איך להגות שמות עבריים
How to pronounce Hebrew names

אלון איתי וגיא שקד

הפקולטה למדעי המחשב, טכניון

Israeli Seminar on Computational Linguistics 2010
16/6/2010

מוטיבציה

ניקוד של טקסט עברי הוא השלב הראשון בדרך להגייה. כשמדובר בטקסט כללי ניתן להשתמש במנתח מורפולוגי או במילון, אך לא כשמדובר בשמות.

מוטיבציה

ניקוד של טקסט עברי הוא השלב הראשון בדרך להגייה. כשמדובר בטקסט כללי ניתן להשתמש במנתח מורפולוגי או במילון, אך לא כשמדובר בשמות.

- שמות פשוטים - גיא, שקד, אלון, איתי

אופן ההגייה ברור לנו כדוברי השפה.

מוטיבציה

ניקוד של טקסט עברי הוא השלב הראשון בדרך להגייה. כשמדובר בטקסט כללי ניתן להשתמש במנתח מורפולוגי או במילון, אך לא כשמדובר בשמות.

- שמות פשוטים - גיא, שקד, אלון, איתי

אופן ההגייה ברור לנו כדוברי השפה.

- שמות מורכבים - אייאפיאטלאייקוטל, אצ'נפה, אלמאצ'אש

מציבים אתגר משמעותי גם בפני בני אדם.

בעיות מיוחדות בעברית

- שמות ממגוון גדול של מקורות - שפות, תרבויות, ארצות מוצא.

בעיות מיוחדות בעברית

- שמות ממגוון גדול של מקורות - שפות, תרבויות, ארצות מוצא.
- בטקסט חסר ניקוד - אין תנועות.

בעיות מיוחדות בעברית

- שמות ממגוון גדול של מקורות - שפות, תרבויות, ארצות מוצא.
- בטקסט חסר ניקוד - אין תנועות.
- א, ו, י - אמות קריאה או עיצורים?

בעיות מיוחדות בעברית

- שמות ממגוון גדול של מקורות - שפות, תרבויות, ארצות מוצא.
- בטקסט חסר ניקוד - אין תנועות.
- א, ו, י - אמות קריאה או עיצורים?
- ב, כ, פ - יש או אין דגש?

בעיות מיוחדות בעברית

- שמות ממגוון גדול של מקורות - שפות, תרבויות, ארצות מוצא.
- בטקסט חסר ניקוד - אין תנועות.
- א, ו, י - אמות קריאה או עיצורים?
- ב, כ, פ - יש או אין דגש?
- שׂין או שׁין?

הגדרת הבעיה

בעיה

בהינתן שם בעברית לא מנוקדת -

- להוסיף לכל אות סימן ניקוד - בָּ, בֶּ, בִּ, בּוֹ, בּוּ או בְּ - לציון התנועות.
- לכל מופע של ב, פ, כ - להכריע האם הן דגושות או לא.
- לכל מופע של האות שי"ן - להכריע האם מדובר בשׁין ימנית או שׂמאלית.

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית

- לכל שפה שיטת כתיבה שונה.
- לרוב, הגיית שמות זהה או דומה בשפות שונות.
- בא"ב הלטיני -
  - תנועות נכתבות כאותיות.
  - לרוב העיצורים יש אות נפרדת - למשל, ב ו-בּ ייכתבו ע"י אותיות שונות.

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית

- לכל שפה שיטת כתיבה שונה.
- לרוב, הגיית שמות זהה או דומה בשפות שונות.
- בא"ב הלטיני -
  - תנועות נכתבות כאותיות.
  - לרוב העיצורים יש אות נפרדת - למשל, ב ו-בּ ייכתבו ע"י אותיות שונות.

דוגמא

איך הוגים "אביתר"? האם ה-ב' דגושה? האם ה-י' עיצורית או אם קריאה? אילו תנועות מתאימות?

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית

- לכל שפה שיטת כתיבה שונה.
- לרוב, הגיית שמות זהה או דומה בשפות שונות.
- בא"ב הלטיני -
  - תנועות נכתבות כאותיות.
  - לרוב העיצורים יש אות נפרדת - למשל, ב ו-בּ ייכתבו ע"י אותיות שונות.

דוגמא

איך הוגים "אביתר"? האם ה-ב' דגושה? האם ה-י' עיצורית או אם קריאה? אילו תנועות מתאימות?

- Evyatar - הב' לא דגושה.

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית

- לכל שפה שיטת כתיבה שונה.
- לרוב, הגיית שמות זהה או דומה בשפות שונות.
- בא"ב הלטיני -
  - תנועות נכתבות כאותיות.
  - לרוב העיצורים יש אות נפרדת - למשל, ב ו-בּ ייכתבו ע"י אותיות שונות.

דוגמא

איך הוגים "אביתר"? האם ה-ב' דגושה? האם ה-י' עיצורית או אם קריאה? אילו תנועות מתאימות?

- Evyatar - הב' לא דגושה.
- Evyatar - ה-י' עיצורית.

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית

- לכל שפה שיטת כתיבה שונה.
- לרוב, הגיית שמות זהה או דומה בשפות שונות.
- בא"ב הלטיני -
  - תנועות נכתבות כאותיות.
  - לרוב העיצורים יש אות נפרדת - למשל, ב ו-בּ ייכתבו ע"י אותיות שונות.

דוגמא

איך הוגים "אביתר"? האם ה-ב' דגושה? האם ה-י' עיצורית או אם קריאה? אילו תנועות מתאימות?

- Evyatar - הב' לא דגושה.
- Evyatar - ה-י' עיצורית.
- Evyatar - אֶבְיָתָר.

שימוש בטרנסקריפציה באנגלית

האלגוריתם

הקלט - זוגות של שמות, בעברית ואנגלית.

שימוש בטרנסקריפציה באנגלית
האלגוריתם

הקלט - זוגות של שמות, בעברית ואנגלית.

- נתאים בין העיצורים בשם העברי לעיצורים בטרנסקריפציה האנגלית.

שימוש בטרנסקריפציה באנגלית

האלגוריתם

הקלט - זוגות של שמות, בעברית ואנגלית.

- נתאים בין העיצורים בשם העברי לעיצורים בטרנסקריפציה האנגלית.
- נסיק את אופן הגיית האותיות ב, כ, פ ו-ש ע"פ הטרנסקריפציה האנגלית.

שימוש בטרנסקריפציה באנגלית

האלגוריתם

הקלט - זוגות של שמות, בעברית ואנגלית.

- נתאים בין העיצורים בשם העברי לעיצורים בטרנסקריפציה האנגלית.
- נסיק את אופן הגיית האותיות ב, כ, פ ו-ש ע"פ הטרנסקריפציה האנגלית.
- נוסיף את שאר התנועות ונזהה אמות קריאה.

שימוש בטרנסקריפציה באנגלית
האלגוריתם

הקלט - זוגות של שמות, בעברית ואנגלית.

- נתאים בין העיצורים בשם העברי לעיצורים בטרנסקריפציה האנגלית.
- נסיק את אופן הגיית האותיות ב, כ, פ ו-ש ע"פ הטרנסקריפציה האנגלית.
- נוסיף את שאר התנועות ונזהה אמות קריאה.

דוגמא

אבא	אבן
Aba	Even
⇓	⇓
אבּא	אבן
⇓	⇓
אָבָּא	אֶבֶן

שימוש בטרנסקריפציה באנגלית

ניתן לקבל זוגות של שמות בעברית ובאנגלית ממספר מקורות, למשל -

- ויקיפדיה
- רשימות סטודנטים

שימוש בטרנסקריפציה באנגלית

ניתן לקבל זוגות של שמות בעברית ובאנגלית ממספר מקורות, למשל -

- ויקיפדיה
- רשימות סטודנטים

בעיות בשיטה -

- שמות שהגייתם בעברית ובאנגלית שונה - אברהם / Abraham
- מוגבל לשמות שמופיעים ברשימות

אלגוריתם 2: תבניות נפוצות

שמות ממקורות דומים חולקים בינהם תבניות דומות בכתיבה ובהגייה, בפרט - רישא וסיפא.

אלגוריתם 2: תבניות נפוצות

שמות ממקורות דומים חולקים בינהם תבניות דומות בכתיבה ובהגייה, בפרט - רישא וסיפא.

דוגמא

גולדברג

רישא - גולדמן, גולדשטיין
סיפא - גוטנברג, חיימברג

אלגוריתם 2: תבניות נפוצות

שמות ממקורות דומים חולקים בינהם תבניות דומות בכתיבה ובהגייה, בפרט - רישא וסיפא.

דוגמא

גולדברג

רישא - **גולד**מן, **גולד**שטיין
סיפא - גוטנברג, חיימברג

אלגוריתם 2: תבניות נפוצות

שמות ממקורות דומים חולקים בינהם תבניות דומות בכתיבה ובהגייה, בפרט - רישא וסיפא.

דוגמא

גולד**ברג**

רישא - גולדמן, גולדשטיין
סיפא - גוטנ**ברג**, חיימ**ברג**

אלגוריתם 2: תבניות נפוצות

שמות ממקורות דומים חולקים בינהם תבניות דומות בכתיבה ובהגייה, בפרט - רישא וסיפא.

דוגמא

גולדברג

רישא - גולדמן, גולדשטיין
סיפא - גוטנברג, חיימברג

רישאות

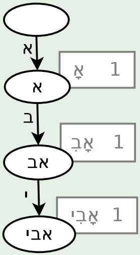
בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

רישאות

בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

דוגמא

אבי

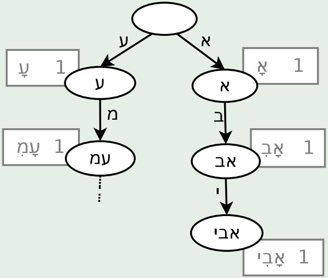


רישאות

בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

דוגמא

אבי
עמי

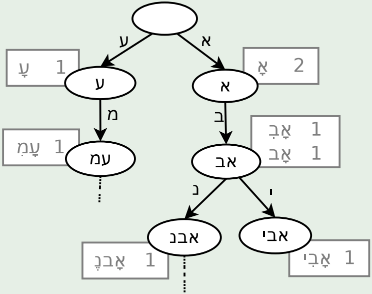


רישאות

בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

דוגמא

אבי
עמי
אבנר

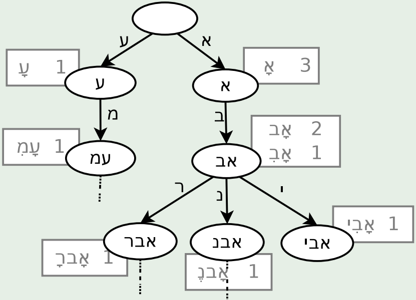


רישאות

בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

דוגמא

אבי
עמי
אבנר
אברהם

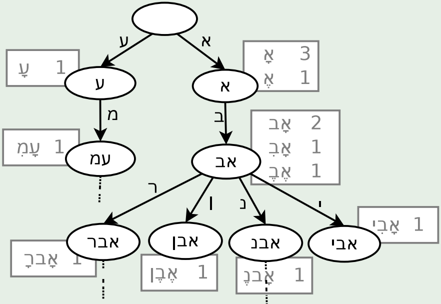


רישאות

בהינתן רשימה של שמות שאופן ההגייה שלהם ידוע - ניצור Trie ובו לכל צומת נמנה את מספר המופעים של כל הגייה.

דוגמא

אבי
עמי
אבנר
אברהם
אבן



רישאות וסיפאות
האלגוריתם

בדומה לרישאות, נבנה Trie של סיפאות.

רישאות וסיפאות

האלגוריתם

בדומה לרישאות, נבנה Trie של סיפאות.

בהינתן שם שהגייתו לא ידועה, נמצא את הרישא והסיפא הארוכות ביותר המופיעות במבני הנתונים -

רישאות וסיפאות
האלגוריתם

בדומה לרישאות, נבנה Trie של סיפאות.

בהינתן שם שהגייתו לא ידועה, נמצא את הרישא והסיפא הארוכות ביותר המופיעות במבני הנתונים -

- הגיית הרישא והסיפא תבחר לפי ההגייה הנפוצה ביותר
  גוֹלדברג + גולד**בֶּרג** ← גוֹלדבֶּרג

רישאות וסיפאות
האלגוריתם

בדומה לרישאות, נבנה Trie של סיפאות.

בהינתן שם שהגייתו לא ידועה, נמצא את הרישא והסיפא הארוכות ביותר המופיעות במבני הנתונים -

- הגיית הרישא והסיפא תבחר לפי ההגייה הנפוצה ביותר
  גוֹלדברג + גולד**בֶּרג** ← גוֹלדבֶּרג
- במקרה של חפיפה בין הרישא והסיפא - החפיפה תתחשב בשתיהן
  אָהָרוֹני + אה**רוֹנִי** ← אָהָרוֹנִי

רישאות וסיפאות
האלגוריתם

בדומה לרישאות, נבנה Trie של סיפאות.

בהינתן שם שהגייתו לא ידועה, נמצא את הרישא והסיפא הארוכות ביותר המופיעות במבני הנתונים -

- הגיית הרישא והסיפא תבחר לפי ההגייה הנפוצה ביותר
  גוֹלדברג + גולד**בֶּרג** ← גוֹלדבֶּרג
- במקרה של חפיפה בין הרישא והסיפא - החפיפה תתחשב בשתיהן
  אָהָרוֹני + אה**רוֹנִי** ← אָהָרוֹנִי
- במקרה של פער בין הרישא לסיפא (נדיר) - הפעלה רקורסיבית
  אָיֶיאפיאטלא**ייקוּטָל** ←
  אֶפִּיאטלָ**אִי** ...

תיקונים

במקרים מסויימים יש לבצע התאמות של הפלט -

תיקונים

במקרים מסויימים יש לבצע התאמות של הפלט -

- ו"ו עם חולם אחרי חולם חסר
  יָעֲקֹובִּי ← יָעֲקוֹבִּי

תיקונים

במקרים מסויימים יש לבצע התאמות של הפלט -

- ו"ו עם חולם אחרי חולם חסר
  יַעֲקֹובִּי ← יַעֲקֹובִּי
- יו"ד בסוף שם, ולפניה עיצור ללא תנועה
  כַּרְמֶלִי ← כַּרְמֶלְִי

תיקונים

במקרים מסויימים יש לבצע התאמות של הפלט -

- ו"ו עם חולם אחרי חולם חסר
  יָעֲקֹובִּי ← יָעֲקוֹבִּי
- יו"ד בסוף שם, ולפניה עיצור ללא תנועה
  כָּרמֶלִי ← כָּרמֶלְי
- תנועות אִ, אָ, אֶ בעיצור אחרון
  שִׁירָ ←שִׁיר

שילוב של שני האלגוריתמים

שני האלגוריתמים משלימים זה את זה -

שילוב של שני האלגוריתמים

שני האלגוריתמים משלימים זה את זה -

- שמות באנגלית מאפשרים למצוא את אופן ההגייה עבור רשימה גדולה של שמות, אך האלגוריתם מוגבל לשמות שמופיעים באותה רשימה.

שילוב של שני האלגוריתמים

שני האלגוריתמים משלימים זה את זה -

- שמות באנגלית מאפשרים למצוא את אופן ההגייה עבור רשימה גדולה של שמות, אך האלגוריתם מוגבל לשמות שמופיעים באותה רשימה.
- לאלגוריתם הרישאות והסיפאות דרושה רשימה ראשונית, ממנה הוא יכול לנחש את ההגייה של שמות נוספים.

שילוב של שני האלגוריתמים

שני האלגוריתמים משלימים זה את זה -

- שמות באנגלית מאפשרים למצוא את אופן ההגייה עבור רשימה גדולה של שמות, אך האלגוריתם מוגבל לשמות שמופיעים באותה רשימה.
- לאלגוריתם הרישאות והסיפאות דרושה רשימה ראשונית, ממנה הוא יכול לנחש את ההגייה של שמות נוספים.

אנו מפעילים את אלגוריתם 1 על רשימה דו לשונית של שמות על מנת למצוא את ההגייה הנכונה שלהם, ואז מזינים את אלגוריתם 2 ברשימה הזו על מנת למצוא את אופן ההגייה של שמות שלא הופיעו ברשימות המקוריות.

כלי הערכה

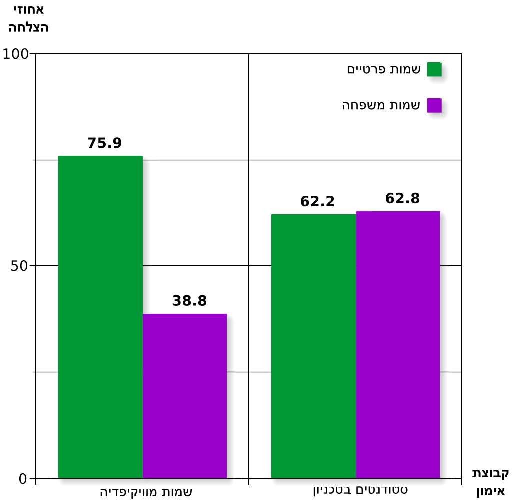
על מנת לבחון את ביצועי האלגוריתמים, בחרנו באקראי 1000 רשומות מתוך ספר טלפונים, הפרדנו את השמות המופיעים בהם לשמות פרטיים ושמות משפחה והגדרנו ידנית את אופן ההגייה הנכון לכל שם.

בתור קבוצת אימון (Training set) השתמשנו במספר מקורות -

- שמות אקראיים (אחרים) מספר טלפונים - תוייגו ידנית
- שמות אישים מוויקיפדיה - בעברית ובאנגלית
- שמות משפחה של סטודנטים בטכניון - בעברית ובאנגלית

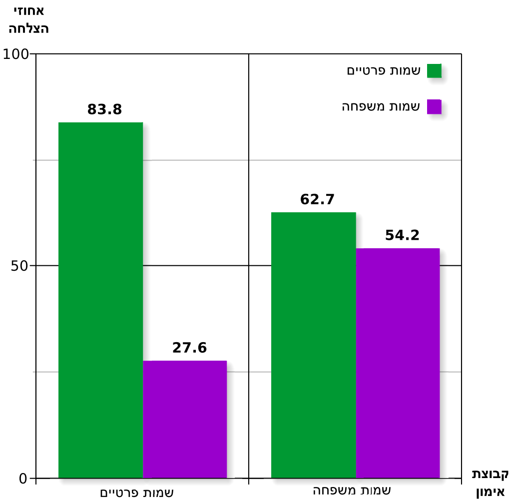
תוצאות

אלגוריתם 1: שימוש בטרנסקריפציה באנגלית



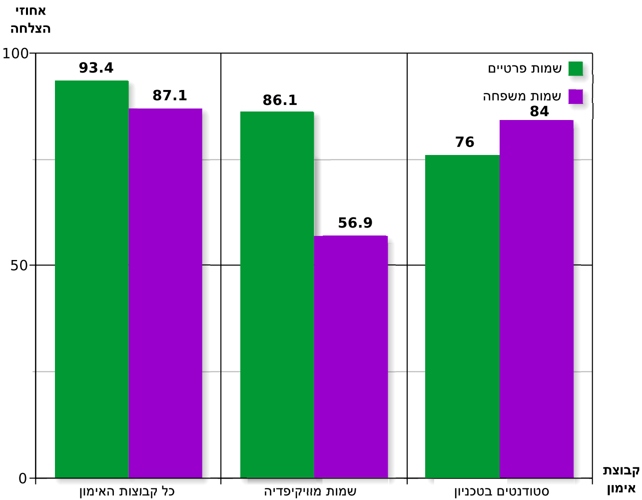
תוצאות

אלגוריתם 2: רישא-סיפא



תוצאות

אלגוריתם משולב



סיכום ומסקנות

- ניתן להגיע לתוצאות משמעותיות בשילוב שני האלגוריתמים.

סיכום ומסקנות

- ניתן להגיע לתוצאות משמעותיות בשילוב שני האלגוריתמים.
- המגוון והגודל של קבוצת האימון משפיעים בצורה משמעותית על הביצועים.

סיכום ומסקנות

- ניתן להגיע לתוצאות משמעותיות בשילוב שני האלגוריתמים.
- המגוון והגודל של קבוצת האימון משפיעים בצורה משמעותית על הביצועים.
- מבט לעתיד -

סיכום ומסקנות

- ניתן להגיע לתוצאות משמעותיות בשילוב שני האלגוריתמים.
- המגוון והגודל של קבוצת האימון משפיעים בצורה משמעותית על הביצועים.
- מבט לעתיד -
  - מיקום הטעם.

סיכום ומסקנות

- ניתן להגיע לתוצאות משמעותיות בשילוב שני האלגוריתמים.
- המגוון והגודל של קבוצת האימון משפיעים בצורה משמעותית על הביצועים.
- מבט לעתיד -
  - מיקום הטעם.
  - שימוש בשפות נוספות.