חלק הדיבר של המילה הוא התכונה המורפולוגית החשובה ביותר. המנתח מצליח לזהות אחוז מסוים של ביטויי שם פרטי המורכבים ממילה אחת ומתייג אותם כ"שם עצם פרטי" (proper name). שמות רבים מזוהים ע"י המנתח כ"לא ידוע", דבר המקשה על התיוג. שמות פרטיים רבים, בעיקר שמות ארגונים וביטויי זמן ומספר, אינם "שם עצם פרטי". למרות כל אלה תרומת מאפייני חלקי הדיבר למערכת רבה. הוגדרו מאפייני חלקי דיבר עבור המילה הנוכחית וחלון של ±2 מילים סביבה. ערך המאפיין הוא חלק הדיבר אותו קיבלה המילה.
- למה (lemma): למה היא הכניסה המילונית של המילה. לדוגמא הלמה של הפועל "כשנכנסה" היא "נכנס". שימוש בלמה מאפשר נטרול של המילה מהתחיליות שלה ומהטיותיה השונות. ללמה אין משמעות רבה בביטויי שם פרטי, אולם יש לה משמעות עבור המילים הסובבות את הביטוי. הוגדר מאפיין המקבל את ערך הלמה של המילה הנוכחית. קבוצת הערכים עבור מאפיין זה כוללת את הכניסות המילוניות של כל המילים בקורפוס האימון.
- מאפיין המקבל את ערך התחילית (prefix) הראשונה של המילה הנוכחית.

- מאפיין המקבל את ערך הסופית (suffix) של המילה הנוכחית.
- מאפיינים בינאריים המציינים האם המילה הנוכחית או המילה הקודמת לה הן בצורת סמיכות.

6.3.8 עיבוד מוקדם

לפני אימון המערכת התבצע שלב של עיבוד מוקדם על קורפוס האימון. נאספו רשימות שונות על בסיס התיוגים הידועים וחלקי הדיבר של המילים. סוגי הרשימות:

- רשימת מילים נפוצות: נאספה רשימה של מילים אשר הופיעו בלפחות חמש כתבות שונות בקורפוס האימון.
- רשימות ביטויים נפוצים: עבור כל סוג תג נאספה רשימה של ביטויים אשר הופיעו שלוש פעמים או יותר בקורפוס האימון.
- רשימות שמות עצם: עבור כל סוג תג נאספו רשימות של שמות עצם אשר הופיעו בתדירות גבוהה לפני, אחרי או בתוך הביטוי.

ערכי הסף של הרשימות נקבעו אמפירית.

הוגדר מאפיין בינארי המציין האם המילה הנוכחית נמצאת ברשימת המילים הנפוצות.

אם המילה היא חלק מביטוי נפוץ, הוגדר מאפיין המקבל את ערך סוג הרשימה אליה הוא שייך.

הוגדרו מאפיינים הבודקים האם המילה הנוכחית והמילים סביב לה נמצאות ברשימות שמות העצם. מאפיינים אלה מקבלים את ערך סוג הרשימה.

6.3.9 מכפלות מאפיינים

על מנת להדגיש את הקשר בין מאפיינים מסוימים, הוגדרו מאפיינים המקבלים ערך מכפלת מאפיינים אחרים.

דוגמאות למכפלות מאפיינים:

- סמיכות המילה הקודמת עם התג שהיא קיבלה
- המילה הנוכחית עם חלק הדיבר שלה
- חלק הדיבר של המילה הנוכחית עם חלקי הדיבר של המילה הקודמת לה והמילה אחריה
- הלמה של המילה הנוכחית עם חלק הדיבר שלה.

6.4 ניתוח תוצאות

הטבלה הבאה מציגה את התוצאות שהתקבלו ע"י המערכת המשלבת את כל המאפיינים שהוצגו לעיל:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	89.05	67	76.47
PERS	91.6	69.36	78.94
LOC	92.13	74.53	82.4
ORG	82.96	53.71	65.2
DATE	92.04	75.93	83.21
MONEY	89.01	80.36	84.46
TIME	47.14	39.58	43.03
PERCENT	95.24	81.67	87.93

טבלה 6-3: תוצאות מודל האנטרופיה המקסימלית

אחוז השגיאה הכולל של המערכת הוא כ-23.5%, צמצום של כ-20% מאחוז השגיאה של קו ההתחלה. אחוז הדיוק של המערכת גבוה יחסית, ואילו ה-recall נמוך יותר. תוצאה זו מצביעה על כך שכ-33% מהביטויים אינם מזוהים אולם אלו שמזוהים, לרוב מזוהים נכון.
הביטויים המזוהים ברמה הגבוהה ביותר הם ביטויי שם מקום, לביטויים אלו אחוז שגיאה נמוך יחסית גם בקו ההתחלה. רמת השגיאה של ביטויי מקום היא כ-17.5%, של ביטויי שם אדם כ-22% ושל ביטויי ארגון כ-35%. ביטויי ארגון מזוהים ברמה נמוכה יחסית ואחוז השגיאה הוא כ-35%.

סוגי הטעויות מגוונים.
להלן מספר טעויות נפוצות בביטויי ישויות (התיוג מצויין רק עבור השם הפרטי על מנת להקל את הקריאה):

- טעויות בשמות מקוצרים. לדוגמא:

"...הוא חסיד הפתרון הדיפלומטי למשבר <OTHER> המפרץ </OTHER>" (הכוונה היא לקיצור שם המקום "המפרץ הפרסי")

" מצד אחד רוצה <OTHER> האוצר </OTHER> להוריד את שכר המינימום..." (קיצור שם הארגון "משרד האוצר")

" זו אחת הסיבות לכך שמענקי <PERSON> מקארתור</PERSON> מבוקשים כל כך"

(המילה "מקארתור" היא קיצור לשם הארגון "קרן מקארתור" שהוזכר במהלך הכתבה)

- טעויות אשר נעשו על מילים נפוצות אשר בדרך כלל אינן משמשות כשם פרטי:

" ... אמרו אנשי <OTHER> כך </OTHER> בעת מסע ההלוויה"

(כאשר המילה "כך" מופיעה עם מרכאות סביבה היא כן מזוהה כשם ארגון)

- טעויות בשמות לועזיים לא נפוצים. לדוגמא:

" בימי המלחמה הקרה היה <OTHER> טוקי </OTHER> אנטי - קומוניסט נלהב"

"... העמדתו של <OTHER> בוסקה </OTHER> לדין תגרום נזק ציבורי"

- טעויות בביטויים המכילים מקף:

"<LOCATION> כביש באב </LOCATION> <OTHER> אל – וואד </OTHER>"

אחוז גבוה מהביטויים שקיבלו תיוג שגוי תויגו חלקית נכון. ביטויים אלה נחשבים כטעות במבחן המשולב. למשל:

"<PERSON>מרטין</PERSON> <OTHER> לותר </OTHER> <PERSON>קינג</PERSON>"

"<ORGANIZATION> איגוד הפועלים</ORGANIZATION > <OTHER> החקלאיים </OTHER> "

ביטויי תאריך, כסף ואחוז מזוהים ברמה גבוהה יותר מקו ההתחלה. כלומר רמת הזיהוי שלהם טובה יותר משימוש בביטויים רגולריים בלבד.

דוגמא לטעות בביטוי אחוזים:

"<OTHER>מ5 <OTHER/> <PERCENT>ל30 אחוזים <PERCENT/>

הטעויות שנעשו בביטויי כסף הן כאשר המטבע לא מצוין במפורש או שהביטוי תויג חלקית לדוגמא:

"<OTHER>כמה <OTHER/> <MONEY>מיליוני דולרים <MONEY/>".

"<OTHER>...הוא הרווח כ-30 מיליון <OTHER/>"

בנוסף נעשו טעויות בשל בעיות בפורמט הטקסט. לדוגמא הביטוי: "כ30 – אחוזים" לא תויג בשל המיקום השגוי של המקף (הביטוי הנכון הוא: "כ-30 אחוזים").
תוצאות הזיהוי של ביטויי זמן נמוכות מאוד. זאת בשל ביטויים של מספרים בעלי ארבע ספרות (למשל 1950) המזוהים ע"י המערכת כביטויי תאריך שכן ביטויי תאריך נפוצים יותר בקורפוס האימון (פי 4 מביטויי זמן).

בהמשך נבחן את השפעת המאפיינים השונים על המערכת.

6.4.1 חלקי דיבר

כאמור, המאפיינים הדומיננטיים ביותר הם מאפייני חלקי הדיבר.
תוצאות מערכת המתבססת רק על חלקי דיבר (חלון של ±2 סביב המילה הנוכחית):

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	50.32	22.68	31.27
PERS	67.56	39.22	49.63
LOC	57.19	38.81	46.24
ORG	15.83	4.81	7.38
DATE	35.48	17.96	23.85
MONEY	9.13	9.82	9.46
TIME	10	6.25	7.69
PERCENT	16.67	11.54	13.64

טבלה 6-4: תוצאות מערכת אנטרופיה מקסימלית מבוססת חלקי דיבר

אחוז השגיאה של המערכת גבוה מזה של קו ההתחלה פרט לזיהוי שם אדם. התיוגים הנכונים ניתנו הודות לזיהוי שמות מסוימים כ"שם עצם פרטי". זיהוי ביטויי ארגון, כסף וזמן נמוך מאוד.

תרומתם של מאפייני חלקי הדיבר באה לידי ביטוי בעיקר בשילובם עם שאר המאפיינים.

תוצאות מערכת המכילה את כל המאפיינים פרט לחלקי דיבר:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	94.11	48.43	63.95
PERS	91.78	24.56	38.75
LOC	96.42	60.31	74.21
ORG	91.04	43.68	59.04
DATE	92.29	70.17	79.72
MONEY	89.29	83.93	86.53
TIME	7.14	6.25	6.67
PERCENT	98	90	93.83

טבלה 6-5: תוצאות מערכת אנטרופיה מקסימלית ללא חלקי דיבר

ניתן לראות שמאפייני חלקי הדיבר מצמצמים את אחוזי השגיאה בכ-13%. בהיעדר מאפיינים אלה דיוק המערכת עלה אך זוהו כ-20% פחות ביטויים. חלקי הדיבר משמעותיים בזיהוי שמות אדם ומצליחים לצמצם את אחוז השגיאה בזיהוי ביטויים אלה בכ-40%. בזיהוי ביטויי ארגון ומקום צמצום אחוז השגיאה הוא בכ-6%. זיהוי ביטויי הכסף והאחוזים טוב יותר בהיעדר מאפייני חלקי דיבר. ניתן להניח שבהיעדר מאפיינים אלה הביטויים הרגולריים מקבלים משקל רב יותר.

6.4.2 מילון

תוצאות מערכת המתבססת רק על מאפייני מילון (חלון של ± 2 סביב המילה הנוכחית):

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	66.41	31.18	42.44
PERS	34.83	8.29	13.39
LOC	77.6	59.09	67.09
ORG	69.26	49.97	58.05
DATE	73.76	32.41	45.03
MONEY	23.33	16.96	19.64
TIME	0	0	0
PERCENT	33.33	5.51	9.46

טבלה 6-6: תוצאות מערכת אנטרופיה מקסימלית מבוססת מילון

אחוז השגיאה של זיהוי שמות מקומות זהה לזה של קו ההתחלה ואילו אחוז השגיאה של זיהוי שמות ארגונים צומצם בכ-7% מקו ההתחלה. בכל שאר סוגי הביטויים אחוז השגיאה גבוה בהרבה מקו ההתחלה.

ניתן להסיק מכך כי אחוז גבוה של ביטויי שמות מקומות וארגונים בקורפוס נמצא במילון. עם זאת, למרות שמילון שמות אדם הוא הגדול ביותר, התוצאות עבור זיהוי שמות אדם הוא בין הנמוכים. בהיעדר מאפיינים אחרים התיוג של שם אדם אינו מקבל הסתברות גבוהה למרות הימצאות המילה במילון (ניתן לראות זאת ע"י ה-recall הנמוך מאוד).

לא זוהו ביטויי זמן שכן ביטויים אלה לא מופיעים במילונים.

תוצאות מערכת המכילה את כל המאפיינים פרט למאפייני מילון:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	89.87	59.7	71.74
PERS	89.5	58.37	70.66
LOC	92.16	68.48	78.57
ORG	81.81	45.61	58.57
DATE	94.1	78.1	85.36
MONEY	91.67	73.21	81.41
TIME	47.14	39.58	43.03
PERCENT	90	70	78.75

טבלה 6-7: תוצאות מערכת אנטרופיה מקסימלית ללא מילון

ניתן לראות ששימוש במאפייני מילון מצמצם את אחוז השגיאה של המערכת כולה בכ-5%. צמצום השגיאה ניכר בכל סוגי הביטויים ברמה זהה. בהעדר מאפיינים אלה רמת הדיוק של המערכת לא נפגעה, אולם זוהו פחות ביטויים.

נעשה ניסוי נוסף לבדיקת השפעת גודל המילון על תוצאות המערכת. לשם כך התבצעה חלוקה אקראית של המילון לחצי ונעשה שימוש רק בחצי מהמילון. התוצאות הראו על עליה של כ-2% באחוזי השגיאה של המערכת בכל סוגי הביטויים. בניסוי נוסף הוגדל מילון שמות האדם בכ-50%. לא נצפה שינוי באחוזי השגיאה בעקבות הגדלת המילון. מניסויים אלה ניתן להסיק כי שינוי גודל המילון יכול להשפיע על המערכת עד לרמה מסוימת. אפילו בהינתן מילונים גדולים ומקיפים, לא ניתן לבנות מערכת טובה המתבססת על מילונים בלבד.

6.4.3 למה

לכניסה המילונית של המילה משמעות רבה בהבנת פירוש המילה. היא יכולה לסייע במשימות רבות של ניתוח טקסט. נבדוק עד כמה היא משמעותית במשימת תיוג השמות הפרטיים. תוצאות מערכת המבוססת אך ורק על מאפיין הלמה של המילה הנוכחית:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	15.48	0.13	0.26
PERS	0	0	0
LOC	0	0	0
ORG	0	0	0
DATE	83.33	2.58	5.01
MONEY	0	0	0
TIME	0	0	0
PERCENT	0	0	0

טבלה 6-8: תוצאות מערכת אנטרופיה מקסימלית מבוססת למה

ניתן לראות שהתוצאות של מודל זה אפסיות. בהיעדר מאפיינים נוספים התיוג "לא שם פרטי" מקבל את ההסתברות הגבוהה ביותר עבור כל למה בקורפוס. הייחודיות של עברית היא בכך שמקורם של שמות הוא ממילים שונות בעלי תפקידים שונים בשפה ולכן הלמה אינה מספיקה לזיהוי תפקיד המילה במשפט. יש זיהוי מסוים (נמוך מאוד) של ביטויי תאריך, כנראה בשל השימוש החוזר במילה "יום" בתוך ביטויי תאריך.

תוצאות מערכת המכילה את כל המאפיינים פרט ללמה:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	91.27	63.91	75.18
PERS	91.8	66.1	76.86
LOC	93.6	73.02	82.04
ORG	83.76	49.18	61.97
DATE	94.9	75.52	84.11

MONEY	89.01	80.36	84.46
TIME	47.14	39.58	43.03
PERCENT	97.37	80	87.83

טבלה 6-9: תוצאות מערכת אנטרופיה מקסימלית ללא למה

שימוש בלמה מצמצם בכ-1% את אחוזי השגיאה של המערכת כולה. בהיעדר למה הדיוק של המערכת עלה וה-recall ירד. זיהוי ביטויי אדם, מקום וארגון נפגע אולם לא נראה שינוי משמעותי בזיהוי תאריך, זמן, אחוזים וכסף.

הלמה מצליחה לצמצם את אחוז השגיאה אולם אינה משמעותית כמאפייני חלקי הדיבר והמילונים.

6.4.4 עיבוד מוקדם

קו ההתחלה מציג מערכת המשתמשת ברשימות עיבוד מוקדם (רשימות שונות במקצת מאלה המשמשות את מערכת האנטרופיה המקסימלית).

נבחן את תוצאות המערכת בהיעדר המאפיינים המשתמשים ברשימות העיבוד מוקדם:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	88.61	63.65	74.08
PERS	91.85	69.57	79.17
LOC	91.36	74.71	82.2
ORG	78.55	51.26	62.04
DATE	91.17	75.93	82.86
MONEY	85.71	80.36	82.95
TIME	35.71	39.58	37.55
PERCENT	91.3	81.67	86.22

טבלה 6-10: תוצאות מערכת אנטרופיה מקסימלית ללא רשימות עיבוד מוקדם

שימוש ברשימות העיבוד המוקדם מאפשר צמצום השגיאה בכ-2%. בזיהוי שמות ארגונים השגיאה צומצמה בכ-3% כנראה בשל ביטויי ארגון החוזרים על עצמם בקורפוס האימון וזיהוי שמות עצם נפוצים הבאים לפני או בתוך ביטויי ארגון. העיבוד המוקדם אינו עוזר לזיהוי שמות אדם ומקומות. זיהוי שאר הביטויים גם הוא משתפר שוב בשל זיהוי שמות עצם המופיעים בתדירות גבוהה לפני, אחרי או בתוך ביטוי השם.

6.4.5 השפעת וסוג גודל הקורפוס

הקורפוס שהשתמשנו בו נלקח ממקורות שונים בעלי אופי שונה. כמו כן, מכל מקור לקחנו כמות שונה של כתבות. כאמור, לכתבות מ"מעריב" ומ"הארץ" אופי דומה מבחינת אוצר המילים, הניסוח והביטויים

הנפוצים, בעוד לכתבות מ"ערוץ 7" ולכתבות בנושאי כלכלה אופי משלהם. נבחנה השפעה הסוגים השונים של הכתבות על המערכת.

תוצאות מערכת שאומנה ללא הכתבות מ"מעריב":

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	89.49	63.51	74.29
PERS	91.53	68.87	78.6
LOC	91.56	74.73	82.29
ORG	80.63	51.97	63.2
DATE	93.7	73.87	82.61
MONEY	85.71	80.36	82.95
TIME	40.48	39.58	40.02
PERCENT	97.62	83.33	89.91

טבלה 6-11: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "מעריב"

הכתבות מ"מעריב" מהוות כ-10% מהקורפוס כולו. ישנה עליה באחוזי השגיאה הכוללים בכ-2%. נפגעה רמת זיהוי כל הביטויים פרט לביטויי שם אדם ומקום.

תוצאות מערכת שאומנה ללא הכתבות מ"הארץ":

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	87.58	60.5	71.56
PERS	91.3	63.79	75.11
LOC	91.07	74.03	81.67
ORG	76.24	51.06	61.16
DATE	94.12	79.03	85.92
MONEY	92.86	52.68	67.22
TIME	37.14	31.25	33.94
PERCENT	89.58	81.67	85.44

טבלה 6-12: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "הארץ"

כתבות "הארץ" מהוות כ-40% מהקורפוס כולו. בהיעדרן ניתן לראות עליה באחוזי השגיאה של כ-5%. זיהוי שמות אדם, ארגונים וכסף נפגע ברמה משמעותית. במפתיע עלתה רמת הזיהוי של ביטויי תאריך. כנראה שביטויי התאריך בקורפוס "הארץ" מגוונים ואינם בעלי תבנית מוגדרת ולכן מוסיפים אי סדר למערכת.

תוצאות מערכת שאומנה ללא הכתבות מ"ערוץ 7":

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	87.72	59.96	71.23
PERS	90.54	68.87	78.23
LOC	89.96	71.13	79.44
ORG	80.22	47.75	59.87
DATE	88.14	75.72	81.46
MONEY	85.71	80.36	82.95
TIME	37.5	25	30
PERCENT	93.18	81.67	87.05

טבלה 6-13: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס "ערוץ 7"

כתבות "ערוץ 7" גם הן מהוות כ-40% מהקורפוס כולו. נצפתה עליה באחוזי השגיאה של המערכת ברמה דומה לזו שנצפתה בהעדר כתבות "הארץ". עם זאת סוג השגיאות שונה. בהיעדר כתבות "ערוץ 7" כמעט לא נפגע זיהוי שמות אדם. ניתן להסיק מכך שבכתבות אלה אין תוספת משמעותית ללקסיקון שמות האדם במערכת או ששמות האדם המופיעים בהן לא מופיע בתדירות גבוהה בקורפוס כולו. כתבות "ערוץ 7" תורמות יותר לזיהוי שמות מקומות וארגונים מכתבות "הארץ".

תוצאות מערכת שאומנה ללא כתבות הכלכלה:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	88.9	62.8	73.6
PERS	91.45	70.19	79.42
LOC	91.56	74.54	82.18
ORG	78.22	52.14	62.57
DATE	94.73	72.75	82.3
MONEY	93.75	50	65.22
TIME	33.33	33.33	33.33
PERCENT	95	30.38	46.04

טבלה 6-14: תוצאות מערכת אנטרופיה מקסימלית ללא קורפוס כלכלה

כתבות הכלכלה מהוות כ-10% מהקורפוס. כצפוי, תרומתן ניכרת בזיהוי שמות ארגונים וביטויי כסף ואחוזים. בהעדרן נצפתה עליה של כ-3% באחוז השגיאה של זיהוי שמות אחוזים, עליה של 20% באחוזי השגיאה של ביטויי כסף ועליה של 40% בביטויי אחוזים. כתבות אלו כמעט ולא משפיעות על זיהוי שמות אדם ומקום.

6.4.6 סיכום התוצאות

מהניסויים הרבים נובע כי המאפיינים הדומיננטיים ביותר הם מאפייני המילון וחלקי הדיבר. מאפייני המילון משמעותיים בזיהוי שמות מקומות וארגונים, ואילו מאפייני חלקי הדיבר משמעותיים בזיהוי שמות אדם. שאר המאפיינים מצליחים לצמצם את אחוזי השגיאה, אך ברמות נמוכות הרבה יותר (בין 0.5% ל-2% לכל מאפיין). התוצאות הטובות ביותר עבור כל התגים יחד מתקבלות ע"י שילוב המאפיינים כולם. גודל הקורפוס וסוגו משפיעים רבות על המערכת וראינו כי לכל סוג קורפוס תרומה שונה למערכת.

7 זיהוי שמות פרטיים באמצעות מודל מרקוב חבוי (HMM)

העבודות אשר הציגו את התוצאות הגבוהות ביותר באנגלית השתמשו במערכות אשר שילבו מספר מודלים שונים. מערכות שהשתמשו במודל מרקוב עבור השפה האנגלית הציגו תוצאות טובות יחסית. למודל מרקוב תכונות שונות ממודל האנטרופיה המקסימלית באופן ההתמודדות שלו עם כמות גדולה של מאפיינים ודלילות המידע. כמו כן מודל מרקוב משתמש בשיטת חיפוש שונה ממודל האנטרופיה המקסימלית ולוקח בחשבון הסתברויות מעבר ממצב למצב, בעוד מודל האנטרופיה המקסימלית מקבל החלטה מקומית. ניתן להשתמש בתכונות השונות של המודלים על מנת לשפר את ביצועי המערכת. ניסינו לבנות מערכת לזיהוי שמות פרטיים המבוססת על HMM. קבוצת המצבים וא"ב המערכת הוגדרו בצורות שונות.

7.1 המודל הבסיסי

המודל הראשון הינו המודל הבסיסי ביותר. נגדיר מודל מרקוב חבוי בו קבוצות המצבים הם התגים עצמם. עבור כל אחד משבעת התגים השונים של שמות פרטיים נגדיר 2 מצבים: תחילת ביטוי ואמצע ביטוי. נגדיר מצב נוסף עבור התיוג "אחר" וכן 2 מצבים מיוחדים עבור תחילת משפט וסופו (בסה"כ 17 מצבים) . א"ב הסמלים הוא בפשטות המילים עצמן. כלומר, א"ב הסמלים הוא קבוצת כל המילים בקורפוס האימון. קבוצות הסתברויות מעבר בין מצבים והסתברויות הפליטה עבור כל מצב חושבו על בסיס קורפוס האימון (קורפוס זהה לקורפוס ששימש למערכת האנטרופיה המקסימלית) ע"פ הנוסחאות המתוארות בפרק עבודות קודמות-מודלי מרקוב.
בהינתן טקסט חדש נשתמש בחיפוש ויטרבי כדי לקבל את רצף המצבים (כלומר התגים) בעל ההסתברות הגבוהה ביותר.

תוצאות המודל הבסיסי:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	31.79	1.23	2.37
PERS	0	0	0
LOC	0	0	0
ORG	0	0	0
DATE	77.35	16.71	27.48
MONEY	0	0	0
TIME	24.29	20.83	22.43
PERCENT	5.56	1.67	2.57

טבלה 7-1: תוצאות מערכת HMM הבסיסי

התוצאות של מודל זה אפסיות, הרבה מתחת לקו ההתחלה. רוב המילים קיבלו את התיוג "אחר" כלומר, כמעט ולא ניתן תיוג של שם פרטי. ניתן להסביר זאת ע"י כך שבהיעדר מידע נוסף, המצב "אחר" הוא בעל הסתברויות הפליטה הגבוהות ביותר עבור כל המילים בקורפוס. בשפה העברית מילים רבות משמשות הן כשמות פרטיים והן כשמות עצם ("אילן"- שם אדם, "שינוי"- שם ארגון), פעלים ("אסף" – שם אדם) שם תואר ("יפה"). לכן בהיעדר מידע נוסף התיוג הסביר ביותר לכל מילה הוא "אחר". ניתן לראות רמה מסוימת של זיהוי במקרים של תאריך זמן ואחוזים, בקטגוריות אלה יש מילים מסוימות החוזרות על עצמן בהסתברות גבוהה.

7.2 חלקי דיבר כמצבים

המסקנה המתבקשת מהמודל הראשון היא שהמילה עצמה אינה מידע מספק למערכת על מנת למצוא שמות פרטיים בטקסט. יש לבנות מערכת עם מידע מעמיק יותר על תכונות המילה והקשרה במשפט. המידע הבסיסי ביותר בהבנת הקשר של מילה במשפט הוא חלק הדיבר של המילה. נעשה ניסיון לבניית HMM בו א"ב הסמלים הוא מכפלה של המילה עם חלק הדיבר שלה. מודל כזה הניב תוצאות טובות יותר מהמודל הבסיסי, אולם המודל המתואר בהמשך הוא המודל המשלב את חלקי הדיבר בצורה הטובה ביותר.

במודל זה קבוצת המצבים הוגדרה בצורה שונה. כעת קבוצת המצבים תהיה מכפלה של קבוצת התגים (כפי שהוגדרה במודל הקודם) עם קבוצת חלקי הדיבר האפשריים. כלומר, יהיה מצב עבור שם אדם+שם עצם, שם אדם+פועל וכו'. בנוסף נגדיר מצבים עבור תחילת משפט וסופו. בסך הכל מתקבלים 212 מצבים. האינטואיציה להגדרה זו נובעת מהעובדה כי למבנה התחבירי של המשפט יש משמעות רבה במשימת התיוג ועל כן יש להדגיש את מבנה הרצף באמצעות הסתברויות המעבר של ה-HMM.

א"ב הסמלים שוב יהיה המילים עצמן.

לקבלת ניתוח חלקי הדיבר של המילים בטקסט נעשה שימוש במנתח צורני לעברית הכולל הפגת עמימות מורפולוגית של מני אדלר [22].

קבוצות הסתברויות המעבר בין מצבים והסתברויות הפליטה עבור כל מצב חושבו שוב על בסיס קורפוס האימון והמנתח הצורני.

בהינתן טקסט חדש נותח הטקסט תחילה ע"י המנתח הצורני ולאחר מכן התבצע חיפוש ויטרבי המתבסס על חלקי הדיבר הידועים.

תוצאות המודל מוצגות בטבלה הבאה:

TEXT & TYPE			
	% Precision	**% Recall**	**% F-measure**
TOTAL	33.67	26.35	29.56
PERS	49.97	34.58	40.87
LOC	45.01	36.56	40.35
ORG	14.64	7.22	9.67
DATE	28.69	33.53	30.92
MONEY	4.07	25.89	7.03
TIME	8.01	20.83	11.57
PERCENT	1.79	1.67	1.73

טבלה 2-7: תוצאות מערכת HMM – חלקי דיבר כמצבים

התוצאות טובות בהרבה מהמודל הבסיסי, אך עדיין מתחת לקו ההתחלה. התוצאות בזיהוי שם אדם גבוהות מקו ההתחלה אולם שמות ארגונים ומקומות זוהו ברמה נמוכה מקו ההתחלה. הרבה משמות האדם מזוהים ע"י המנתח הצורני כ"שם עצם פרטי" (proper name) דבר המעלה את ההסתברות לזיהוי ע"י מודל ה-HMM. שמות ארגונים ומקומות מזוהים ע"י המנתח הצורני ברמה נמוכה ומכאן ההבדל ברמת הזיהוי. כמו כן לגבי ביטויי אחוזים, כסף, תאריך וזמן תוצאות הזיהוי בעזרת ביטויים רגולריים מוצלחות הרבה יותר.

חלקי הדיבר הם מאפיין חשוב במשימת התיוג כמו שנצפה גם במודל האנטרופיה המקסימלית. השימוש בחלקי דיבר כמצבים במודל טוב יותר משימוש בהם כחלק מהא"ב משום שהמבנה התחבירי של המשפט כולו חשוב מאוד בהבנת תפקיד כל מילה בו. יחסי הגומלין בין חלקי הדיבר באים לידי ביטוי בחישוב הסתברויות המעבר בין המצבים. תכונת הרצף של המשפט והמבנה התחבירי שלו באים לידי ביטוי בצורה טובה יותר כאשר חלקי הדיבר מוגדרים כמצבים ולא רק כסמל הנפלט ממצב מסוים. מצבים המשלבים חלקי דיבר מאפשרים חלוקת המשפט לביטויים ארוכים ממילה אחת ומתן תיוג זהה לכל המילים השייכות לאותו הביטוי שכן סוג התג גם הוא חלק מהגדרת המצב.

7.3 HMM המשלב מאפיינים

נראה כי יש להוסיף מידע על מנת להגיע למודל מוצלח. נעשו ניסיונות עם קבוצות מצבים שונות וא"ב סמלים שונים. המודל המוצג בהמשך הוא המודל אשר הניב את התוצאות הטובות ביותר.

קבוצת המצבים הוגדרה בדומה לקבוצת המצבים במודל הקודם: מכפלת התגים עם חלקי הדיבר. א"ב הסמלים הוגדר באופן שונה. במקום שכל מצב יפלוט מילה, כל מצב כעת פולט מחרוזת המייצגת אוסף מאפיינים. הקו המנחה לבחירה זו היה שלמטרת זיהוי השמות הפרטיים, המילה עצמה פחות חשובה מהמאפיינים שלה. א"ב סמלים כזה יהיה מצומצם יותר. בכך נצליח לשלב יותר מידע במודל וכן להקטין

את פיזור ההסתברויות במודל. כלומר, כאשר נעבוד עם א"ב מצומצם נצפה למצוא יותר עדויות לכל סמל.

א"ב הסמלים הורכב משרשור המאפיינים הבאים עבור כל מילה בקורפוס:

- ביטויים רגולריים עבור ביטויי תאריך, זמן, אחוזים וכסף כפי שהוגדרו בקו ההתחלה. כל מאפיין קיבל ערך בינארי: האם הביטוי בטקסט מתאים לביטוי הרגולרי או לא.
- מאפיין עם ערך בינארי עבור ביטוי רגולרי המזהה ביטוי בתוך מרכאות.
- שימוש במילונים חיצוניים: מילון שמות אדם פרטיים ושמות משפחה, ערים, ארצות ארגונים וחברות. הוגדר מאפיין "מילון" אשר מקבל את הערך של חיפוש המילה הנוכחית או חלון של ± 2 סביבה במילון. אם הביטוי לא נמצא במילון המאפיין קיבל את הערך "אחר".
- מאפיין בינארי עבור ביטוי רגולרי המזהה שם מקום: שם עצם מתוך רשימה סגורה של שמות עצם הנפוצים בביטויי מקומות כשלאחריו שם אדם המופיע במילון שמות אדם (כפי הוגדר עבור מערכת האנטרופיה המקסימלית).
- מאפיין בינארי עבור ביטוי רגולרי המזהה שם ארגון: שם עצם מתוך רשימה סגורה של שמות עצם הנפוצים בביטויי ארגון כשלאחריו שם אדם או שם מקום המופיע במילון (כפי שהוגדר עבור מערכת האנטרופיה המקסימלית).
- נעשה עיבוד מוקדם על קורפוס האימון אשר במהלכו נאספו רשימות שונות: מילים נפוצות, ביטויים נפוצים עבור כל תג בנפרד. עבור כל תג נאספה רשימה של ביטויים המופיעים בתדירות גבוהה לפניו או אחריו (חלון של ± 2). הוגדרו מספר מאפיינים אשר ערכם הוא הערך המוחזר מחיפוש המילה הנוכחית או חלון של ± 2 ממנה ברשימות הנ"ל.

אימון המודל ואלגוריתם החיפוש התבצעו בצורה זהה לשני המודלים הקודמים.

תוצאות:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	80.04	59.41	68.2
PERS	82.41	55.47	66.31
LOC	86.81	71.48	78.4
ORG	73.4	47.3	57.53
DATE	85.58	71.82	78.1
MONEY	87.12	66.96	75.72
TIME	24.42	68.75	36.04
PERCENT	73.17	63.59	68.04

טבלה 7-3: תוצאות מערכת HMM – שילוב מאפיינים

אחוז השגיאה ירד בכ-11% מאחוז בשגיאה בקו ההתחלה. זיהוי שמות אדם טוב בכ-30% מקו ההתחלה, זיהוי שמות מקומות טוב ב-10%, ושמות ארגונים ב-6%. זיהוי ביטויי תאריך טוב מזיהוי באמצעות ביטויים רגולריים בלבד. אולם עבור ביטויי אחוזים, זמן וכסף עדיין הזיהוי היעיל ביותר הוא באמצעות ביטויים רגולריים בלבד כמו במודל של קו ההתחלה.

יש לציין כי נעשה ניסיון להרחיב את קבוצת המאפיינים כפי שנעשה עבור מודל האנטרופיה המקסימלית. תוספת מאפיינים הביאה לעליה באחוזי השגיאה שכן היא גרמה לפיזור גדול יותר של העדויות. מכיוון שבמודל זה יש התייחסות לאיחוד המאפיינים ולא לכל מאפיין בנפרד, תוספת מאפיינים מביאה למצב בו יש מספר עדויות נמוך יותר לכל סמל ומכאן לפיזור גבוה יותר של ההסתברות.

המסקנה העיקרית של תוצאות מודל זה היא החשיבות של המאפיינים על פני המילה עצמה. תהליך הלמידה נעשה באופן יעיל יותר ע"י שימוש בקבוצת מאפיינים אשר לכל אחד מהם מספר קבוע של ערכים אפשריים. נעשה שימוש במאפיינים מקומיים בלבד עבור כל מילה. כלומר, במילה עצמה ובחלון קבוע סביבה על מנת לקבוע את ערך המאפיינים.

כאמור, השימוש במכפלת חלקי דיבר והתגים כמצבים במודל מאפשר חלוקת המשפט לביטויים ומתן תיוג זהה לביטוי כולו. כלומר, הביטוי כולו תויג נכון או לא נכון. רק 15% מהטעויות שנעשו ע"י המערכת (מתן תג שם שגוי או מתן תיוג "אחר" לשם פרטי) על ביטויים הארוכים ממילה אחת נעשו רק על חלק מהביטוי.

החיסרון של המודל הוא שאוסף המאפיינים מוגדר כמחרוזת אחת עבור כל מילה. אין התייחסות לכל מאפיין בנפרד אלא לאיחוד שלהם. הבעיה נובעת מהעובדה שב-HMM הקלאסי כל מצב פולט סמל אחד בלבד בכל פעם.

בעיה נוספת במודל זה בפרט ובשימוש במודל מרקוב בכלל היא ההנחה הבסיסית היא שהסתברויות המעבר והסתברויות הפליטה הן בלתי תלויות ולכן מחושבות בנפרד, הנחה שככל הנראה לא מתקיימת במקרה של זיהוי שמות פרטיים.

נעשו ניסויים עם קורפוס אימון בגדלים שונים. קורפוס האימון המקורי מנה כ-43,000 מילים בכל ניסוי (כ- 3500 ביטויי שם פרטי). ירידה של 10% בגודל קורפוס האימון כמעט ולא הביאה לשינוי בתוצאות. ירידה של 20% בגודל קורפוס האימון הביאה לעליה באחוזי השגיאה בכ-2%. ירידה של 40% בגודל קורפוס האימון הביאה לעליה של כ-3% באחוזי השגיאה. הפגיעה הגדולה ביותר הייתה בתוצאות זיהוי שמות אדם, מקומות וארגונים. הייתה ירידה של כ-5% ברמת הדיוק (precision) של המערכת. אימון המערכת ללא המאמרים בנושא כלכלה (כ-5000 מילים) הביא לעליה בשגיאות שנעשו בזיהוי ביטויי האחוזים בלבד, עליה של כ-20% באחוזי השגיאה.

נעשה ניסיון לבנות מודל מרקוב דומה (קבוצת מצבים וא"ב פליטה דומים) אך מסדר 2. במודל כזה הסתברות המעבר למצב הבא תלויה בשני המצבים הקודמים והמילה הנוכחית. מודל כזה הניב תוצאות נמוכות יותר.

7.4 השוואה למודל האנטרופיה המקסימלית

ה-HMM הסופי לא הציג תוצאות טובות יותר מתוצאות מודל האנטרופיה המקסימלית בשום קטגוריה. אולם מודל זה מצליח לזהות ביטויים שונים מאלה המזוהים ע"י מודל האנטרופיה המקסימלית. ההבדל נובע מהתכונות השונות של כל מערכת.

ב-HMM שולבו פחות מאפיינים, תוספת מאפיינים הביאה לירידה בתוצאות. לכן במודל זה יש דגש על מאפיינים של מילון וביטויים רגולריים. תכונה זו מביאה לזיהוי שונה של ביטויים, לא תמיד הזיהוי הוא נכון.

דוגמא לזיהוי נכון של ה-HMM המתבסס על מילון (ביטוי שלא זוהה ע"י מערכת ה-ME):

" ... שוררת בין דמוקרטים התחושה שיש להם סיכוי ממשי להוציא את בוש מן <I_ORG> **הבית הלבן** <I_ORG/> בעוד שנתיים. "

למרות שהביטוי "הבית הלבן" קיים במילון הארגונים מערכת האנטרופיה המקסימלית לא זיהתה אותו. כנראה שלצירוף של המילה "בית" עם שם תואר עוקב אין הסתברות גבוהה להיות שם פרטי.

דוגמא לזיהוי לא נכון של ה-HMM על פי המילון (זוהה נכון ע"י מערכת ה-ME):

" אגרתי כוחות חדשים ואני חוזר במשנה מרץ <I_ORG> **לעבודה** <I_ORG/> שאני אוהב. "

דוגמא לזיהוי נכון של ה-HMM המתבסס על הביטוי הרגולרי עבור שם מקום (ביטוי שלא זוהה ע"י מערכת ה-ME):

" החל משעות הצהרים התקבצו אלפים סביב ישיבת הרעיון היהודי <I_LOC> **בשכונת שמואל הנביא** <I_LOC/> . "

דוגמא לזיהוי לא נכון של ה-HMM על פי הביטוי הרגולרי עבור שם מקום (זוהה נכון ע"י מערכת ה-ME):

"... הוא מצדו אינו ממליץ לחון את <I_LOC> **הר שפי** <I_LOC/> . "

הבדל משמעותי בין שני המודלים הוא שמודל האנטרופיה המקסימלית מתייחס לכל מאפיין בנפרד. ה-HMM מתייחס רק לאיחוד המאפיינים, אין למודל אפשרות לזהות מאפיין בודד אם הוא מופיע בהקשר של מאפיינים שונים. מודל האנטרופיה המקסימלית מסוגל להתמודד עם קבוצה גדולה של מאפיינים ולתת לכל אחד מהם משקל לפי ההשפעה שלו על מתן כל תג. הגדלת אוסף המאפיינים אינו מביא למצב בו יש מספר עדויות קטן מדי עבור כל מאורע, בניגוד למתרחש ב-HMM.

הבדל נוסף בין המערכות הוא שיטת החיפוש. מערכת ה-HMM משתמש בחיפוש ויטרבי בעוד מערכת האנטרופיה המקסימלית משתמשת בשיטת Beam Search . חיפוש ויטרבי לוקח בחשבון את הסתברויות המעבר בין המצבים, דבר שלא קיים במודל האנטרופיה המקסימלית. מעבר בין תגים במודל האנטרופיה המקסימלית בא לידי ביטוי בצורת מאפיינים ולכן פחות דומיננטי.

היתרון היחיד של ה-HMM הוא ההמשכיות שלו. כאמור, קביעת מצבי ה-HMM כמכפלת חלקי הדיבר והתגים מביאה לכך שאין הרבה טעויות תיוג בתוך הביטוי עצמו והביטוי כולו מתויג ע"י אותו תג. במודל האנטרופיה המקסימלית יש הרבה יותר ביטויים המתויגים חלקית. במודל האנטרופיה המקסימלית כ-35% מהטעויות שנעשו ע"י המערכת על ביטויים הארוכים ממילה אחת נעשו רק על חלק מהביטוי לעומת 15% ב-HMM.

8 שילוב מערכות

כאמור בפרק עבודות קודמות, מערכות אשר שילבו בין מודלים שונים הצליחו להגיע לתוצאות הטובות ביותר עבור השפות אנגלית וגרמנית.
במחקר זה המערכת אשר הציגה את התוצאות הטובות ביותר היא מערכת האנטרופיה המקסימלית. על מנת לשפר את תוצאות הזיהוי נבנתה מערכת המשלבת את שלוש המערכות שהוצגו עד כה: מערכת קו ההתחלה, מערכת ה-HMM ומערכת האנטרופיה המקסימלית. לכל מערכת תכונות ויתרונות משלה:

- מערכת קו ההתחלה מצליחה לזהות ברמה טובה ביטויי תאריך, זמן, כסף ואחוזים בשל הדומיננטיות של השימוש בביטויים רגולריים.
- מערכת ה-HMM מצליחה לזהות ביטויים במלואם וכן בשל תכונותיה השונות מצליחה לזהות ביטויים שלא מזוהים ע"י מערכת האנטרופיה המקסימלית.
- מערכת האנטרופיה המקסימלית מצליחה לשלב מספר רב של מאפיינים ולהגיע לתוצאות טובות. עדיין יש מקום לשיפור בעיקר בביטויי ארגון וזמן. כמו כן התוצאות מציגות אחוזי דיוק גבוהים ו-recall נמוך יותר. יש לשאוף לאיזון בין שני מדדים אלה.

אופן השילוב המוצלח ביותר הוא מתן משקל "הצבעה" שונה לכל מערכת בכל קטגוריה. משקל זה נקבע סטטיסטית על בסיס קורפוס אימון כפי שהוצג ב-[6]. על קורפוס אימון כזה להיות שונה מהקורפוס ששימש את המערכות השונות לאימון. בשל מגבלות הגודל של הקורפוס שלנו לא ניתן היה לבצע אימון על קורפוס שונה. לכן אופן המיזוג נקבע אמפירית. כלומר, נבחרה דרך המיזוג אשר הציגה את התוצאות הטובות ביותר בניסויים שבוצעו.
כל מערכת אומנה בנפרד על אותו קורפוס. בהינתן טקסט חדש, כל מערכת תייגה אותו. לאחר קבלת שלושת התיוגים השונים התבצע מיזוג שלהם. משום שמערכת האנטרופיה המקסימלית הציגה את התוצאות הטובות ביותר נעשה שימוש בתיוגים האחרים רק במקרים של מתן תיוג "לא שם פרטי" על ידה. לכלל זה ישנו יוצא דופן אחד.
להלן חוקי המיזוג:

- אם מערכת האנטרופיה המקסימלית תייגה מילה כתאריך:
    - אם גם ה-HMM תייגה כתאריך המילה מקבלת את התיוג תאריך
    - אם מערכת קו ההתחלה תייגה כביטוי זמן המילה מקבלת תיוג זמן

  (חוק זה התקבל מכיוון שאחוזי השגיאה של מערכת האנטרופיה המקסימלית גבוהים מאוד וקיימים ביטויי זמן רבים המקבלים תיוג תאריך)
- אחרת אם מערכת האנטרופיה המקסימלית נתנה למילה תג שונה מהתג "לא שם פרטי" זהו התג שיינתן למילה.
- אחרת אם ה-HMM נתן למילה תיוג שם אדם, מקום או ארגון זהו התג שהמילה תקבל.

- אחרת אם ה-HMM ומערכת קו ההתחלה נתנו למילה תיוג זהה זהו התג שיינתן למילה.
- אחרת יינתן למילה התיוג "לא שם פרטי".

מכיוון שהתיוגים של מערכת האנטרופיה המקסימלית לא השתנו (פרט לשינויים בביטויי תאריך) לא יתכן שה-recall ירד (שכן לא יתכן שנזהה כעת פחות ביטויים). עם זאת, מכיוון שאחוזי הדיוק של ה-HMM (80%) ושל קו ההתחלה נמוכים יותר (71%) נצפה שאחוזי הדיוק של המערכת המשולבת יהיו נמוכים יותר מאלו של מערכת האנטרופיה המקסימלית. כאמור, המטרה היא להביא לאיזון ככל האפשר בין הדיוק של המערכת וה-recall שלה.

תוצאות המערכת המשולבת מוצגות בטבלה הבאה:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	84.54	74.31	79.1
PERS	90.66	73.82	81.38
LOC	83.09	82.8	82.94
ORG	77.14	62.03	68.77
DATE	90.2	85.18	87.62
MONEY	85.71	85.71	85.71
TIME	77.78	87.5	82.35
PERCENT	97.83	86.67	91.91

טבלה 8-1: תוצאות המערכת המשולבת

המערכת המשולבת מצליחה לצמצם את אחוז השגיאה שהושג ע"י מערכת האנטרופיה המקסימלית בכ-2.5%. כצפוי, רמת הדיוק של המערכת ירדה וה-recall עלה עבור כל סוגי התגים. רמת הזיהוי השתפרה עבור כל הביטויים פרט לביטויי שם מקום, שם היא נשארה זהה. אחוז השגיאה בזיהוי ביטויי שם אדם וארגון הצטמצם בכ-2.5%. זיהוי ביטויי זמן השתפר משמעותית ואחוז השגיאה הצטמצם בכ-40%.
אחוז השגיאה עבור כל סוגי הביטויים נמוך מ-19% פרט לביטויי ארגון.
זיהוי ביטויי ארגון משיג את התוצאה הנמוכה ביותר עבור כל השפות. אחוז השגיאה בזיהוי שמות ארגונים גבוה מזיהוי שאר הביטויים בכ-5%-10% עבור אנגלית וגרמנית. ממבחני ההסכמה ניתן לראות שרמת ההסכמה בקרב מתייגים ידניים עבור ביטויי ארגון גם היא נמוכה מרמת ההסכמה על שאר הביטויים. ביטויי ארגון הם הקשים ביותר לזיהוי בשל המגוון הרב שלהם והיעדר תבנית תחבירית מוגדרת. ביטויים אלה לרוב מתויגים חלקית ע"י המערכת כאשר המילים אשר מקבלות תיוג שגוי הן בד"כ שמות עצם בתוך הביטוי.

9 השוואה למערכות קיימות

9.1 השוואה למערכת אנטרופיה מקסימלית של גנדי למברסקי

כאמור בפרק עבודות קודמות, **גנדי למברסקי** מאוניברסיטת בן גוריון פיתח בשנת 2003 מערכת אנטרופיה מקסימלית לזיהוי שמות פרטיים [8]. המערכת הייתה מבוססת על קורפוס קטן יחסית (קורפוס של כ-25 אלף מילים ששימש הן לאימון והן לבדיקה) ועל קבוצת מאפיינים קטנה.

קו ההתחלה במערכת זו חושב בעזרת מערכת מבוססת מילון בלבד. תוצאות קו ההתחלה היו F=58.6%. תוצאות אלו גבוהות בהרבה מהתוצאות שהתקבלו באמצעות מערכת דומה על הקורפוס והמילונים הנוכחיים. במחקר זה, כמתואר בסעיף התוצאות של מערכת האנטרופיה המקסימלית, התקבלה תוצאה של F=42.4% בהתבסס על מילון בלבד. ניתן להסיק מכך שחלק גדול מביטויי השמות בקורפוס ששימש לעבודתו של למברסקי היה מצוי במילונים.

נשווה בין מערכת האנטרופיה המקסימלית שהוצגה כאן למערכת של למברסקי. ההבדל בין שתי מערכות האנטרופיה המקסימלית נעוץ בקבוצות מאפיינים שונות, שכן נעשה שימוש באותו כלי מדף ובאותה שיטת חיפוש.

בטבלה הבאה מתוארים המאפיינים השונים של שתי המערכות:

גנדי למברסקי	שלי
המילה הנוכחית	מילה נוכחית ומילים בחלון של ±2 סביבה
מילונים. ערך מילוני עבור המילה הנוכחית	מילונים מורחבים. ערך מילוני עבור המילה הנוכחית וחלון של ±1 סביבה
	לקסיקון שמות עצם נפוצים
ביטויים רגולריים: • ביטוי תאריך ("20 באוגוסט", "20.8.05", "1999") • אחוזים • מועדון ספורט ("מכבי ת"א) • ביטוי כסף ("מיליון דולר") • זמן ("12:30") • מילים לועזיות	ביטויים רגולריים (כמתואר בפרק מאפיינים) • ביטויי תאריך • אחוזים • ביטויי כסף • זמן • מילים לועזיות • שמות מקומות • שמות ארגונים • מילים המכילות סימנים מיוחדים • מילים המופיעות בתוך מרכאות • ביטויי מספר

למה	למה
חלק דיבר של המילה הנוכחית	חלקי דיבר של המילה הנוכחית וחלון של ±2 סביבה
	תחיליות
	סופיות
תג קודם של הלמה	תג קודם של המילה
	תיוג קודם של המערכת (2 מילים קודמות למילה הנוכחית)
	סמיכות
	עיבוד מוקדם
	מכפלות מאפיינים

טבלה 9-1: השוואת מאפיינים עם גנדי למברסקי

בעבודת התזה שלו דיווח למברסקי על הקטנת אחוזי השגיאה בכ-20% מקו ההתחלה. אם נעשה השוואה דומה בין תוצאות המערכת מבוססת מילון בלבד למערכת שלי (מערכת האנטרופיה המקסימלית בלבד) נמצא ירידה של כ-35% באחוזי השגיאה.

לשם השוואה מדויקת יותר נבדקה המערכת של למברסקי על קבצי הבדיקה של המערכת הנוכחית. התקבלו תוצאות נמוכות מאוד. לשיפור המערכת, היא אומנה על הקורפוס החדש (הגדול יותר). נעשה שימוש במילונים המורחבים ובמפיג העמימות המורפולוגית של מני אדלר [22] (אשר אחוז השגיאה שלו נמוך מאחוז השגיאה של המנתח בו השתמש למברסקי).

התוצאות שהתקבלו מוצגות בטבלה הבאה:

TEXT & TYPE			
	% Precision	% Recall	% F-measure
TOTAL	84.28	51.73	64.11
PERS	89.26	57.81	70.17
LOC	80.4	57.45	67.01
ORG	70.38	30.9	42.95
DATE	89.61	57.05	69.72
MONEY	50	45.54	47.67
TIME	0	0	0
PERCENT	50	30.77	38.1

טבלה 9-2: תוצאות המערכת של גנדי למברסקי

ניתן לראות כי גודל וסוג הקורפוס והמילונים משמעותיים מאוד, שכן בעזרתם הצטמצם אחוז השגיאה בכ-30% מהבדיקה ההתחלתית (לפני האימון המחודש). זיהוי שמות אדם הוא הגבוה ביותר, אולם זיהוי שמות מקומות אינו גבוה מקו ההתחלה כפי שהוגדר בעבודה זו.
אחוז השגיאה גבוה בכ-12% מאחוז השגיאה של מודל האנטרופיה המקסימלית שלי (וב-15% מהמערכת המשולבת). יש לציין כי חלק מהטעויות שנעשו נבעו מכך שהגדרת התיוג הייתה קצת שונה בין שתי המערכות. למשל למברסקי תייג ביטויים כגון "אתמול" ו-"בערב" כביטויי תאריך ובעבודה זו הם לא הוגדרו כך.
קבוצת המאפיינים שהוצגה בעבודה זו רחבה יותר ומאפשרת למידה מעמיקה יותר של ההקשר של המילה הנוכחית. הבדל זה מאפשר דיוק גבוה יותר של המערכת ואפשרות לזהות יותר ביטויים.
הבדל בולט בין העבודות הוא חוסר במאפיינים עבור המילים סביב המילה הנוכחית בעבודה של למברסקי. בעבודה זו נאספו מאפיינים שונים גם עבור מילים בחלון של 2± סביב המילה הנוכחית (מאפיינים כגון ביטויים רגולריים, חלקי דיבר וערכי לקסיקון). שימוש במאפייני חלקי דיבר בחלון סביב המילה הנוכחית חשוב מאוד שכן הם מספקים מידע על המבנה התחבירי של המשפט כולו. כמו כן הערך המילוני של המילים סביב המילה הנוכחית חשוב עבור ביטויים שאורכם ארוך ממילה אחת.
הבדל חשוב נוסף הוא שימוש בתחיליות, סופיות וסמיכות כפי שהתקבלו ממפיג העמימות המורפולוגית. מאפיינים אלה מאפשרים פירוש נכון של המילה ותפקידה במשפט (עד כדי אחוז השגיאה של מפיג העמימות).
מעבר למערכת האנטרופיה המקסימלית במחקר זה נבנתה מערכת המשלבת מודלים שונים. שילוב המודלים אפשר צמצום אחוז השגיאה אף יותר.

9.2 השוואה למערכות עבור השפה האנגלית

נעשתה עבודה רבה על מערכות עבור השפה האנגלית. במשימת CoNLL2003 [23] הוצגו מספר עבודות והתאפשרה השוואה הוגנת בין המערכות השונות.
התוצאות הטובות ביותר עבור זיהוי שמות פרטיים באנגלית שנצפו ב-CoNLL2003 היו של המערכת של **Florian** , **Ittaycheriah**, **Jing**, **Zhang** [6]. המערכת הציגה מודל ששילב מערכות אנטרופיה מקסימלית, מודל מרקוב ועוד (ראה פרק 2: עבודות קודמות). המערכת הציגה תוצאות של F=88.76%.
מערכת מבוססת אנטרופיה מקסימלית בלבד שהוצגה ב-CoNLL2003 של **Chieu** ו- **Ng** [7] הגיעה למקום השני עם תוצאות של F=88.31%.

יש לציין שבמשימת CoNLL2003 אומנו המערכות על קורפוס אימון בגודל 250,000 מילים לעומת הקורפוס הצנוע שלנו שגודלו כ-57,000 מילים.

בשפה האנגלית מאפיינים של שימוש באותיות גדולות בתחילת מילה או בתוך מילה משמעותיים ביותר. המשימה העיקרית הנותרת היא להבחין בין סוגי הביטויים השונים. מאפיינים של התחלת מילה באות גדולה, רצפים של מילים המתחילות באותיות גדולות, מילה אשר כולה באותיות גדולות וכו' אינם רלוונטיים עבור עברית אולם תורמים רבות עבור אנגלית.

מאפיינים מסוימים שהוזכרו בעבודתם של Chieu ו- Ng שימשו גם בעבודה זו (למשל רשימות העיבוד המוקדם, ביטויים רגולריים מסוימים ועוד). מאפיינים אחרים אשר תרמו למערכת עבור השפה האנגלית לא נמצאו שימושיים עבור השפה העברית. Chieu ו- Ng ביצעו עיבוד מוקדם רחב מאוד על קורפוס האימון. הם אספו רשימות של תחיליות וסופיות נפוצות וכן צירופי מילים נפוצים עבור כל קטגוריה. מאפיינים אלו לא תרמו למערכת העברית שכן בעברית יש תפקיד שונה לתחיליות ולסופיות מאשר באנגלית. קשה למצוא בעברית תבניות חוזרות בתחיליות ובסופיות של מילים וביטויים באותה הקטגוריה. כמו כן, בעבודתם נעשה שימוש לא רק במאפיינים מקומיים עבור כל מילה אלא גם במאפיינים עבור הופעות נוספות של אותה המילה במאמר. גם למאפיין זה לא נראתה השפעה על המערכת העברית. הוספה של מאפיינים גלובליים לא העלתה ולא הורידה מביצועי המערכת. התיוג הינו החלטה מקומית, לרב גם אדם שקורא טקסט אינו נזקק למידע על הטקסט כולו אלא רק למשפט עצמו על מנת לקבל החלטה. לידע הגלובלי המתקבל אין השפעה רבה על החלטת המערכת.

בשל המידע הרב שמספקות האותיות הגדולות באנגלית, טבעי לצפות שהתוצאות בשפה בה יש שימוש באותיות אחידות עבור כל הביטויים התוצאות יהיו נמוכות יותר.

Andrew Borthwick [5] ביצע ניסויים בעזרת מערכת אנטרופיה מקסימלית עבור טקסט שכולו אותיות גדולות. ע"י שימוש באותה קבוצת מאפיינים (ראה פרק 2: עבודות קודמות) התקבלה תוצאה של F=77.98%. מערכת **IsoQuest** המבוססת על כללים ידניים השיגה תוצאות של F=81.96% עבור טקסט באותיות גדולות. תוצאות אלו כבר ברות השוואה לתוצאות העברית. תוצאות המערכת שלנו גבוהות מתוצאותיו של Borthwick בכאחוז ונמוכות בכ-3% מתוצאות המערכת הידנית.

הבעיה עבור השפה העברית עדיין קשה יותר מאשר הבעיה עבור טקסט אנגלי שכולו באותיות גדולות. זאת בשל תופעת הצמדת התחיליות והסופיות למילה עצמה וכן תופעת הטיית המילים. בעיה מרכזית בשפה העברית היא בעיית פירוק המילה: הפרדת התחיליות והסופיות מהמילה עצמה ומציאת הלמה.

9.3 השוואה למערכות עבור שפות נוספות

בשפות אחרות נעשתה פחות עבודה וכן דרוש ידע לשוני מקצועי על מנת להשוות את קושי משימת התיוג בשפות השונות.

במשימת CoNLL2003 הוצגו תוצאות המערכות השונות עבור השפה הגרמנית והתוצאה הטובה ביותר הייתה של מערכת ששילבה מספר מודלים F=72.41% [6].

במשימת CoNLL2002 נבדקו ביצועי מערכות שונות בשפות ספרדית והולנדית. בשפה הספרדית התוצאה הטובה ביותר הייתה F=81.39%. בהולנדית התוצאה הטובה ביותר הייתה F=77.05%. מעניין לציין שעבור שפות אלו התוצאות הטובות היותר לא הושגו ע"י מערכות אנטרופיה מקסימלית.

בהשוואה לתוצאות המתוארות ניתן לומר שהתוצאות אשר הושגו עבור השפה העברית טובות, אולם לא ניתן לנתח את ההבדלים ללא ידע מקצועי נוסף.

10 סיכום ומסקנות

משימת זיהוי שמות פרטיים בטקסט יכולה להיות משימה קלה עבור אנשים אולם היא קשה יחסית עבור מחשב. השפה העברית מציגה אתגר נוסף במשימה זו שכן השמות הפרטיים בה אינם מתחילים באות גדולה ולכן קשים לזיהוי בצורה אוטומטית. מורכבות השפה מביאה פעמים רבות למצב בו לכל מאפיין יש מספר עדויות מועט, דבר המקשה על מערכת סטטיסטית. ממחקרים קודמים עלה כי מודל האנטרופיה המקסימלית הוא המתמודד עם בעיה זו בצורה הטובה ביותר.

כשלב מקדים לבניית המערכת נוסחו הנחיות לתיוג ספציפיות לשפה העברית. על הנחיות אלה הושגה הסכמה גבוהה בקרב מתייגים אנושיים (ממוצע של כ-99.5% הסכמה).

במהלך המחקר נבנתה מערכת אוטומטית לזיהוי שמות פרטיים המשלבת גישות שונות.

המערכת הציגה את התוצאות הטובות ביותר עבור השפה העברית עד כה. אחוז השגיאה בקו ההתחלה עבור משימה זו הוא כ-42.5%. המערכת הסופית הצליחה לצמצם את אחוז השגיאה בכ-21%. תוצאות אלו טובות יחסית לתוצאות שהושגו עבור טקסט באנגלית שכולו באותיות גדולות ויחסית לתוצאות בשפות אחרות. רמת הדיוק של המערכת גבוהה – 84.5% ואילו ה-recall נמוך יותר – 74.3%. כלומר, כ-25% מהביטויים אינם מזוהים אך אלה שמזוהים מזוהים עם אחוז שגיאה של כ-15%.

במהלך המחקר נבנו שני מודלים אוטומטיים, HMM ומודל אנטרופיה מקסימלית. מודל האנטרופיה המקסימלית היה מוצלח הרבה יותר. המערכת אשר הציגה את התוצאות הטובות ביותר היא המערכת אשר שילבה שלושה מודלים: אנטרופיה מקסימלית, HMM ומודל קו ההתחלה המבוסס על ביטויים רגולריים ולקסיקון הנבנה מקורפוס האימון. מיזוג תוצאות שלושת המודלים הביאו לצמצום אחוזי השגיאה של מודל האנטרופיה המקסימלית בכ-2.5%.

המחקר התמקד בבחינת השפעת מאפיינים שונים על המודלים. נמצא כי המאפיינים החשובים ביותר למשימת התיוג הם מאפיינים מקומיים אשר מתחשבים במילה הנוכחית וחלון של עד ±2 מילים סביבה. ראינו כי המאפיינים הדומיננטיים ביותר הם מאפייני חלקי הדיבר ומאפייני המילון. עם זאת, התוצאות הטובות ביותר מתקבלות משילוב קבוצה גדולה ומגוונת יותר של מאפיינים ומשילוב המודלים השונים. זיהוי ביטויי שם ארגון הוא הקשה ביותר והציג תוצאות נמוכות יחסית. זאת בשל מורכבותם וגיוונם של ביטויים אלה. וכן בשל העובדה שבתוך ביטויים אלה לרוב יש שימוש בשמות עצם ובמילים נפוצות אשר לא מופיעות תמיד בהקשר של שם פרטי.

בניתוח תוצאות המערכת יש לקחת בחשבון את אחוז השגיאה של מנתח חלקי הדיבר והמנתח המורפולוגי (5% - 8%). שמות פרטיים הם בעיה קשה גם עבור המנתח המורפולוגי והוא נוטה לסמן שמות פרטיים רבים כמילים לא מוכרות.

11 הצעות למחקר עתידי

- הקורפוס אשר שימש אותנו לאימון ובדיקה הוא בגודל של כ-57,000 מילים. קורפוס זה קטן יחסית. למשל במשימת CoNLL2003 השתמשו בקורפוס גדול פי 5. ראינו כי גודל הקורפוס וסוגו משפיעים רבות על המערכת, שכן ניתן למצוא עדויות רבות יותר לכל מאורע. עבודה עם קורפוס גדול יותר יכולה לשפר את תוצאות המערכת. תיוג של קורפוס גדול היא משימה ידנית הדורשת עבודה רבה של מספר אנשים.
- הגדלת הקורפוס יכולה לתרום לשיפור שיטת המיזוג בין המודלים השונים. בהינתן קורפוס גדול ניתן להקצות חלק ממנו לאימון מערכת סטטיסטית המחליטה איזה מהתיוגים של המודלים השונים לבחור.
- במהלך מחקר זה ניסינו להרחיב את משימת התיוג לסוגי תגים שונים. בהנחיות לתיוג הוגדרו גם קטגוריות "שונות": שמות אירועים, דגמים, סרטים, שירים וביטויי השתייכות לקבוצה. בניסיון לשלב תגים אלה, לא נפגעה רמת הזיהוי של התגים הקיימים אולם המערכת לא הראתה תוצאות טובות בזיהוי ביטויי "שונות". ביטויים אלה הם בעלי אופי שונה ודורשים מאפיינים מיוחדים. אנחנו העדפנו להתרכז תחילה בביטויים בסיסיים יותר. כעת, בהינתן מערכת הנותנת תוצאות טובות יחסית ניתן להתמקד בסוגים אחרים של ביטויים.
- כאמור בפרק עבודות קודמות, מערכות ידנית הן אשר הציגו את התוצאות הטובות ביותר. פיתוח מערכות כאלה דורש ידע לשוני מקצועי. עם זאת ניתן לשלב כמות מסוימת של כללים ידניים במודל האנטרופיה המקסימלית בצורת מאפיינים. רוב המאפיינים ששימשו למחקר זה היו אוטומטיים. מאפיינים ששילבו מעט ידע חיצוני על השפה היו ביטויים רגולריים עבור זיהוי שמות מקום וארגון. מאפיינים אלה הצליחו לצמצם את אחוז השגיאה של המערכת ולהעלות את רמת הדיוק שלה.
- מערכת ה-HMM שהוצגה במחקר זה מבוססת על הגדרת ה-HMM הקלאסית. ניתן לבנות HMM מורכב יותר בדומה למודל ה-"IdentiFinder " שהוצג בפרק עבודות קודמות. בנוסף פותחו מודלים עבור תהליכים מרקוביים המשלבים מאפיינים (למשל DBN). יתכן ומערכות אלו יצליחו לצמצם את אחוזי השגיאה אף יותר.
- ניתן למקצע את המערכת ע"י התמקדות בסוג מסוים של טקסטים. אם אנו רוצים לתייג טקסטים בעלי אופי מוגדר (למשל טקסטים צבאיים, כתבות רכילות) ניתן לאמן מערכת על טקסטים בעלי אופי זה. ניתן להניח כי בקורפוס המכיל כתבות עם אופי דומה יש פחות אי סדר ולכן נצפה לתוצאות טובות יותר.

12 נספח 1: קוים מנחים לתיוג שמות פרטיים בעברית

1. כללי

1.1 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

<MISC> - שונות

דוגמאות:

"ממשלת קלינטון"

ממשלת <PERS> קלינטון </PERS>

"יבואני ארה"ב"

יבואני <LOC> ארה"ב </LOC>

"מחשבי אפל"

מחשבי <ORG> אפל </ORG>

"כנסת ישראל"

<ORG> כנסת ישראל </ORG>

"יו"ר הליכוד, אריאל שרון, ..."

יו"ר <ORG> הליכוד </ORG> , <PERS> אריאל שרון </PERS>, ...

1.2 תחיליות יכללו בתיוג

"יו"ר הליכוד"

יו"ר <ORG> הליכוד </ORG>

1.3 ניתן לפרק ביטויי סמיכות ולתת תיוגים שונים לכל חלק

"משפחת כהן"

משפחת <PERS> כהן </PERS>

1.4 במקרים מסוימים ביטויים באורך מספר מילים יכילו תתי מחרוזות שיכולות לקבל תיוג שונה. במקרים כאלה לא יהיה תיוג נפרד לתתי המחרוזות.

"עוף הנגב"

<ORG> עוף הנגב </ORG>

(אין תיוג נפרד ל"נגב" כשם מקום).

"פסטיבל עכו"

<MISC> פסטיבל עכו </MISC >

"מכבי חיפה מוזילה כרטיסים"

<ORG> מכבי חיפה </ORG> מוזילה כרטיסים

1.5 בביטויי שייכות תתויג כל תת מחרוזת.

"בית הספר לניהול של אוניברסיטת בן גוריון"

<ORG> בית הספר לניהול </ORG> של <ORG> אוניברסיטת בן גוריון </ORG>

1.6 כינויים וקיצורים (למשל: ארה"ב, אי.בי.אם., "התפוח הגדול" (כינוי לניו יורק) יתויגו כשמות.

"הזמרת רוני סופרסטאר תופיע הערב"

הזמרת <PERS> רוני סופרסטאר </PERS> תופיע הערב

"הבית הלבן מוסר..."

<ORG> הבית הלבן </ORG> מוסר...

1.7 מרכאות יכללו בתיוג רק אם הן מופיעות באמצע השם.

"ויטו "הסנדק" קורליאונה"

<PERS> ויטו "הסנדק" קורליאונה </PERS>

2. PERS

2.1 קידומות כגון: מר, גברת, יו"ר, פרופ', רב לא יתויגו כחלק מהשם. בדומה עבור סיומות כגון ז"ל.

2.2 שמות משפחה יתויגו כ-PERS.

"הפרידמנים ישבו לארוחת הצהריים"

<PERS> הפרידמנים </PERS> ישבו לארוחת הצהריים

"משפחת כהן יצאה לטייל"

משפחת <PERS> כהן </PERS> יצאה לטייל

2.3 שמות בעלי חיים יתויגו כ-PERS.

"כלבו של ראש הממשלה, רקסי, נלקח לווטרינר"

כלבו של ראש הממשלה, <PERS> רקסי </PERS>, נלקח לווטרינר"

2.4 דמויות בדיוניות (גיבורי טלוויזיה, סרטים מצוירים וכ"ו) יתויגו כ-PERS.

"באטמן הוא דמות מפורסמת"

<PERS> באטמן </PERS> הוא דמות מפורסמת

"הלכנו לראות את הסרט באטמן"

הלכנו לראות את הסרט <MISC> באטמן </MISC>

"גרפילד אוהב לזניה"

<PERS> גרפילד </PERS> אוהב לזניה

2.5 לא יתויגו כ-PERS שמות הניתנים על-שם אדם : פסקי דין, מחלות, פרסים וכו' (למשל: "פרס נובל", "הוריקן מייקל" – ראה סעיף שונות).

LOC .3

3.1 מקומות: מקומות גיאוגרפיים, פוליטיים, אסטרונומיים, מדינות, יבשות, מחוזות, ערים, כפרים, שכונות, כבישים, שדות תעופה, גשרים, בסיסים צבאיים, רחובות (יכול לכלול גם מספר+רחוב), ימים, חופים, הרים, מבנים ("מגדל אייפל"...), תעלות, אגמים.

3.2 שם מקום יכול להיכלל בשם הארגון ולא יתויג בנפרד

"אוניברסיטת בן גוריון בנגב"
<ORG> אוניברסיטת בן גוריון בנגב </ORG>

"אפריקה ישראל להשקעות בע"מ היא חברת השקעות בינלאומית"
<ORG> אפריקה ישראל להשקעות בע"מ </ORG> היא חברת השקעות בינלאומית

3.3 בביטוי המורכב משמות מקומות עוקבים (מופרדים בפסיק או לא) יתויג כל מקום בנפרד.

"ירושלים, ישראל"
<LOC> ירושלים </LOC> , <LOC> ישראל </LOC>

"רח' ביאליק 114 באר שבע"
<LOC> רח' ביאליק 114 </LOC> <LOC> באר שבע </LOC>

3.4 שמות עצם:

"רחוב" "שדרות", "שכונה", "צומת" ההכרחיים לזיהוי המקום יתויגו כחלק מהביטוי ("עיר", "ישוב" וכו' לא יתויגו).
(אם ניתן להשמיט את שם העצם ושם המקום עדיין נשאר ברור, הוא אינו חלק מהביטוי)

"שדרות רוטשילד"

<LOC> שדרות רוטשילד </LOC>

"ר"ח שאול המלך"

<LOC> ר"ח שאול המלך </LOC>

"צומת גילת"

<LOC> צומת גילת </LOC>

"בסמוך לעיר באר שבע...."

בסמוך לעיר <LOC> באר שבע </LOC>....

"היישוב רהט"

היישוב <LOC> רהט </LOC>

ישנם שמות עצם שנהוג להצמידם לשם המקום (צורת דיבור) – גם אלה יכללו כחלק מהביטוי.

"קיבוץ עין גדי"

<LOC> קיבוץ עין גדי </LOC>

שמות עצם המזוהים עם שם המקום יתויגו יחד עם השם (חלק מהשם הרשמי, או צורת דיבור).

"ים המלח"

<LOC> ים המלח </LOC>

"עמק יזרעאל"

<LOC> עמק יזרעאל </LOC>

"נהר הירדן"

נהר <LOC> הירדן </LOC>

"כדור הארץ"

<LOC> כדור הארץ </LOC>

"ארץ ישראל"

<LOC> ארץ ישראל </LOC>

3.5 שמות כבישים יתויגו כביטוי אחד הכולל את שם העצם "כביש".

"כביש ירושלים תל-אביב"

<LOC> כביש ירושלים תל-אביב </LOC>

3.6 לא יתויגו שמות עצם המתארים את המקום

"חופי הים התיכון"

חופי <LOC> הים התיכון </LOC>

3.7 ביטויי כיוון (צפון, דרום וכו') יתויגו רק כאשר הם חלק מהשם הרשמי

"מזרח אירופה"

<LOC> מזרח אירופה </LOC>

"איזור הדרום"

איזור הדרום (ללא תיוג)

"המזרח התיכון"

<LOC> המזרח התיכון </LOC> (כינוי)

"מרכז אסיה"

מרכז <LOC> אסיה </LOC>

3.8 שמות מקומות בצורת תואר השם לא יתויגו כשם מקום (ראה סעיף 9 - שונות).

"חיל האוויר האמריקאי"

<ORG> חיל האוויר </ORG> <MISC_AFF> האמריקאי </MISC_AFF>

"סטודנטים ישראלים"

סטודנטים <MISC_AFF> ישראלים </MISC_AFF>

3.9 כינויים וקיצורים יתויגו כשמות (ראה 1.6).

"מהומות בשטחים"

מהומות <LOC> בשטחים </LOC>

"ישובי הבקעה"

ישובי <LOC> בשטחים </LOC>

(קיצור לבקעת ירדן)

3.10 ארצות הברית" ו-"ברית המועצות" יתויגו כשמות מקומות .

3.11 כאשר שם מקום מופיע לבד השם תמיד יתויג כשם מקום.

"ארה"ב תקפה את עיראק"

"<LOC> ארה"ב </LOC> תקפה את <LOC> עיראק </LOC>

"במשחק הכדורגל בשבת סכנין ניצחה את בני יהודה"

במשחק הכדורגל <DATE> בשבת </DATE> <LOC> סכנין </LOC> ניצחה את <ORG> בני יהודה </ORG>

"ישראל תשחרר אסירים"

<LOC> ישראל </LOC> תשחרר אסירים

"מדינת ישראל עומדת בפני הכרעה היסטורית"

<ORG> מדינת ישראל </ORG> עומדת בפני הכרעה היסטורית

4. ORG

4.1 שמות ארגונים: עסקים, תחנות רדיו/טלוויזיה, עיתונים, ארגונים פוליטיים, מפלגות,דתות, קבוצות דתיות, ארגונים דתיים, בנקים, תזמורות, משרדים ממשלתיים, קבוצות ספורט, צבאות, חטיבות צבאיות, חיילות, ארגונים כלכליים, מלונות, אוניברסיטאות, בתי חולים, ועדות.

"ממשלת צפון קוריאה"

<ORG> ממשלת צפון קוריאה </ORG>

"מכבי תל אביב"

<ORG> מכבי תל אביב </ORG>

"מאה ושתיים אפ.אם."

<ORG> מאה ושתיים אפ.אם. </ORG>

"ערוץ 2"

<ORG> ערוץ 2 </ORG>

"נסדאק"

<ORG> נסדאק </ORG>

" הפצועים פונו לבית החולים הדסה בירושלים"

הפצועים פונו <ORG> לבית החולים הדסה </ORG> <LOC> בירושלים </LOC>

"אוניברסיטת תל-אביב"

<ORG> אוניברסיטת תל-אביב </ORG>

"משרד העבודה והרווחה"

<ORG> משרד העבודה והרווחה </ORG>

"אלוף פיקוד דרום"

אלוף <ORG> פיקוד דרום </ORG>

4.2 סיומות ארגונים (למשל "בע"מ") יתויגו כחלק מהשם.

4.3 בביטויים מורכבים תתויג כל תת מחרוזת (ראה 1.5).

"בית המשפט המחוזי בירושלים"

<ORG> בית המשפט המחוזי </ORG> <LOC> בירושלים </LOC>

"ועדת חוץ וביטחון של הכנסת"

<ORG> ועדת חוץ וביטחון </ORG> של <ORG> הכנסת </ORG>

4.4 שמות אירועים לא מתויגים כארגונים (ראה סעיף 9 - שונות), אולם ועדות וארגונית הקשורים אליהם יתויגו כשם ארגון.

"הועד האולימפי הישראלי"

<ORG> הועד האולימפי הישראלי </ORG>

4.5 במקרים בהם שם יצרן ומוצר מוזכרים, שם היצרן לא יתויג בנפרד (ראה סעיף 9 - שונות).

"פורד פוקוס"

<MISC> פורד פוקוס </MISC >

4.6 שמות כללים כגון "הממשלה", "הצבא", "המשטרה" לא יתויגו (לעומת "ממשלת ישראל", "צה"ל", "משטרת ישראל", שיתויגו כ-ORG).

4.7 שמות עצם יתויגו רק אם הם חלק מהשם הרשמי של הארגון.

"מפלגת הליכוד"

מפלגת <ORG> הליכוד </ORG>

"ערוץ 7"

<ORG> ערוץ 7 </ORG>

"הארגון הצבאי הלאומי"

<ORG> הארגון הצבאי הלאומי </ORG>

"ארגון החמאס"

ארגון <ORG> החמאס </ORG>

4.8 כינויים וקיצורים יתויגו כשמות (ראה 1.6).

"ההסתדרות תשבית את הנמלים"

<ORG>ההסתדרות</ORG> תשבית את הנמלים

(קיצור של הסתדרות העובדים)

"האוצר החליט על קיצוצים..."

<ORG>האוצר</ORG> החליט על קיצוצים...

(קיצור של משרד האוצר)

4.9 הביטוי "עם ישראל" יתויג כשם ארגון.

5. TIME

5.1 ביטוי מדויק של זמן: "10 דקות אחרי 5" (לעומת "מספר דקות אחרי 5"), "חצות" (לעומת "בוקר", "ערב").

"הכנס יתחיל בשעה 8:10"

הכנס יתחיל <TIME>בשעה 8:10</TIME>

"25/11/04 8:00"

<DATE>25/11/04</DATE> <TIME>8:00</TIME>

"המסיבה מתחילה בחצות"

המסיבה מתחילה <TIME>בחצות</TIME>

5.2 במקרים של ביטויי שייכות הביטוי יתויג כביטוי אחד

"ארבע לפנות בוקר"

<TIME>ארבע לפנות בוקר</TIME>

5.3 המילה "שעה" על כל צורותיה תתויג כחלק מהביטוי במידה והיא מופיעה בצמוד לביטוי שעה.

"קבענו לשעה 12:00"

קבענו <TIME>לשעה 12:00</TIME>

5.4 אזורי זמן יתויגו כחלק מהביטוי

"GMT 5:00"

<TIME> GMT 5:00 </TIME>

"23:00 לפי זמן ישראל"

<TIME> 23:00 לפי זמן ישראל </TIME>

"בישראל, ביום שלישי בשעה שמונה..."

<LOC> בישראל </LOC>, < DATE> ביום שלישי < /DATE> < TIME> בשעה
שמונה </TIME> ...

6. DATE

6.1 ביטוי תאריך מדייק: ציון יום מדויק ("ראשון"), תאריך מדויק ("1 בנובמבר" לעומת "ראש חודש"), עונה ("אביב"), שנה מדויקת ("2005" לעומת "שנה הבאה", "השנה"), עשור ("שנות ה-80" לעומת "עשר השנים האחרונות" "המאה ה-20" לעומת "המאה הנוכחית").

"ינואר 2005"

<DATE> ינואר 2005 </DATE>

"בסביבות ד' בתשרי"

בסביבות ה- <DATE> ד' בתשרי </DATE>

"אוגוסט שנה שעברה"

<DATE> אוגוסט </DATE> שנה שעברה

"במהלך החורף השנה ירדו משקעים רבים"

במהלך <DATE> החורף </DATE> השנה ירדו משקעים רבים

"האביב הקרוב"

<DATE> האביב </DATE> הקרוב

6.2 שמות עצם לא יתויגו כחלק מהשם ("חודש", "עונה" וכו'), פרט למילה "יום" המופיעה בצמוד לציון אחד מימות השבוע.

"יום ג' הקרוב"

<DATE> יום ג' </DATE> הקרוב

"החגיגות יערכו במשך חודש יולי"

החגיגות יערכו במשך חודש <DATE> יולי </DATE>

6.3 במקרים של ביטויי שייכות מדויקים הביטוי יתויג כביטוי אחד

"רבע שלישי של 1990"

<DATE> רבע שלישי של 1990 </DATE>

"שלישי שעבר"

<DATE> שלישי </DATE> שעבר

6.4 ימים מיוחדים, חגים למשל , יתויגו כתאריך. במקרה של ביטוי תאריך המציין חג, שם העצם "חג" יתויג כחלק מהביטוי.

"ביום העצמאות האחרון התקיימה תהלוכה"

<DATE> ביום העצמאות </DATE> האחרון התקיימה תהלוכה

"כוננות גבוהה לקראת חג הפורים"

כוננות גבוהה לקראת <DATE> חג הפורים </DATE>

"יום הזיכרון לחללי צה"ל"

<DATE> יום הזיכרון לחללי צה"ל </DATE>

6.5 ביטויי תאריך בהתאם ללוחות השנה השונים (סיני, עברי וכו') יתויגו.

MONEY .7

דוגמאות:

"עשרים מיליון שקלים חדשים"

<MONEY> עשרים מיליון שקלים חדשים </MONEY>

"$100"

<MONEY> $100 </MONEY>

"הפרויקט מוערך בשווי של מיליוני דולרים"

הפרויקט מוערך בשווי של <MONEY> מיליוני דולרים </MONEY>

7.1 ביטויי כסף ביחידות נפרדות יתויגו בנפרד

"$20 (90 ש"ח)"

<MONEY> $20 </MONEY> (<MONEY> 90 ש"ח </MONEY>)

7.2 הערכות לא יתויגו (פרט לתחיליות-ראה סעיף 1.2)

"יותר ממאה שקלים"

יותר <MONEY> ממאה שקלים </MONEY>

"כמאה שקלים"

<MONEY> כמאה שקלים </MONEY>

7.3 כמתים יתויגו רק אם ניתן להחליפם במספר

"כמה אלפי דולרים"

<MONEY> כמה אלפי דולרים </MONEY>

8. PERCENT

8.1 המילה "אחוז" על כל צורותיה תתויג כחלק מהביטוי

"שלושת רבעי אחוז"

<PERCENT> שלושת רבעי אחוז </PERCENT>

"10%"

<PERCENT> 10% </PERCENT>

"התוצאות השתפרו משמונים אחוזים בשנה שעברה לתשעים השנה "

התוצאות השתפרו <PERCENT> משמונים אחוזים </PERCENT>

בשנה שעברה <PERCENT> לתשעים </PERCENT> השנה

"סיכויים של 50-50"

סיכויים של <PERCENT> 50 </PERCENT> - <PERCENT> 50
</PERCENT>

(המקף לא מתויג כחלק מהביטוי שכן הוא מפריד בין שני ביטויים שונים.)

"40-30 אחוזים"

<PERCENT> 30 </PERCENT> - <PERCENT> 40 אחוזים </PERCENT>

8.2 המילה מינוס (סימן -) תתויג כחלק מהביטוי

"מינוס 10 אחוז"

<PERCENT> מינוס 10 אחוז </PERCENT>

8.3 הערכות לא יתויגו (פרט לתחיליות-ראה סעיף 1.2)

"בערך 70%"

בערך <PERCENT> 70% </PERCENT>

"כ70%"

<PERCENT> כ 70% </PERCENT>

9. MISC

9.1 ביטויים המציינים שמות אירועים יתויגו כשונות (<MISC_EVENT>). שמות עצם יתויגו כחלק מהשם רק כאשר הם חלק מהשם הרשמי.

"גביע אירופה לאלופות"

<MISC_EVENT> גביע אירופה לאלופות </MISC_EVENT >

"פסטיבל ערד"

<MISC_EVENT> פסטיבל ערד </MISC_EVENT >

"מלחמת המפרץ"

<MISC_EVENT> מלחמת המפרץ </MISC_EVENT >

"רם וארליך העפילו לחצי גמר ווימבלדון"

<PERS> רם </PERS> <PERS> וארליך </PERS> העפילו לחצי גמר <MISC_EVENT> ווימבלדון </MISC_EVENT >

"האינתיפאדה אשמה בהפסדי המשק"

<MISC_EVENT> האינתיפאדה </MISC_EVENT > אשמה בהפסדי המשק

9.2 ביטויי השתייכות לקבוצה (למקום או לארגון) יתויגו כשונות (<MISC_AFF>) .

"אלפי מוסלמים משתתפים בתפילות בהר הבית"

אלפי <MISC_AFF> מוסלמים </MISC_AFF > משתתפים בתפילות <LOC> בהר הבית </LOC>

"הפרדה בין יהודים לערבים"

הפרדה בין <MISC_AFF> יהודים </MISC_AFF> <MISC_AFF> לערבים </MISC_AFF>

"עלייה במספר הפלשתינים התומכים בשלום"

עלייה במספר <MISC_AFF> הפלשתינים </MISC_AFF> התומכים בשלום

"המורדים הכורדים"

המורדים <MISC_AFF> הכורדים </MISC_AFF>

"האיחוד האירופי"

<ORG > האיחוד האירופי <\ORG >

"יורש העצר האוסטרי"

יורש העצר <MISC_AFF> האוסטרי </MISC_AFF >

"מאות חרדים הפגינו..."

מאות <MISC_AFF> חרדים </MISC_AFF > הפגינו...

9.3 שמות סרטים, תכניות טלוויזיה, ספרים, שירים, דוחות, הסכמים, פסקי דין וחוקים יתויגו כשונות (<MISC_ENT>).

"לונדון את קירשנבאום"

<MISC_ENT> לונדון את קירשנבאום </MISC_ENT >

"טולסטוי כתב את מלחמה ושלום"

<PERS> טולסטוי </PERS> כתב את <MISC_ENT> מלחמה ושלום </MISC_ENT>

"בסיום הטקס שרנו את 'התקווה'"

בסיום הטקס שרנו את ' <MISC_ENT> התקווה </MISC_ENT> '

"דו"ח מצא"

"דו"ח מצא"(ללא תיוג)

"חוק 'פינוי פיצוי' "

"חוק 'פינוי פיצוי' "

"הסכם אוסלו"

"הסכם אוסלו"

"תכנית 'מפת הדרכים' "

"תכנית 'מפת הדרכים' "

13 נספח 2: דוגמא לטקסט מתויג

דוגמא לטקסט מתוך קורפוס האימון. כל שורה מכילה טוקן, כאשר העמודה הראשונה היא הטוקן עצמו והשנייה היא התיוג הידני אשר ניתן לו. משפטים מופרדים ע"י שורה ריקה.

```
ב- O
25 I_DATE
לאוגוסט I_DATE
עצר O
השב"כ I_ORG
את O
מוחמד I_PERS
אבו-ג'וייד I_PERS
, O
אזרח O
ירדני O
, O
שגויס O
לארגון O
הפת"ח I_ORG
והופעל O
על O
ידי O
חיזבאללה I_ORG
. O

אבו-ג'וייד I_PERS
התכוון O
להקים O
חוליות O
טרור O
בגדה I_LOC
ובקרב O
ערביי O
ישראל I_LOC
, O
לבצע O
פיגוע O
ברכבת I_ORG
ישראל I_ORG
בנהריה I_LOC
, O
לפגוע O
במטרות O
ישראליות O
בירדן I_LOC
ולחטוף O
חיילים O
```

כדי O
לשחרר O
אסירים O
ביטחוניים O
. O

אבו-ג'וייד I_PERS
, O
נולד O
בסוריה I_LOC
, O
ובגיל O
7 O
עבר O
לירדן I_LOC
. O

ב- O
1998 I_DATE
הגיע O
אבו-ג'וייד I_PERS
להתגורר O
בשטחים I_LOC
תוך O
שהוא O
מבקר O
תכופות O
בירדן I_LOC
. O

בשנת O
2001 I_DATE
שב O
אבו-ג'וייד I_PERS
לירדן I_LOC
שם O
פגש O
פעיל O
של O
פלג O
" O
אבו-מוסא I_ORG
" O
בפת"ח I_ORG
שהציע O
לו O
להצטרף O
ל O
" O
מאבק O
הפלשתיני O
" O

. O

אבו-ג'וייד I_PERS
הסכים O
, O
ויצא O
לאימונים O
בסוריה I_LOC
, O
שם O
הוכשר O
בירי O
נשק O
כבד O
, O
ייצור O
מטענים O
ושימוש O
בחומרי O
נפץ O
. O

לאחר O
סיום O
האימונים O
קיבל O
אבו-ג'וייד I_PERS
תשלום O
ושב O
לירדן I_LOC
. O

למרות O
שהובהר O
לו O
שהוכשר O
כדי O
לסייע O
למאבק O
הפלשתיני O
, O
ב- O
2003 I_DATE
החליט O
אבו-ג'וייד I_PERS
לצאת O
לעיראק I_LOC
כדי O
לסייע O
בלחימה O
נגד O
האמריקנים O

. O

כשבועיים O
לחם O
אבו-גוויד I_PERS
עם O
המורדים O
העיראקיים O
אך O
נפצע O
קל O
מירי O
בראשו O
וחזר O
לירדן I_LOC
. O

14 סימוכין

[1] Grishman, R: The NYU system for MUC-6 or where's the syntax?, In: *Proceedings of the Sixth Message Understanding Conference* (November 1995), Morgan Kaufmann.

[2] D. Klein, J. Smarr, H. Nguyen and C. Manning: Named Entity Recognition with Character-Level Models. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003.

[3] K. Saito, M. Nagata: Multi-Language Named Entity Recognition System based on HMM. NTT Corporation, Japan.

[4] Robert Malouf: Markov models for language-independent named entity recognition. In: *Proceedings of CoNLL-2002*, Taipei, Taiwan, 2002, pp. 187-190.

[5] A. Borthwick: A Maximum Entropy Approach to Named Entity Recognition. Ph.D. New York University. Department of Computer Science, Courant Institute, 1999.

[6] R. Florian, A. Ittycheriah, H. Jing and T. Zhang: Named Entity Recognition through Classifier Combination. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003.

[7] H. Chieu, H. Ng: Named Entity Recognition with a Maximum Entropy Approach. Department of Computer Science, National University of Singapore, 2003.

[8] G. Lembersky: Named Entity Recognition in Hebrew & Hebrew Multiword Expressions: Approaches and Recognition Methods. Department of Computer Science, Ben Gurion University, 2003.

[9] R. Florian: Named Entity Recognition as a House of Cards: Classifier Staking. Center for Language and Speech Processing John Hokins University, 2002.

[10] D. Klein, C. Manning: Maxent Models, Conditional Estimation, and Optimization. HLT-NAACL 2003 and ACL Tutorial.

[11] Tong Zhang and David Johnson: A Robust Risk Minimization based Named Entity Recognition System. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 204-207.

[12] Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., Weischedel, R., and the annotation group: Algorithms that learn to extract information – BBN: Description of the SIFT system as used for MUC-7. *In proceedings of the Seventh Message Understanding Conference (MUC-7) (April 1998)*

[13] Della Pietra, Stephen A., Vincent J. Della Pietra and Adam L. Berger "A Maximum Entropy Approach To Natural Language Processing", *IBM T. J. Watson Research Center*, 1996

[14] Darroch. J., and Ratcliff, D., "Generalized iterative scaling for log-linear models." *The Annals of Mathematical Statistics 43* (1972) 1470-1480

[15] Ristad, E.S. Maximum Entropy Modeling Toolkit (MEMT), release 1.6 beta, February 1998. Includes documentation, which has an overview of MaxEnt modeling.

[16] OpenNLP Project Homepage: http://opennlp.sourceforge.net

[17] OpenNLP MaxEnt Project Homepage: http://maxent.sourceforge.net

[18] Nancy Chinchor, Erica Brown, Lisa Ferro and Patty Robinson, *1999 Named Entity Recognition Task Definition*, MITRE, 1999.

[19] wordfreak Project Homepage: http://wordfreak.sourceforge.net

[20] Jean Carletta, "Assessing Agreement on Classification Tasks: The Kappa Statistics", 2004.

[21] C. Manning, H. Schutze: Foundations of Statistical Natural Language Processing. The MIT Press. Cambridge, MA. 1999.

[22] Adler M., A Hebrew morphological disambiguator based on an unsupervised morpheme-based stochastic model, ISCOL 2005, Haifa.

[23] Language-Independent Named Entity Recognition II (CoNLL2003) Homepage: http://www.cnts.ua.ac.be/conll2003/ner/#CN03

[24] Adwait Ratnaparkhi. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. *Ph.D. thesis*, University of Pennsylvania.

[25] Adwait Ratnaparkhi: A Maximum Entropy Model for Part-Of-Speech Tagging, 1996

[26] מנחם אדלר, מודל מרקוב לבעיית התיוג של טקסט בעברית, תזת תואר שני אוניברסיטת בן גוריון 2001

Ben-Gurion University

Faculty of Natural Sciences

Department of Computer Sciences

Hebrew Named Entity Recognition

Thesis submitted as part of the requirements for the M.Sc.

degree of Ben-Gurion University of the Negev

By

Naama Ben Mordecai

September 2005

Hebrew Named Entity Recognition

This thesis is submitted as part of the requirements for the M.Sc. degree

Written by Naama Ben - Mordecai

Advisor Dr. Michael Elhadad

Department Computer Science

Faculty Natural Sciences

Ben-Gurion University of the Negev

Abstract

Named entity recognition is a computational linguistics task in which we seek to classify every word in a document into one of eight categories: person, location, organization, date, percentage, monetary value and "other". In the taxonomy of computational linguistics, it falls under the domain of "information extraction". Named Entity recognition is a foundation for work on more complex information extraction tasks.

There has been a considerable amount of work on named entity taggers in recent years in many languages. High accuracy results have been achieved for the English language.

In non-English languages, reported performance is significantly lower than in English. The problem in English is greatly simplified by the fact that most of the named entities start with a capital letter.

The Hebrew language presents a lot of morphological ambiguity, which makes automatic processing difficult. There are many other qualities, unique to Hebrew language and culture that might influence the named entity recognition problem. For example: "smichut", agglutination. In addition to these difficulties Hebrew doesn't have the advantage of using capital letters.

Much work has been done on named entity systems for the English language. Different researches have shown different approaches to this problem. Automatic approaches (such as Hidden Markov models, Maximum Entropy models) have the advantage of being dynamic and their development doesn't require much professional linguistic knowledge.

The Maximum Entropy probabilistic modeling technique has proved to be the most powerful one, which can handle large amount of statistical information. Such a system constructs a statistic-probabilistic model that is able to evaluate the likelihood of every word to be in one of mentioned above categories. The system estimates probabilities based on the principle of making as few assumptions as possible, other than constrains imposed. Such constraints are derived from training data, expressing some relationship between features and outcome. We look for the probability distribution with the highest entropy. After creating a model, a dynamic search algorithm is used to find the most probable sequence of outcomes.

The first step in approaching the Named Entity Recognition problem in Hebrew was to define the tagging task for Hebrew. The task definition has to be as unambiguous as possible. Tagging guidelines where phrased and we preformed some agreements tests among human taggers to make sure that the task is well defined.

In this research we present three models for approaching the Named Entity Recognition problem in Hebrew: Maximum Entropy model, Hidden Markov Model and a simple model based on regular expressions and a lexicon extracted from the training data.

This research focused on finding features that are suitable for this problem and for the Hebrew language. We found out that local features (the current word and 1-2 words around it) are the most significant. Important features are dictionaries and grammatical and morphological features. The best results are achieved by combining a large set of features. The training corpus has also a great impact on the performance of the system.

The Maximum Entropy model has shown the best results out of the three models. However, best results are archived when combining the three models.

The combined system has archived good results, the best results achieved so far for Hebrew.

Table of Contents

1. **Background and Introduction** 10

1.1 Named Entity Recognition 10

1.2 Named Entity Recognition Application 11

1.3 Uniqurness of the Hebrew Language 12

1.4 Named Entity Recognition System Evaluation 14

2. **Prior Work** 16

2.1 The Handcrafted Approach 16

2.2 Automatic Approach: Hidden Markov Models 18

2.2.1 Hidden Markov Models 18

2.2.2 Named Entity Recognition System Based on HMM 19

2.2.3 Character Level HMM 21

2.3 Automatic Approach: Maximum Entropy Model 23

2.3.1 Maximum Entropy 23

2.3.2 Model Constrains 24

2.3.3 GIS Algorithm for Building a Maximum Entropy Model 25

2.3.4 How to Use a Model 26

2.3.5 Named Entity Recognition System Based on a Maximum Entropy Model 26

2.4 Automatic Approach: Rubust Risk Minimization 29

2.5 Classifier Combination 31

3. **Agreement** 33

3.1 Named Entity Recognition Task Definition 33

3.2 Agreement Tests 33

3.2.1 The "Text" Test 33

3.2.2 The "Type" Test 34

3.2.3 The Combined Test 34

3.3 Agreement Estimation 35

3.3.1 Partial Agreement 35

3.3.2 Kappa Statistics 35

3.4 The Experiment 37

3.5 Results and Conclusions 38
3.5.1 The t Test 38
3.5.2 Results of the two first stages 38
3.5.3 Results of the last stage 41
4. **The Corpus** 43
5. **Baseline** 45
6. **Named Entity Recognition via Maximum Entropy Model** 47
6.1 The MaxEnt Package 47
6.2 Beam Search 49
6.3 Features 51
6.3.1 Structural 51
6.3.2 Lexicon 51
6.3.3 Previous Tags 51
6.3.4 Dictionary 51
6.3.5 Local Regular Expressions 52
6.3.6 Long Regular Expressions 53
6.3.7 Part of Speech Tagger and Morphological Disambiguator 54
6.3.8 Preprocessing 54
6.3.9 Features product 55
6.4 Results 56
6.4.1 Part of Speech 58
6.4.2 Dictionary 59
6.4.3 Lemma 60
6.4.4 Preprocessing 61
6.4.5 Corpus 62
6.4.6 Results Summery 64
7. **Named Entity Recognition via HMM** 65
7.1 The Basic Model 65
7.2 Part of Speech Tags as Model Nodes 66
7.3 HMM Combining Features 67
7.4 Comparison to the Maximum Entropy Model 70
8. Classifier Combination 72
9. **Comparisons** 74

9.1 Comparison to Gennady Lambersky's Work 74
9.2 Comparison to English Systems 76
9.3 Comparison to System for Other Languages 77
10. **Summery and Conclusions** 79
11. **Suggested Future Work** 80
12. **Appendix A: Tagging guidelines** 81
13. **Appendix B: Tagged Article Sample** 96
14. **Reference** 100