**מחשבות על מציאת אנלוגיות**

ישנם כמה (3-4) מאמרים.

**הראשון:**  <http://www.cs.huji.ac.il/~dshahaf/crowd-machine-learning16.pdf> - Scaling up Analogy with Crowdsourcing and Machine Learning

עקרונית נתנו לאנשים (אמזון מכניקל טורק) למצוא מסמכים אנלוגיים למסמך כלשהו (seed) שהם קיבלו.

מה שהם עושים זה:

1. מחפשים אחר התאמות
2. מסננים ומעבדים את ההתאמות שמצאו
3. מתארים את האנלוגיה (מדוע ההתאמה טובה).
4. מדגישים מילות מפתח.

בעזרת הפיצ'רים המתוחכמים האלו ניסו לפתור משימות similarity בין מילים – זאת בעזרת גרף בו הקודקודים הם queries, שהכניסו היוזרים על מנת למצוא אנלוגיות, ויש ביניהן קשת אמ"מ החיפושים האלו הובילו לאנלוגיה לאותו מוצר.

בנוסף, על מנת ללמוד אנלוגיות, השתמשו באנלוגיות שסימנו היוזרים (בינארי – 1 אם יש אנלוגיה, 0 אם אין) על מנת ליצור dataset מתויג. עליו התאמנו בעזרת רשת CNN סיאמית (כלומר, המשקולות לה זהות). הדאטה הוא מהצורה – כאשר ה- הם המסמכים, ו- אומרת אם יש ביניהן אנלוגיה. אימנו את זה בעזרת contractive loss.

#### השני: <http://www.hyadatalab.com/papers/analogy-kdd17.pdf> - [Accelerating Innovation Through Analogy Mining](http://www.hyadatalab.com/papers/analogy-kdd17.pdf) KDD 2017

במאמר הזה יש פירוט מעט שונה – עכשיו מניחים כי כל מוצר מורכב מ- purpose ומ-mechanism. אלו שני וקטורים שונים שנרצה לאסוף מהדאטה.

הדאטה שעובדים עליו הוא מ-quirky. איסוף ה-purpose וה-mechanism הוא בעזרת משימה שנתנו, שוב, באמזון מכניקל טורק. במשימה הזו המשתתפים סימנו משפטים שנראו להם קשורים לתכלית, ומשפטים שנראו קשורים למכניזם. בנוסף, התבצע איסוף אנלוגיות מהעובדים, בדומה למאמר הקודם, רק בסקייל נמוך יותר.

**שיטת הלמידה:**

הדאטה שלנו הוא מהצורה כאשר כל אחד כזה הוא טקסט על מוצר. כל טקסט על מוצר הוא למעשה:

* *– סדרה של טוקנים.*

*אנחנו אומרים כי ה*-purpose הוא למעשה וקטור בינארי שהוא 0 במקום בו הטוקן לא שייך ל-purpose, ו-1 במקומות שכן.

לכאורה המטרה כאן היא להתאים בין סיקוונס הטוקנים שקיבלנו לבין סיקוונס הפרפוס (למצוא את המתאים, אחרי שלמדנו על הדוגמאות. אגב, אותו הדבר תקף גם למכניזם). אבל השוני הוא שכאן לא מעניין אותנו בהכרח הוקטור עבור מוצרים שלא ראינו, אלא אולי איזשהו ייצוג שתופס אותם באופן רחב יותר. אז מה שמנסים לחזות זה איזו מניפולציה כזו (TD-IDF) על שרשור של ייצוגי המילים ב- , כאשר השרשור הוא של כל הפעמים שהופיעו עבור המדרגים השונים ב-mturk.

איך לומדים? משתמשים ב- bi-rnn . מה שאנו רוצים ללמוד הוא למעשה . הרשת מקבלת כקלט את .