⚡️🚀המאמר היומי של מייק 21.06.24:⚡️🚀

Named Entity Recognition as Structured Span Prediction

היום נסקור מאמר בנושא שלא סקרתי הרבה מאוד זמן והוא Named Entity Recognition או NER. מטרת משימה זו היא לזהות בטקסט עצמים(מילים וקבוצות מילים רצופות) מסוגים מסוימים כמו שמות פרטיים, כתובות מגורים, מספרי ת״ז וכדומה. קיימים מודלי NER המתמחים בזיהוי שמות חברות, רשומות רפואיות וכדומה.

מחד גיסא משימת NER היא משימה דיסקרימינטיבית והתוצאה שלה היא סיווג של כל טוקן במשפט לקטגוריה שהוא שייך או לקטגוריה "O" אם הוא לא שייך לאף קטגוריית יעד(נציין כי הזיהוי מתבצע פר מילה ולא פר טוקן מכיוון שמילה עשויה להיות מורכבת מכמה טוקנים). מאידך גיסא ניתן בקלות להפוך אותה לבעייה גנרטיבית כאשר המודל יגנרט את העצמים השייכים לקטגוריות יעד.

בעבר פותחו מגוון שיטות למשימה הזו, חלקם rule-based, חלקם סטטיסטיים אך לאחרונה רשתות השתלטו לנו על NLP וגם המשימה הזו לא הצליחה לברוח מהן. הוצאו לא מעט רשתות שהגיעו לביצועים די יפים במשימה הזו.

המאמר שנסקור היום מציע גישה מעניינת לבעיית NER. כמו שאמרתי ניתן לפתור את הבעיה הזו באופן דיסקרימינטיבי וגנרטיבי אך המאמר הזה לוקח גישה בין אלו וקורא לה Structured Span Prediction.

בגדול הגישה עובדת באופן הבא. מעבירים את כל שמות הקטגוריות יחד (מופרדים עם טוקן מיוחד) דרך טוקנייזר משלהם. לאחר מכן מעבירים את הטקסט דרך טוקנייזר משלו ומכניסים את שניהם דרך מודל שפה דו-כיווני ( bidirectional או encoder) כמו BERT או DeBerta. המודל מפיק ייצוגי הטוקנים תלוי הקשר (גם עבור קטגוריות וגם עבור הטקסט) בתור פלט.

החידוש האמיתי בא לאחר מכן. הרי המטרה של NER היא לזהות כמה מילים רצופות השייכים לאותה קטגוריה. נגיד אנו לוקחים את המילים מ 1 עד 4 ומנסים לזהות מה הסיכוי שהם שייכים לקטגוריה c. המאמר מציע לקחת את הייצוגים תלויי הקשר של מילה 1, מילה 4 (המחברים מציעים להשתמש בייצוג של הטוקן הראשון של כל מילה לייצוג המילה) וגם קטגוריה c ובונים (מאמנים) מודל קטן לשערוך הסתברות זו. יש כמובן הרבה גישות לארכיטקטורה של מודל דליל זה. אפשר לעשות את עם רשת קונבולוציות פשוט, ניתן לקנקט את הייצוגים ולהוסיף שכבה לינארית ואני יכול לחשוב על כמה אופציות נוספות.

עכשיו השאלה האחרונה היא איך לבחור קטגוריות לכל המילים. השיטה הנאיבית היא לחשב את ההסתברויות האלו עבור כל תת סדרה של מילים רצופות החל מהמילה הראשונה ועבור כל תת סדרה לבחור את הקטגוריה בעלת הסתברות הגבוהה ביותר אם היא עולה על סף מסוים או קטגוריה ריקה אם זה לא. הבעיה עם הגישה הזו שכך נוכל לפספס spans ארוכים אחרי שסימנו את ה-span קצר יותר שיש לו חיתוך עם ה-span הארוך.

עקב כך המאמר מציע כמה גישות שונות לבעיה הלא פשוטה זו וביניהם Conditional Random FIelds וגם Maximum Weight Independent Set Method. גם Exhaustive Search יכול לעבוד עבור טקסטים קצרים. יש כאן טעימה נחמדה של שיטות מעניינות הלא קשורות לרשתות.

ואיך מאמנים את זה? האמת זה די פשוט - לוקחים את כל הקטגוריות המסומנות בטקסט ומריצים cross-entropy loss על כולם.

מאמר מאוד מעניין ומאיר עיניים - מחר ההמשך…

https://aclanthology.org/2022.umios-1.1/