נושאים מתקדמים במדעי המחשב 11

(גישות מבוססות קורפוסים לעיבוד שפות טבעיות)

האם שימוש בשפה שלישית מועיל לפתרון בעיית ה-word alignment (ברמת התמניות)?

מגישים:

בני אפלבאום 038584348

יאן ציטרין 312009632

1. הקדמה ורקע

התפתחות רשת האינטרנט העניקה לנו אוצר של טכסטים אלקטרוניים בכמויות שהיו דמיוניות בעבר, ובפרט גם מספר רב של טקסטים מקבילים, דהיינו טקסטים המופיעים במספר שפות שונות. בעקבות זאת, נראה כי ניתן לנצל את המאגרים הללו כדי לאפיין את הקשר בין שפות שונות, ובפרט לייצר (או לשפר) באופן אוטומטי מילון דו-לשוני ע"י מציאת ההתאמות בין הטכסטים ברמת המילה.

בעיה זו, הידועה בשם word alignment, זכתה להתייחסות נרחבת במחצית הראשונה של שנות התשעים [fung-church94], [fung-keown94], [smadja94] והפכה יחד עם שאר הגרסאות של בעיית ה- alignment לאחד הנושאים המרכזיים בעיבוד שפה טבעית.

<u>הגדרה:</u> - ניתן לנסח את בעיית ה-*word alignment* באופן הבא:
<u>קלט</u>: טקסט T_a בשפה A, ותרגומו - טקסט T_b לשפה B.
<u>פלט</u>: לקסיקון דו לשוני מ- A ל- B (תרגום של תת קבוצה של מילות T_a למילות של T_b).

חשוב להפריד בין alignment של תמניות (token level) כפי שהן מופיעות בטכסט לבין alignment של הכניסות המילוניות עבור התמניות הללו (type level), אשר לוקחת בחשבון את השינויים הצורניים [ראה, melamed2000].

החוקרים בתחום מציינים כי לא ניתן להגדיר את איכות הפלט במדויק בגלל שיקול הדעת הסובייקטיבי של השופט האנושי, הבא לידי ביטוי בכך שהקורלציה בין השיפוטים של מומחים שונים נמוכה יחסית. אך בכל זאת, היא ניתנת להערכה במונחים של precision ו-recall [turian2003]. ההגדרה המדויקת של המונחים האלה עבור הבעיה שלנו תבוא בהמשך.

חלק מן הגישות שנוסו בעבר השתמשו בידע בלשני מוקדם על הקשר בין השפות (ראה למשל [church93] או [dagan93]) וכך, התאימו בין מילים על סמך רצפים דומים של אותיות, כך לדוגמא בצמד *government* ו-*gouvernement*. גישות אלה מוגבלות מטבען לשפות קרובות ובוודאי שאינן ישימות למעבר בין שפות בעלות אלף-בית שונה כמו אנגלית-עברית, או צרפתית-יפנית. במקרים כאלה נראה כי כדאי להתייחס לטכסטים השונים כאל רצפים סטטיסטים, ולהצמיד מילה לתרגומה על סמך דפוס הופעות סטטיסטי דומה, תחת ההנחה שמילה ותרגומה יופיעו בערך באותה תדירות ובאותם מקומות בטכסטים המקבילים. גישה כזו תעניק לכל זוג מלים ציון התאמה המבטא בדרך כלשהי את מידת הדמיון בין המופעים של שתי המילים, ותבנה מילון על סמך הזוגות שזכו לציון גבוה.

שאלת הניסוי

ציון הדמיון שמחושב ע"י אלגוריתם הדוגל בגישה הנ"ל מושפע מרעשים שונים:

- חוסר עקביות המתרגם,
- מבנה תחבירי ומורפולוגי שונה של שתי השפות,
- יחסים מורכבים בתרגום של מילים כמו one to many או many to one, וכו'.

יתכן, אם כן, שניתן לשפר אותו באמצעות ידע סטטיסטי נוסף הנגזר משימוש בשפה שלישית.[1]
כלומר, נרצה לבדוק האם ניתן לשפר את תוצאות ה- word alignment מ-A ל-B בעזרת שימוש בשפה שלישית C. כך למשל, עבור מילה a בשפה A נכריע בין כמה תרגומים מתחרים בשפה B b_1, b_2 שציון הדמיון שלהם ל-a דומה, באמצעות מידת הדמיון שלהם למילה c בשפה C שקרובה למילה a. יש

[1] בבעיות אחרות בעיבוד שפה טבעית כבר הוכח כי שימוש בשפה נוספת משפר את התוצאות. כך למשל ב- [dag-itai-91].

לשים לב כי השיטה שאנו מציעים הינה שקופה לאלגוריתם הסטטיסטי שמחשב את מידת הדמיון בין זוג המילים.

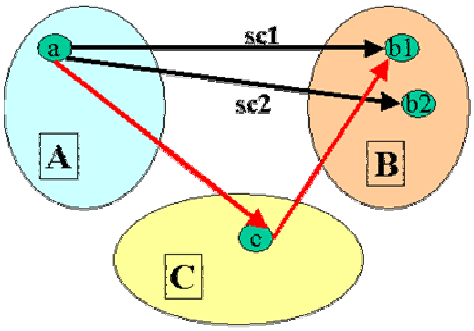

Figure 1

האלגוריתם בו השתמשנו לחישוב מידות הדמיון בין זוגות הוא k-vec (fung-church94).
אך משום שהגישה שלנו אינה תלויה באלגוריתם לא נתאר אותו כאן (לתיאור מפורט של האלגוריתם ראה appendix-k-vec).

שקלול השפה השלישית

<u>הגדרה:</u> בהינתן זוג מילים אנגלי-עברי e,h, נגדיר את ציון ההתאמה שלהן בתיווך המילה הרוסית r כ- $T(e,r,h)=\sqrt{T(e,r)\cdot T(r,h)}$, כאשר $T(e,r),T(r,h)$ הם ציוני של מידת הדמיון האלגוריתם הסטטיסטי מחשב בהרצה על אנגלית-רוסית, ורוסית-עברית בהתאמה.

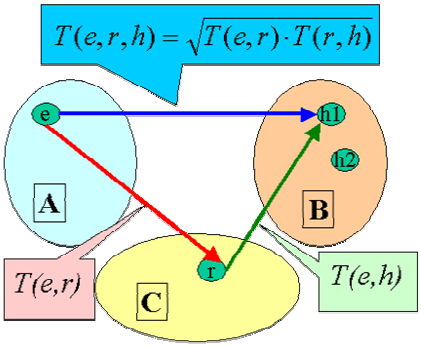

Figure 2

לפי הגדרה זו אם מילה אנגלית ומילה רוסית תואמות מבחינת מופעיהן (בעלות ציון דמיון גבוה), ואותה מילה רוסית תואמת גם למילה עברית, הרי שהמילה האנגלית והמילה העברית זוכות לציון קרבה גבוה.

<u>הגדרה:</u> בהינתן פרמטר β נאמר שמילה רוסית r היא *מתווכת* β-*מהימנה* עבור הזוג r,h אם $T(e_i,r_k,h_j)\geq\beta$.

נשים לב כי לעיתים יתכנו כמה מתווכים β- מהימנים; כך למשל, עבור הצמד <*מלך*, *king*> גם *carja*, וגם *car'* הם מתווכים β-מהימנים. על כן, נתחשב בכל המתווכים ה-β-מהימנים בציון הדמיון המשוקלל.

<u>הגדרה:</u> *ציון ההתאמה המשוקלל* של זוג מילים אנגלי-עברי e,h, הוא:

$$S(e,h)=T(e,h)+\alpha\sum_{r\ \text{s.t.}\ T(e,r,h)>\beta}T(e,r,h)$$

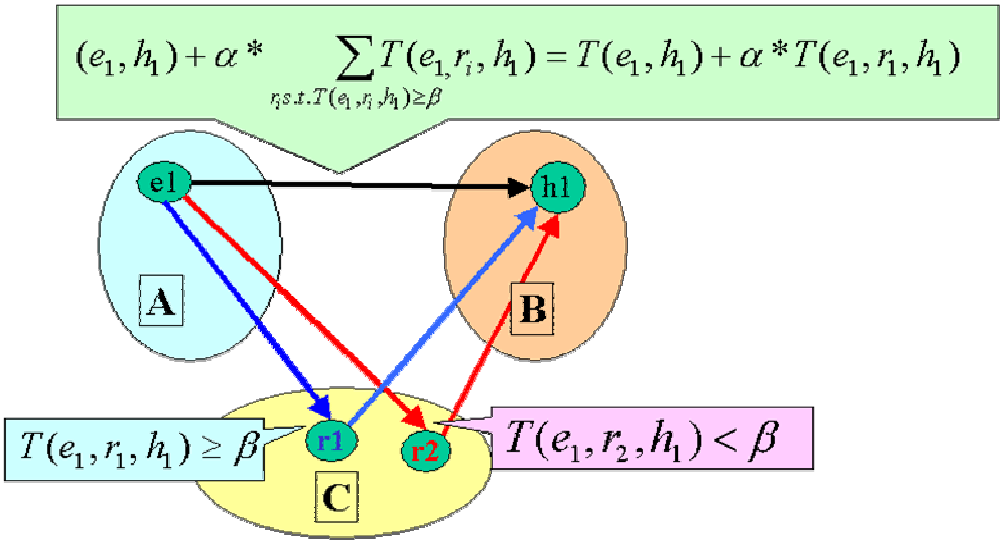

Figure 3

כאשר $T(e,h)$ הוא ציון שהאלגוריתם הסטטיסטי מחשב על אנגלית-עברית.

ציון ההתאמה המשוקלל משקלל את ציון האלגוריתם הסטטיסטי עם ציוני התיווך של המתווכים המהימנים. המשקולת α מאפשרת להעדיף את ציוני התיווך על פני ציון המקורי או להפך.

2. הניסוי

כדי לבדוק האם שימוש בשפה שלישית מועיל, השוונו בין התוצאות שהתקבלו מהרצת אלגוריתם ה--k vec על קורפוס עברי ואנגלי, לבין תוצאות האלגוריתם שלנו המשתמש גם בקורפוס רוסי, אשר מקבילי לשניהם.

תיאור הקורפוס

הקורפוס בו השתמשנו הוא ספר קהלת שגודלו הוא כדלקמן:

שפה	אנגלית	עברית	רוסית
גודל הטכסט (תמניות)	2,991	5,579	4,161

הטקסטים עצמם (אחרי שהורדנו מהם סימני פיסוק, מספרי פרקים וכו') נמצאים בנספחים (eng-final-txt, heb-final-txt, rus-final-txt).

כלי תוכנה בהם השתמשנו

לשם הרצת אלגוריתם ה-k-vec השתמשנו במימוש של ה-kvec ([varma2002]), שהתבצע על ידי Varma Nitin במסגרת עבודת המסטר שלו. בעזרת הכלי שכתב בדק Varma מספר רב של מדדים סטטיסטיים ומצא כי המדד T-score הוא האפקטיבי ביותר לצורך פתרון בעית words allignment. לכן בחרנו להשתמש במדד זה כאל הציון שהאלגוריתם הסטטיסטי נותן למידת הדמיון בין זוג המילים.

הערכת תוצאות האלגוריתם

כפי שמקובל בתחום של מערכות לעיבוד שפות טבעיות ([turian2003], [melamed2003],[white93]) השתמשנו במונחים recall ו- precision כדי להעריך את תוצאות ההרצה של האלגוריתם שלנו.
recall ו- precision חושבו בהשוואה למילון שיצרנו על יד יישור ידני של הטקסט האנגלי מול הטקסט העברי (ראה נספח [gold-standard-txt]).

<u>הגדרה</u>: יהי GST (golden standard translations) אוסף כל התרגומים שקיבלנו באופן ידני.
AT[***A***] (automatic translation of the algorithm ***A***) אוסף כל התרגומים שנותן האלגוריתם ***A***, אזי

$$\text{precision} = |\, GST \cap AT[\boldsymbol{A}] \,| \,/\, AT[\boldsymbol{A}] \;;\quad \text{recall} = |\, GST \cap AT[\boldsymbol{A}] \,| \,/\, GST$$

מהלך הניסוי

1. עיבוד מקדים של הטקסט בשלושת השפות שכלל השמטת סימני פיסוק ומספרים, והשמטת מילים נפוצות מדי ומילים נדירות.
2. הרצת k-vec שלוש פעמים על אנגלית-עברית, אנגלית-רוסית, ורוסית-עברית, וקבלת ציוני T-score לכל זוג מילים כפלט.
3. חישוב ציון ההתאמה המשוקלל ולפיו יצירת המילון "שלנו" (רשימה של זוגות מילים בסדר יורד של ציון ההתאמה המשוקלל).
4. יצירת המילון ה"בסיסי" על סמך ציוני ה-T-score מאנגלית לעברית (רשימה של זוגות מילים בסדר יורד של ציוני ה- T-score).
5. חישוב ה-precision וה-recall של המילון "שלנו" ושל המילון ה"בסיסי" בהשוואה למילון הידני.

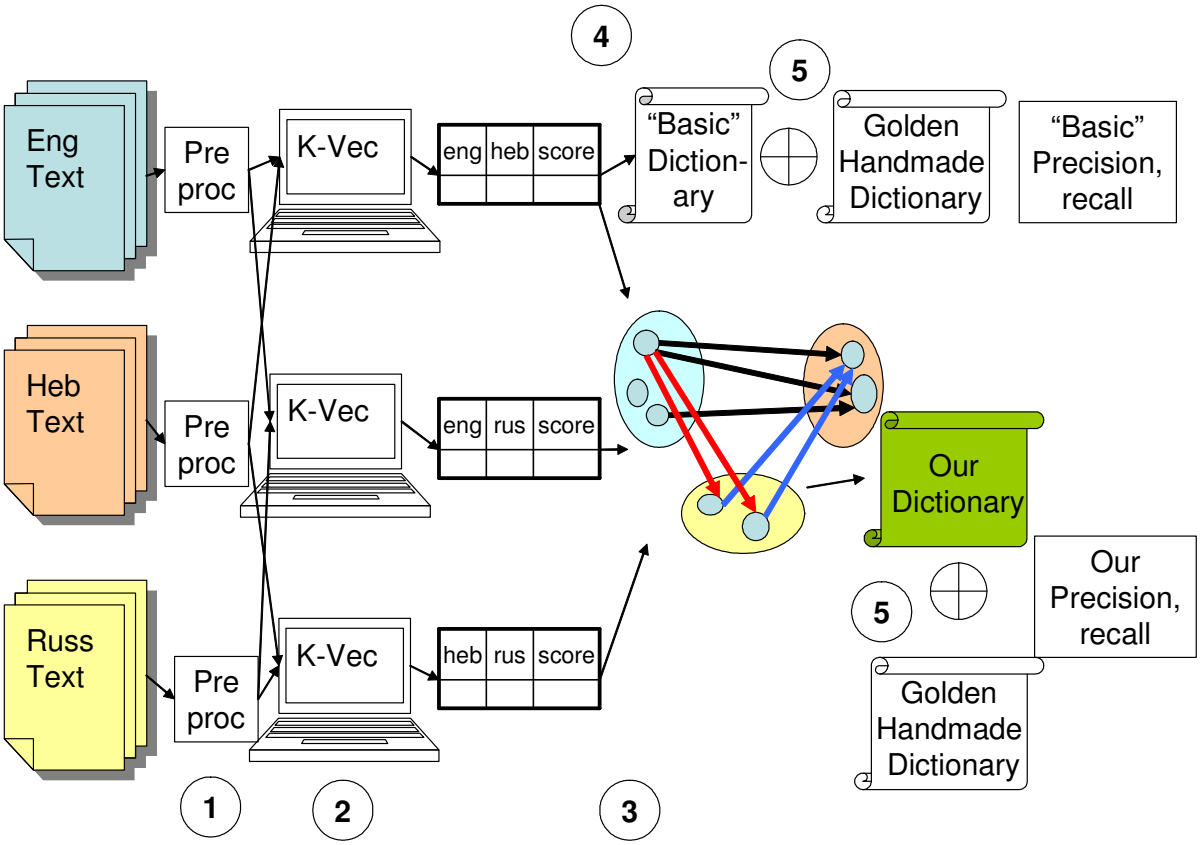

Figure 4

3. תוצאות וניתוח

כזכור הן האלגוריתם "שלנו" והן k-vec מחזירים כפלט רשימת זוגות המדורגת לפי מהימנות יורדת. כיוון שהאלגוריתמים נותנים ציון לכל הזוגות האפשריים ברור שיש "לחתוך" את הרשימה בנקודה כלשהי ולבחון רק את הזוגות שזכו למהימנות גבוהה יחסית.

בדומה למדיניות ההערכה של [varma2002] הסתכלנו על התלות של איכות התוצאות (במונחים של recall ו-precision) בחלק היחסי (באחוזים) של כל התרגומים שהחזירו האלגוריתמים (הk-vec הבסיסי והאלגוריתם שלנו. כלומר, השוונו בין המילונים הכוללים את האחוזון העליון של התוצאות, בין אלה הכוללים את שני האחוזונים העליונים, וכן הלאה במרווחים של 1%. הגרפים שלהלן המתארים את התוצאות שקיבלנו:

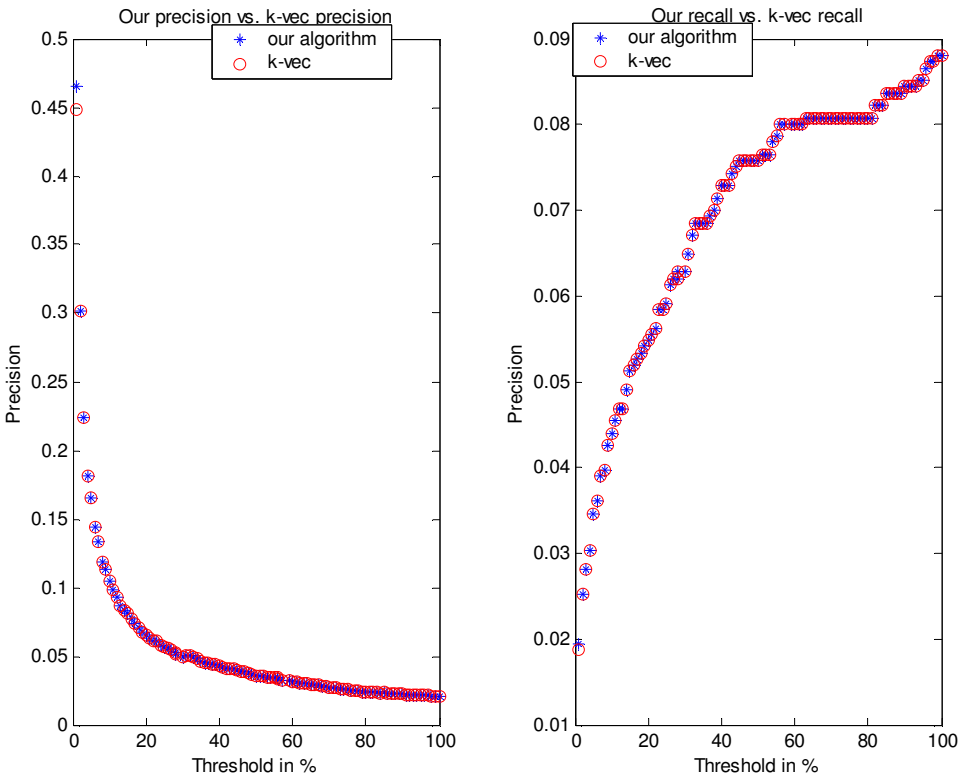

Figure 5

כפי שניתן לראות התוצאות כמעט זהות, פרט ליתרון זעום של האלגוריתם המשוקלל באחוזון העליון. למעשה מדובר בסה"כ בפער של זוג אחד ולכן אין בו כדי להצביע על פער סטטיסטי מובהק. אם כן, בסיכומו של דבר לא הצלחנו לשפר את תוצאות האלגוריתם המקורי באמצעות שימוש בשפה נוספת.

את ערכי α,β שהביאו לתוצאות הנ"ל כוונו באמצעות הרצות חוזרות של הניסוי עם פרמטרים שונים. במהלך בדיקות אלה גילינו כי ה- α אינה משנה מאד את התוצאות לעומת ה- β המשפיעה באופן משמעותי כפי שניתן לראות בטבלה המצורפת (ראה Figure 6).

The recall and precision for the best 1% of the results as a function of β (α =0.5)

β	PRECISION	RECALL
5.00	0.448275862	0.018759019
4.50	0.448275862	0.018759019
4.00	0.448275862	0.018759019
3.50	0.448275862	0.018759019
3.00	0.448275862	0.018759019
2.50	0.448275862	0.018759019
2.00	0.448275862	0.018759019
1.50	0.327586207	0.013708514
1.00	0.120689655	0.005050505
0.50	0.086206897	0.003607504
0.40	0.086206897	0.003607504
0.20	0.086206897	0.003607504
0.10	0.086206897	0.003607504
0.01	0.086206897	0.003607504

Figure 6

ניתוח התוצאות

המדד החדש שהצענו המשקלל את השפה השלישית אינו משפר את התוצאות, את האשמה ניתן לתלות באחת מן הסיבות הבאות:

א. לא "נתנו" מספיק משקל לרכיב שמייצג את השפה השלישית ולכן הוא לא התבטא באופן משמעותי.

ב. הדרך בה נעזרנו בשפה השלישית אינה מוסיפה אינפורמציה חדשה, ולכן הציון המשוקלל דומה למקורי.

בחינה מדוקדקת של התוצאות שוללת את האפשרות הראשונה, שכן, כאמור, הגדלת ה- α (המשקולת של ציון המתווכים) לא שיפרה את התוצאות. נשארנו, אם כן, עם האפשרות הראשונה, הזוכה לתמיכה של הנתונים שכן מסתבר שיש התאמה בין תרגומים בעלי מתווכים מהימנים לתרגומים בעלי ציוני T-score טובים, ולכן הציון המשוקלל אינו שונה בהרבה מן הציון המקורי.

מעבר להסבר זה, לניסוי שביצענו מספר מגבלות שחשוב לעמוד עליהן:

- **בניית המילון**. מסיבות כלכליות לא מסרנו את משימת הבנייה של המילון הידני (שבעזרתו חישבנו את ה-precision וה-recall) לאדם שלישי, אלא עשינו זאת בעצמנו. לכאורה, עובדה זו עלולה להטות את התוצאות, כיוון שיתכן ובנינו את המילון בהתאם לתוצאות האלגוריתם שלנו. ניסינו להימנע מבעיה זו ע"י כך שהרצנו את האלגוריתמים רק לאחר שהמילון הידני היה מוכן, כך שתוצאותיהם לא עמדו לנגד עינינו בזמן הכנתו. עם זאת, יתכן, כמובן, שהערכותינו המוקדמות לגבי תוצאות האלגוריתם השפיעו, בכל זאת, על התהליך.
- **התאמת many to one.** הנחת היסוד לפיה כל מילה מתאימה למילה אחרת היא בעייתית. בשל הבדלי השפות, פעמים רבות ההתאמה היא בין כמה מילים למילה אחת. כך למשל, התמנית "*דברי*" מתאימה לתמניות "*The words of*". עם זאת, סוגיה זו היא אינהרנטית בהגדרת בעיית ה-word alignment, ועל כן משותפת לאלגוריתם שלנו ולאלגוריתם ה- k-vec ואינה רלוונטית להשוואה ביניהם.
- **דלילות הנתונים.** יישרנו את הטקסטים ברמת התמניות ולא השתמשנו במנתח מורפולוגי, ועל כן קיבלנו הרבה ערכים מחד, ומספר מופעים דליל יחסית מאידך. נטייתה של העברית

להשתמש הרבה בתחיליות החריפה את הבעיה - כך למשל, ניקדנו בנפרד את הזוגות: <*כל*, *all*>, <*בכל*, *all* >, < *לכל*, *all* >, < *וכל*, *all* >, < *מהכל*, *all* >, < *מכל*, *all* >. הרצת מנתח מורפולוגי כעיבוד מקדים הייתה מייצרת רק את הזוג <*כל*, *all* > במספר מופעים גדול הרבה יותר, ולכן גם עם ציון התאמה סטטיסטי גבוה יותר. אמנם, ניתן לבטל את הבעיה בטענה לפיה גם סוגיה זו נובעת מהגדרת הבעיה ופוגעת בשני האלגוריתמים גם יחד, אך עם זאת, יתכן שאחד האלגוריתמים נפגע יותר מן השני, ובפרט שיתרון השפה השלישית מתבטא באופן משמעותי יותר כאשר הנתונים פחות דלילים.

סיכום

שאלנו האם ניתן לשפר את תוצאות ה- word alignment משפה A לשפה B בעזרת שימוש בשפה שלישית C. הצענו אלגוריתם כללי המשקלל שפה שלישית באמצעות שימוש באלגוריתם סטטיסטי כלשהו לבעיית היישור כ"קופסא שחורה", ובחנו אותו במקרה של אנגלית, עברית, רוסית, מעל אלגוריתם ה- k-vec. לצערנו, לא הצלחנו לשפר את תוצאות האלגוריתם המקורי, ולכן איננו יכולים לענות בחיוב על שאלת הניסוי. עם זאת, אין הדבר מעיד על כך ששפה שלישית אינה מסוגלת לתרום לתוצאות טובות יותר בפתרון בעיית ה- word alignment. יתכן ששקלול אחר או עיבוד מקדים נוסף כמו ניתוח מורפולוגי ([segal2000]) או יישור לפי פסקאות או משפטים ([gale91]) היו עשויים להביא לתוצאות שונות. אפשרויות אלה מוצעות ככיווני מחקר אפשריים בעתיד.

4. נספחים:

1. תאור של אלגוריתם ה-k-vec:
appendix-k-vec.doc

2. טרנסליטרציה של הכתב העברי :
Heb-Transliteration.txt

3. הטקסטים המקוריים:
eng-final.txt
heb-final.txt
rus-final.txt

4. המילון "המושלם" שנבנה ע"י יישור ידני של הטקסט האנגלי מול הטקסט העברי:
gold-standard.txt

5. האחוז הטוב ביותר של האלגוריתם שלנו מול varma:
dicts-our-vs-varma.xls

5. מקורות

gale91 Gale, W. A. and Church, K. W., “A Program for Aligning Sentences in Bilingual Corpora”, *In Proceedings of ACL-91, Berkeley CA*, 1991

church93 Church K.W., Char_align: A Program for aligning Parallel Texts at the Character Level, *In Proceedings of 31-st Annual Meeting of the Association for Computational Linguistics, pp.1-8,* 1993

dagan93 I. Dagan, K. Church, & W. Gale, "Robust Word Alignment for Machine Aided Translation," *In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, available from the ACL, pp. 1-8*, 1993.

dag-itai91 Dagan, I., Itai, A., and Schwall, U., “Two Languages Are More Informative Than One”, *In Proceedings of the 2gth Annual Meeting of the Association for Computational Linguistics (ACL-91), pp. 130-137*, 1991

fung-church94 Fung Pascale and Church Kenneth. W, **“**K-vec: A New Approach for Aligning Parallel Texts”, *in Proceedings of COLING 94, pp. 1096-1102, Kyoto,* 1994

fung-keown94 Fung Pascale and McKeown Kathleen, “Aligning noisy parallel corpora across language group : Word pair feature matching by dynamic time warping *", in Proceedings of the First Conference of the Association for Machine Translation in the Americas, pp. 81-88, Columbia, Maryland*, 1994

melamed2000 Melamed I. Dan, “Models of Translational Equivalence among *Words”, In Computational Linguistics 26(2), pp. 221-249*, June, 2000

melamed2003 Melamed I. Dan, Green R. and Turian J. P., “Precision and Recall of Machine Translation”, *Proteus technical report #03-004, Presented at NAACL/HLT 2003, Edmonton, Canada*, 2003

m-s99 Manning Chris and Schütze Hinrich, Foundations of Statistical Natural Language Processing, *MIT Press. Cambridge, MA: May*, 1999

segal2000 Segal A., Hebrew Morphological Analyzer for Hebrew Undotted Texts, *M.Sc. thesis, CS Technion*, 2000

smadja94 Smadja F. and McKeown K., “Translating collocations for use in bilingual lexicons”, *In Proceedings of the ARPA Human Language Technology Workshop 94, Plainsboro, New Jersey*, 1994

turian2003 Turian J. P., Shen L. and Melamed I. Dan "Evaluation of Machine Translation and its Evaluation," Proceedings of MT Summit IX, New Orleans, LA., 2003

varma2002 Varma N., *Identifying Word Translations in Parallel corpora Using Measures of Association*, M. Sc. Thesis, University of Minnesota, 2002

white93 White J. S. and O'Connell T. A., "Evaluation of Machine Translation," *In Proceedings of the ARPA HLT Workshop, Princeton, N J*, 1993.