

מידול רשתות נוירונים עמוקות למענה על שאלות טבעיות

על מאגרי נתונים מובנים למחצה / עמרי סויסה

תקציר

עבודת הדוקטורט זו מורכבת מאוסף של שלושה מאמרים הדנים במענה אוטומטי לשאלות בשפה טבעית באמצעות רשתות נוירונים עמוקות בתחום הגנאלוגי (עצי משפחה) כחלק מתחום מדעי הרוח הדיגיטליים. שילוב טכנולוגיות חישוביות ומדעי הרוח הוא מאמץ מתמשך שמטרתו להפוך משאבים כגון טקסטים, תמונות, שמע, ווידאו לזמינים דיגיטליים, ניתנים לחיפוש וניתנים לניתוח. בשנים האחרונות, רשתות נוירונים עמוקות חולשות על תחום ניתוח הטקסט האוטומטי ועיבוד השפה הטבעית (NLP), ובמקרים מסוימים מציגות ביצועים טובים יותר מאשר בני אדם. רשתות נוירונים עמוקות הם האלגוריתמים המתקדמים ביותר ללימוד מכונה, הפותרים משימות עיבוד שפה טבעית רבות הרלוונטיות למחקרים בתחום מדעי הרוח הדיגיטליים, כגון בדיקת איות, זיהוי שפות, חילוף ישויות, זיהוי מחברים, מענה על שאלות ומשימות אחרות. אלגוריתמים מפותחים אלה לומדים דפוסים ממספר רב של דוגמאות "נכונות" ו"לא נכונות" ומיישמים אותם על דוגמאות חדשות. עם זאת, שימוש ברשתות נוירונים עמוקות לניתוח משאבי הטקסט במחקרי מדעי הרוח הדיגיטליים מציב אתגרים, כגון חוסר הזמינות של נתונים מותאמים לאימון מודלים וצורך בהתאמת המודלים לתחום המחקר הספציפי. אוסף מאמרי מחקר זה יחקור אתגרים אלה וכיצד להתגבר עליהם על ידי (1) עיצוב מודל החלטה מעשי עבור מומחים מתחום מדעי הרוח הדיגיטליים בנוגע למתי ואיך לבחור את גישות הלמידה העמוקה המתאימות למחקר. מודל החלטה זה מבוסס על ניתוח מחקרים מתחום מדעי הרוח הדיגיטליים בספרות העדכנית ופתרונות אפשריים. (2) תכנון ואימות אמפירי של ארכיטקטורה (מקצה לקצה) ומתודולוגיה חדשה להתאמת מודל רשתות נוירונים עמוקות למשימת "קריאה קרובה" (מענה על שאלות בשפה חופשית על אדם או קבוצת אנשים ספציפיים) על גרפי ידע גנאלוגיים בשילוב טקסטים לא מובנים. ו(3) תכנון ואימות אמפירי של ארכיטקטורה (מקצה לקצה) ומתודולוגיה חדשה להתאמת מודל רשתות נוירונים עמוקות ל"קריאה רחוקה" (שאלות כמותיות על העץ כולו או מספר גדול של אנשים) באותו התחום.

בסיס הנתונים הגנאלוגי העיקרי עבור מאמרים אלה מורכב מ-3,140 עצים משפחתיים המכילים 1,847,224 אנשים שונים מהמרכז לגנאלוגיה יהודית על שם דאגלס א. גולדמן במוזיאון אנו. כפי שמתואר בפרקי המאמרים, בשל המבנה והמאפיינים הייחודיים של עצי משפחה גנאלוגיים, תחום זה דורש מחקר ומודלים ספציפיים. בסיס נתונים פתוחים נוספים שימשו כבסיס לאימון למודלים בהתאם למאמר הרלוונטי.

מאמרים אלה מציגים את שני האתגרים העיקריים שכמעט כל חוקר DH/LIS יכול לצפות להיתקל בו באמצעות מודלים של DNN בזמן ביצוע מחקר. המאמר הראשון, "Text Analysis Using Deep Neural Networks in Digital Human Sciences and Information Science", מציג את האתגרים העיקריים של שימוש ברשתות נוירונים במחקר במדעי הרוח הדיגיטליים ואת מודל ההחלטה המוצע לטיפול באתגרים אלו. יתרה מכך, מאמר זה טוען כי על חוקרי מדעי המידע \ מדעי הרוח הדיגיטליים להרחיב את ארסנל המיומנויות והשיטות החשובות שלהם.

המאמר השני, "Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs", מציג שיטה המתמודדת עם מספר אתגרים למענה על שאלות עובדתיות בתחום הגנאלוגי, כגון ייצוג גרף גנאלוגי כגרף ידע, אלגוריתם למעבר על גרפים גנאלוגיים המותאם לפרשנות הגנאלוגית של משמעות הקשר בין ה"עלים" (אנשים ומשפחות), יצירת בסיס נתונים מפורמט גנאלוגי (GEDCOM), והגדרת סוגי שאלות גנאלוגיות בתחום. המודל המאומן על בסיס הנתונים הגנאלוגי שנוצר עם קשרים מדרגה שנייה, Uncle-BERT₂, הניב ציון F1 של 81.45 בהשוואה למודל הבסיס BERT (60.12 בלבד) שאומן על בסיס נתונים סטנדרטי של שאלות כלליות. יתרה מכך, מאמר זה בדק את ההשפעה של סוג השאלה על הדיוק של מודלים של רשתות נוירונים במענה לשאלות גנאלוגיות ומצא שלשאלות הקשורות לתאריך יש תכונות כלליות יותר מאשר שאלות אחרות ושאלות הקשורות למקום רגישות יותר "לרעש" מאשר סוגי שאלות אחרות.

המאמר השלישי, "Around the GLOBE: Numerical Aggregation Question-Answering on Heterogeneous Genealogical Knowledge Graphs with Deep Neural Networks", מתאר ומיישם מתודולוגיה רב-שלבית (מקצה לקצה) עבור מענה לשאלות כמותיות באמצעות רשתות נוירונים בתחום הגנאלוגי. התוצאות מראות שהמודל GLOBE, מצליח לענות על שאלות כמותיות ברמת איכות גדולה יותר מהמודל הגנרי המתקדם בתחום. יתרה מכך, מאמר זה בחן את ההשפעה של עיצוב בסיס הנתונים על הדיוק של המודל והראה כי המורכבות של התחום הגנאלוגי דורשת ארכיטקטורה מורכבת יותר שיכולה לפצל ולהרכיב טבלאות מבסיס הנתונים בהתבסס על השאלה. בנוסף נמצא שהעיצוב של בסיס הנתונים המבוסס על דרגת הנרמול השישית (Date et al., 2004) הוא העיצוב האפקטיבי ביותר. יתר על כן, מאמר זה מראה את ההשפעה של הפעולה המתמטית על יכולתו של מודל הבינה המלאכותית לחזות תשובות קרובות לתשובה הנכונה גם כאשר הוא טועה.

לסיכום, התרומות של המאמרים הללו הן (1) מודל החלטה לטיפול באתגרי השימוש ברשתות נוירונים לחוקרי מדעי המידע \ מדעי הרוח הדיגיטליים, (2) ייצוג גרף ידע גנאלוגי על בסיס תקן GEDCOM, (3) אלגוריתם למעבר על גרפים גנאלוגיים (Gen-BFS), (4) בסיס נתונים לאימון מודל בינה מלאכותית שנוצר באופן אוטומטי בסגנון SQuAD לתחום הגנאלוגי (Gen-SQuAD), (5) מודל מבוסס-BERT המותאם למענה על שאלות עובדתיות עבור התחום הגנאלוגי (Uncle-BERT), (6) ייצוג בסיס נתונים גנאלוגי ויצירת מבנה גרף ידע מתקן GEDCOM לבסיס נתונים מובנה עבור תשובות לשאלות כמותיות (GenAgg), (7) הערכה של עיצוב בסיס הנתונים האופטימלי עבור מענה לשאלות כמותיות בתחום הגנאלוגי, (8) ארכיטקטורה למענה לשאלות כמותיות בתחום הגנאלוגי, ו- (9) מודל מבוסס BERT למענה לשאלות כמותיות בתחום הגנאלוגי (GLOBE).

חשוב לציין שניתן ליישם את המתודולוגיות המוצעות על משימות עיבוד שפה טבעיות (NLP) אחרות בתחום הגנאלוגי (ובתחומי מדעי הרוח הדיגיטליים נוספים), כגון חילוץ ישויות, סיווג טקסט, סיכום טקסט וכו'. חוקרים יכולים להשתמש בתוצאות של מאמרים אלה כדי להפחית הזמן, העלות והמורכבות ולשפר את הדיוק במחקר עיבוד שפה טבעיות בתחום הגנאלוגי.

מס' מערכת 9926875312505776