

אופטימיזציית תיקון שגיאות OCR של עיתונות היסטורית בעברית באמצעות רשתות נוירונים ומיקור המונים / עמרי סויסה

תקציר

בשנים האחרונות ישנה מגמת דיגיטציה הולכת ומתעצמת של ספרים ועיתונים. הדיגיטציה מאפשרת הנגשה של ידע היסטורי ותרבותי לציבור תוך שימור הידע באופן אלקטרוני לדורות הבאים. תהליך אוטומטי של הפיכת תמונה לטקסט על ידי קריאת האותיות המוצגות בתמונה נקרא OCR (Optical Character Recognition). למרבה

הצער, לאחר ביצוע OCR נוצרות שגיאות בטקסט המתקבל. בכדי לשפר את התוצר הסופי מתבצע תהליך עיבוד מאוחר אשר מחפש שגיאות בטקסט ומנסה לתקן אותן. בתחום העיתונות ההיסטורית ישנן מספר רב של שגיאות OCR עקב מאפייני התחום הייחודיים. בשנים האחרונות גישת הלמידה העמוקה (deep learning) משולבת בתחום עיבוד השפה טבעית באופן ניכר. אך למרות הפופולאריות הגוברת של הגישה נראה שמעטים המחקרים על השימוש בה לצורך תיקון שגיאות OCR בעיתונות היסטורית בכלל ובעברית בפרט. אחת הסיבות לכך היא העדר כמות מספקת של מידע אמת (golden standard) לאימון רשתות נוירונים לנושא זה.

מטרת מחקר זה הינה לבחון כיצד ניתן לבצע אופטימיזציה לתהליך הלמידה של רשתות נוירונים לצורך תיקון מאוחר של שגיאות OCR. לשם כך נבחנו שלושת החלקים של התהליך: תהליך יצירת מידע אמת על ידי מיקור המונים, עיצוב ומבנה רשת הנוירונים האופטימלית למשימה זאת (נספח א) והתאמת הרשת לתחומי תוכן שונים. לבסוף נערכה השוואה בין מתקנים שונים (בני אדם, מתקני שגיאות כתיב מוכרים בשוק ורשתות נוירונים מותאמות לבעיה).

שאלות המחקר:

1. כיצד משפיעים המאפיינים הייחודיים של תיקון שגיאות OCR על איכות, יעילות ואפקטיביות התוצר

המתקבל על ידי עובדי מיקור המונים?

2. עד כמה משפיע סוג התוכן שעליו לומדת רשת הנוירונים על איכות התוצר (התיקון) החזוי?
3. עד כמה איכותיים מתקני שגיאות שונים (בני אדם, מתקני שגיאות כתיב מובילים בתעשייה ורשתות נוירונים)?

שיטה וממצאים:

לצורך אופטימיזציה של תיקון שגיאות OCR על ידי מיקור המונים נערך ניסוי בפלטפורמת מיקור ההמונים Amazon's Mechanical Turk. ניסוי זה בחן את המאפיינים הייחודיים של תיקון OCR (מתודולוגיית תיקון, תמונת הסריקה ואורך הטקסט) תוך התחשבות במגבלות התחום (שגיאות "הודיות OCR ומגבלות סגמנטציה). כל פריט במאגר הורכב מטקסט תקין, טקסט עם שגיאות OCR נפוצות "ותמונת סריקה" שנוצרה גם כן באופן מלאכותי. הניסוי בחן את השפעת המתודולוגיה, השפעת תמונת הסריקה והשפעת גודל הטקסט על שלושה מדדים: איכות, יעילות ואפקטיביות. בנוסף פותחו שלושה מדדים המתאימים למאפייני התחום הייחודיים. 753 עובדי מיקור המונים תיקנו 3796 טקסטים במהלך הניסוי. מדד האיכות הראה שכאשר הסגמנטציה מאפשרת, פסקאות (טקסט באורך בינוני) הן הגודל האופטימלי לתיקון. הן צריכות להיות מתוקנות במתודולוגיה המחלקת את התיקון לשתי תתי משימות נפרדות (מציאת שגיאות ותיקון שגיאות) ולהיות מוצגות לצד תמונת הסריקה. באופן מפתיע, כאשר הסגמנטציה לא מאפשרת זאת, תיקון כתבות שלמות (טקסט ארוך) בגישה הנאיבית של עריכה (איתור ותיקון במשימה אחת) היא האופטימלית ביותר. יתרה מכך, תיקון טקסטים ארוכים בגישה הנאיבית אופטימלית כאשר יעילות היא המדד הנדרש. לבסוף, כאשר מאזנים בין איכות ליעילות (אפקטיביות), נמצא שמשפטים בודדים מספקים את עלות התועלת הגבוהה ביותר. כאשר הסגמנטציה לא מאפשרת לפצל למשפטים אסטרטגיית התיקון הופכת למורכבת יותר. תוצאות ניסוי זה יכולות לעזור לכל חוקר אשר מבצע תיקון שגיאות כתיב באמצעות מיקור המונים בהפחתת מורכבות תכנון וביצוע ניסויים אילו ולכן יש למסקנות הללו שימוש מעשי חשוב מאוד במחקרים ויישומים עתידיים.

אופטימיזציה רשת הנוירונים נעשתה על ידי בחינת מאפיינים שונים המשפיעים על דיוק הרשת: עומק הרשת, רגולציה (dropout), כמות המאפיינים שהרשת יכולה ללמוד, סוג השכבה העיקרית ברשת, כמות הדוגמאות בכל תקופת למידה, כמות הדוגמאות בכל תהליך למידה והשפעת הלמידה הדו כיוונית. סדרה של ניסויים בוצעו על טקסטים ידועים שנלקחו מפרויקט בן יהודה (המקבילה העברית לפרויקט גוטנברג) ולבסוף גובשה רשת אופטימלית לתיקון שגיאות OCR. אופטימיזציה זאת שיפרה את דיוק הרשת מ-85% דיוק ל-88% ואת דיוק התוצר הסופי ב-4%.

הרשת האופטימלית מבוססת על ארבע שכבות LSTM דו כיוונית בגודל של 500 יחידות. מנגנון השכחה נמצא

אופטימלי כאשר הרשת זורקת 20% מהמידע בכל אימון. אימון אופטימלי מורכב מ-250,000 דוגמאות כאשר בכל תהליך אימון (אצווה) משתתפים 256 דוגמאות.

באמצעות המסקנות מהניסוי במיקור ההמונים בוצע ניסוי עם טקסטים של עיתונים היסטוריים בעברית מתוך מאגר העיתונות ההיסטורית של הספרייה הלאומית (JPress). בניסוי זה השתתפו 75 סטודנטים מתנדבים שתיקנו 150 כתבות. באמצעות מאגר (dataset) זה גובש אלגוריתם הזרקת שגיאות כתיב מותאם לשגיאות נפוצות. OCR בעיתונות היסטורית בעברית. באמצעות אלגוריתם זה נבנה מאגר מלאכותי נוסף על בסיס הטקסטים שנלקחו מפרויקט בן יהודה ואומנה הרשת האופטימלית שנית. בחלק זה נבחנה ההשפעה של סוג התוכן על איכות התיקון על ידי אימון רשת נוספת על התנ"ך והשפעת כמות השגיאות ("הרעש") על איכות הלמידה. נמצא שפער בין השפה התנכ"ית לשפה המודרנית גדול מידי בכדי לאפשר לרשת להכליל את הלמידה שלה ולתקן את OC[^] המודרני. בנוסף, נמצא שהזרקת רעש רב מונע מהרשת ללמוד בצורה אפקטיבית ונדרש מעט מאוד רעש (10%) בכדי להשיג למידה אופטימלית. לסיכום נערכה השוואה בין מתקני שגיאות כתיב ידועים בתעשייה (של חברת Microsoft וחברת Google), הרשתות האופטימליות שאומנו והסטודנטים. נמצא שלמרות ההתקדמות הרבה בתחום למידת המכונה, בני אדם מצליחים להכליל (generalize) בצורה טובה בהרבה כאשר הם מתקנים טקסטים בסוגה (טקסטים היסטוריים) שאינה מוכרת להם. התרומה של בני האדם לתיקון גדולה פי 15 מאשר של רשת הנוירונים אפילו לאחר אופטימיזציה של מבנה הרשת והשגיאות שעליהן למדה. אך בהשוואה למתקני שגיאות כתיב סטנדרטיים הרשת תרמה לסך הדיוק והאיכות הממוצעים בעוד שמתקני שגיאות הכתיב הנפוצים בתעשייה הכניסו יותר שגיאות מאשר תיקונים.

תוצאות סדרת ניסויים אילו יכולות לעזור לבניית כלים לתיקון שגיאות כתיב במאגרי עיתונות היסטורית בעברית ובכך לשפר את שימור המורשת העברית. כמו כן, חוקרים יוכלו להשתמש בתוצאות ניסויים אילו בכדי לצמצם את המורכבות בבניית רשת לתיקון שגיאות כתיב בעברית. בנוסף, טבלת השגיאות הנפוצות בעברית שגובשה יכולה לשמש בסיס למחקרי המשך בתחום ולשיפור תהליך OC[^] עצמו.

תוצרי מחקר זה יכולים לשמש כבסיס למחקרים נוספים בתחום תיקון שגיאות הכתיב בעברית, תיקון שגיאות OCR בעברית (ואולי בשפות נוספות) ושיפור איכות זיהוי האותיות בתהליך OCRⁿ עצמו. בנוסף, ניתן להמשיך את מחקר האופטימיזציה, הן של מיקור ההמונים והן של רשתות הנוירונים, על ידי בחינת פרמטרים נוספים שלא נבחנו במחקר זה והשוואה למאגרי מידע (תחומי תוכן) נוספים.

006.424 סוי.או תשע"ט

מספר מערכת: 9926564169405776