**שילוב של למידה עמוקה בעולם עיבודי הטקסט – האם מהפיכה ב-NLP?**

**אורי חנני**

תקציר

בעת האחרונה פותחו מספר כלים המשלבים את הטכנולוגיה של למידה עמוקה בעולם עיבודי הטקסט. שילוב זה הביא לקפיצה גדולה באיכות הביצועים של מערכות NLP לעומת הכלים הקלסיים הקיימים. החזון של יצירת מכונות בעלות יכולת של Natural Language Understanding נראה קרוב מאי פעם. בצד זה גדל מאד מגוון היישומים האפשריים בשווקים האירגוניים והצרכניים, ויש צפי להורדת מחירי מערכות . במאמר זה אנו מנסים לבחון את התהליך ולתת מענה לשאלה האם אנו במהלך של מהפכה בתחום?

רקע

בשנה האחרונה הופיעו כלים חדשניים בעולם ה-NLP (Natural Language Processing). סבסטיאן רודד, אחד מהחוקרים המובילים בתחום, הגדיר תופעה זאת במשפט הבא: **'ה-ImageNet של השפה הגיע'**. במונח ImageNet הוא מתייחס למערכת המפורסמת של ראייה ממוחשבת וסיווג תמונות שבשנת 2012, הביאה למהפכה בכל הקשור ביישומי תמונה והשלכותיה בכל תחומי החיים, לרבות ההופעה של יישומי תמונה ווידיאו בסמארטפון.

החשיבות של ImageNet הייתה בהמחשה שרשתות ניואורניות שעברו אימון על מאגרי תמונות גדולים, יכולות לשמש אתחול יעיל (כלומר לא ממצב של אפס ידע) ללמידה של מאגרי תמונות שונים ומגוונים, וליצירת מערכות מיון וסיווג ספציפיות של מאגרי תמונות וגם וידיאו שאינן קשורות דווקא במאגרים הכללים עליהם התאמנו. דבר דומה נעשה ומתרחש עתה בשימוש של למידה עמוקה בתחום ה-NLP.

צפוי כי הופעת הכלים החדשים תביא לגל חדש של יישומים של עיבודי טקסט, בכל מגזרי החיים, ובמיוחד במישור הצרכני (B2C), ובמערכות מידע ותקשורת בין אישיות וארגוניות.

הגורם המרכזי להופעת כלים חדשניים אלו הוא השילוב של טכנולוגיות של למידה עמוקה ( Deep Learning) בכלים הקלסיים של NLP, כלים מבוססי מנועי חוקים ומהלכים סטטיסטיים בהם נעשה שימוש זה מספר עשורים.

כלי למידה עמוקה

הכלי הראשון שהופיע באמצע שנת 2018 הוא ULMFiT של חברת fast.ai ואחריו שורה של מוצרים ELMO של Allen Institute, Transformer ו-GPT2 של OpenAI, BERT של גוגל, ובחודשים האחרונים XLNet של גוגל ו-ERNIE 2.0 של ביאדו הסינית.

מאחורי כל הכלים האלו עומדים מספר עקרונות:

- אימון וחיזוי בכלים של רשתות ניארוניות מבוססות למידה עמוקה

- Transfer Learning – הפרדה מבנית בין מערכות ענק של אימון (Pre-Training) על מאגרי טקסטים ענקיים (כגון פרויקט גוטנברג, ויקיפדיה וכו'), לבין מערכות המשך המבצעות fine-tuning ומשמשות לאימון על מאגרים קטנים יחסית שעברו בד"כ תהליך של labelling כלומר, תיוג אנושי.
- Word Embeddings – שימוש בטכנולוגיה חדשה המבוססת על קשר שיש בין מילה במשפטי טקסט נותן ושכנותיה, וגזירת המשמעות הסמנטית של הקשר.
- אימון על מאגרי טקסט ענקיים כגון הוויקיפדיה. החשוב הוא שמאגרים אלו לא עברו שום תיוג אנושי והם מעובדים ונלמדים ע"י מערכות הלמידה העמוקה בצורתם המקורית.
- אימון בעזרת מחשבי ענק בענן והפעלה ע"י המשתמשים במחשבי קצה סטנדרטיים (לרבות הסמארטפון).

היתרונות שעולים מעקרונות אלו מתבטאים בהורדת עלויות, מול העלאה דרמטית ברמות הביצועים. במבחני ביצוע (Benchmarks) שנערכים משפרות מערכות אלו פעם אחר פעם את רמות הביצוע. בחלק מהמבחנים התוצאות מתקרבות (ולפעמים עוברות) ליכולת האנושית, ועומדות לשבור את מחסום מבחן טיורינג (Turing Test) המפורסם.

<u>הדגמות</u>

לשם בחינת התהליכים שתוארו למעלה והשלכותיהם המדעיות והמסחריות, פיתחנו מערכת הדגמה ובה יישומים קלסיים ומתקדמים בתחום ה-NLP.

בין היישומים במערכת נמנה את:

- יצירת אוטומטית של טקסט – קיימות הדגמות בשלשה תחומים של יצירת Fake Text:
  - יצירת חדשות 'מפוברקות'
  - יצירת הודעות 'מפוברקות' של ה- USA Federal Reserve
  - כתיבה של טקסטים מתוך מחזות של שייקספיר
- תרגום אוטומטי בין שפות
- חילוץ ויצירת קו זמן אוטומטי מתוך מסמכים – חילוץ ביטויי זמן (במגוון מופעים והקשרים) תוך הבנת ההקשר (Contextual Inference). בשילוב של תרגום בין שפות ניתן לקבל באופן אוטומטי קווי זמן (Time Line) של מספר מסמכים בשפות שונות, כאשר נתון כי המסמכים קשורים בצורה זאת או אחרת.
- לינגוסטיקה אוטומטית – זיהוי של משפטים נכונים או שאינם נכונים לינגוסטית.
- חילוץ אוטומטי של תשובות לשאלות מול טקסטים – יכולת של מתן תשובות לפי טקסט נתון כאשר נשאלת שאלת הקשר, שיש לה (או שאין) תשובה בתוך הטקסט.
- מנועי חיפוש סמנטיים – מציאת יעדי חיפוש לשאילתות בעזרת הבנה הקשר וסמנטיקה.
- תמציות אוטומטית של טקסטים – יצירת תמצית של טקסטים נתונים, ע"י Extractive Summary כלומר, בחירת המשפטים החשובים ביותר. ויצירת Abstractive Summary שהוא כתיבת תמצית

ע"י text generation ושימוש בטקסט שמן הסתם 'מבין' את המסמך המקורי, ולאו דווקא עושה שימוש בטקסט המקורי.
- ניתוח סנטימנט – גילוי הסנטימנט הנחבא בתוך טקסט נתון.

<u>סכום</u>

לכל היישומים שנמנו למעלה מכנה משותף אחד, שהוא בעצם בסיס המהפכה המתרחשת בתחום ה-NLP. המערכות האוטומטיות 'מבינות' טקסט, 'מבינות' הקשר ו'מבינות' משמעויות גלויות ונחבאות.

'הבנת' הטקסט הנעשית בצורה אוטומטית לאחר אימון על מאגרים עצומים, הינה רק בתחילת דרכה. עם הזמן, סביר שיעשה שיפור בכל הגורמים והפרמטרים שהוזכרו למעלה. תהליכים אלו בעצם מייתרים בעצם את התשובה לשאלה העומדת בבסיס מאמר זה: **המהפכה מתרחשת מול עינינו**.

לסיכום:

- כלים קלסיים של NLP יוחלפו ב-NLP מבוסס למידה עמוקה
- מגוון היישומים ילך ויעמיק בכל תחומי החיים
- עיבודי הטקסט יתרחבו במיוחד בתחום הצרכנים (B2C)
- השיפור הדרמטי בביצועים ילווה בירידת מחירים יחסית
- יישומים חדשים יופיעו בתחומי:
  - חיפוש סמנטי
  - יצירת אוטומטית של אונתולוגיות
  - אונתולוגיות תשולבנה בכלי DL בישומי NLP
- יישומי יצירת טקסט 'מפוברק' ייצרו בעיות אמינות ויהיה צורך במתן תשובות רגולטוריות ואתיות בקנה מידה ובממדים שלא הכרנו קודם.