⚡️🚀המאמר היומי של מייק -07.12.24: ⚡️🚀  
Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering

#### 1. תמצית המאמר

המאמר מציע שיטה המציידת RAG עם מערכת מבוססת גרפי ידע (KG) המותאמת לשירות לקוחות. המערכת, שפותחה על ידי צוות המחקר של LinkedIn, מעשירה LLMs בידע מבני שמקורו בפניות שירות היסטוריות. על ידי שילוב יחסים שונים בין פניות השירות (טיקטים) בגרף, השיטה משפרת באופן משמעותי את דיוק האחזור(retrieval), איכות התשובות והיעילות, עם שיפורים ניכרים במדדים כמו MRR, BLEU ומקטין זמני הטיפול בפניות.

#### 2. תרומות מרכזיות

##### א. שילוב KG במערכות RAG

שימור מידע מבני:  
כל טיקט מיוצג כעץ (יחסים פנימיים בתוכו) ומקושרת לפניות אחרות דרך יחסים סמנטיים או מפורשים. עיצוב זה משמר את ההיגיון הלוגי של הטיקט, כולל תיאור הבעיה והפתרון. כל טיקט מהווה צומת בגרף.

שיפור באחזור ויצירת תשובות:  
המערכת מנווטת בגרף כדי לזהות תתי-גרפים רלוונטיים, המוזנים ל-LLMs לצורך יצירת תשובות איכותיות.

##### ב. בניית גרף הידע:

עץ פנימי לטיקט:  
צמתים מייצגים חלקים כמו סיכומים או שורשי בעיה, וקשתות מציינות יחסים היררכיים.

קשרים בין פניות:

קשרים מפורשים: יחסים כמו (e.g., "clone of" or "caused by").

קשרים סמויים: מחושבים על בסיס דמיון קוסיין בין אמבדינגס.

##### ג. שלבים בתהליך אחזור ותשובות

המערכת פועלת ב 3 שלבים:

זיהוי ישויות(entity) וכוונות:  
המערכת הופכת שאילתות משתמש לישויות וכוונות(intents) באמצעות LLMs וניתוח ותבניות YAML.

אחזור תת-גרפים:  
מתבצע חישוב דמיון בין אמבדינגס לשאילתה לצמתים בגרף לזיהוי תת-הגרפים הרלוונטיים ביותר.

יצירת תשובות:  
המערכת יוצרת תשובות בהתבסס על תת-הגרפים רלוונטיים לשאילתת המשתמש.

4. קצת פרטים על השיטה

השיטה המוצעת כוללת 3 שלבים עיקריים:

a. זיהוי ישויות בשאילתה וזיהוי כוונה(intent):

המערכת מעבדת שאילתות משתמש על ידי חילוץ ישויות מוגדרות וכוונות באמצעות ניתוח תבניות YAML ו-LLMs. ישויות מוגדרות מייצגות אופיינים מהותיים (למשל, "תקציר בעיה" או "תיאור בעיה"), בעוד כוונות(intents) מכילות את מטרת השאילתה (למשל, "פתרון תיקון"). לדוגמה, בהינתן השאילתה "כיצד לשחזר את בעיית ההתחברות כאשר משתמש לא יכול להתחבר ל-LinkedIn?", המערכת מזהה את הישויות כ"בעיית התחברות" ו"משתמש לא יכול להתחבר" ואת הכוונה כ"פתרון תיקון."

b. אחזור מבוסס אמבדינגס (ייצוג):

זיהוי פניות רלוונטיות: מחשבים עד כמה הישויות שחולצו משאילתת המשתמש (למשל, "בעיית התחברות") תואמות את הצמתים ב-KG. עבור כל יישות בשאילתה, השיטה משתמשת בדמיון קוסיין למדידת קרבה בין ייצוג הישות לייצוגים של צמתים בגרף. הציונים מצטברים על פני כל הצמתים השייכים לטיקט מסוים. ככל שלטיקט יש מספר ישויות קרובות לשאילתה, הציון שלו עולה, מה שהופך אותו לסביר יותר להיבחר כרלוונטי.

חילוץ תת-גרף רלוונטי: לאחר זיהוי טיקטים הרלוונטיים ביותר, הם משמשים לבניית שאילותות למסד נתונים (DB) בשפת שאילתות גרפים הנקראת Cypher. שאילתות אלה מאפשרות למערכת לחלץ תת-גרפים מקושרים, כגון תיאורים קשורים או שלבים לשחזור בעיה. תהליך האחזור המובנה הזה מבטיח(סוג של כמו תמיד) שהמערכת אוספת מידע מדויק ורלוונטי מבחינת ההקשר מגרף הידע.

c. יצירת תשובה:

מגנרטת תשובות על ידי קישור נתוני הגרף שאוחזרו עם השאילתה המקורית. LLM מנסח מחדש את השאילתה באופן דינמי ומייצר תשובות מובנות. לדוגמה השאילתה "שגיאת העלאת csv בעדכון אימייל משתמש" מנוסחת מחדש ל-Cypher לאינטראקציה עם DB, מאחזרת פתרונות צעד-אחר-צעד.

#### 6. סיכום

המאמר מציג דרך פורצת דרך לשילוב גרפי ידע במערכות RAG עבור מענה לשאלות בשירות לקוחות. על ידי לכידת יחסים פנימיים וחיצוניים בין פניות, המערכת משפרת משמעותית את דיוק האחזור ואיכות יצירת התשובות, ומציבה כיוון מעניין ביישומים פרקטיים של LLMs.

https://arxiv.org/abs/2404.17723: