ID:84

Erişim Destekli Üretim (RAG) Tabanlı Türkçe Soru-Cevap Sistemlerinde Geri Çağırma Yöntemlerinin Karşılaştırılması

Anar Gurbanov[1], Metin Bilgin[2]

[1, 2] Bursa Uludağ Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye

anarqurbanov0020@gmail.com, metinbilgin@uludag.edu.tr

Özet: Bu çalışmada, Türk tarihi alanına yönelik Erişim Destekli Üretim (RAG) tabanlı bir Türkçe soru-cevap sistemi geliştirilmesi üzerinedir. Sistem kapsamında, Gömme Tabanlı Geri Çağırma (embedding-based retrieval), Yoğun Geçit Erişimi (Dense Passage Retrieval – DPR) ve BM25 yaklaşımları karşılaştırmalı olarak değerlendirilmiştir. Hem geri çağırma (retrieval), hem de üretim (generation) için performans ölçümleri yapılarak performans farklılıkları analiz edilmiştir.

Anahtar Sözcükler: Erişim Destekli Üretim (RAG), Türkçe Soru-Cevap Sistemleri, Yoğun Geçit Erişimi (DPR), Büyük Dil Modelleri (LLM)

1. Giriş

Günümüzde akademik ve uzmanlık gerektiren alanlarda hızlı ve doğru bilgiye erişim kritik önemdedir. Ancak geleneksel LLM'ler, eğitim verileriyle sınırlı olmaları ve halüsinasyon üretmeleri nedeniyle güvenilirlik sorunları yaşamaktadır.

Bu çalışmada, bu sorunları azaltmak için RAG tabanlı bir Türk tarihi soru-cevap sistemi geliştirilmiştir. TQuAD, İslam Ansiklopedisi ve Vikipedi verileri kullanılmış; BM25 ve DPR karşılaştırılmıştır. Sistem, geri çağırma için Top-k ve Recall@k, üretim için BLEU, ROUGE ve F1 metrikleriyle değerlendirilmiştir.

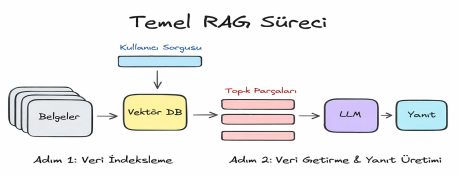
2. Önceki Çalışmalar

Büyük Dil Modelleri (LLM), 2017'de geliştirilen Transformer mimarisiyle NLP alanında köklü bir dönüşüm yaratmıştır. Dikkat (attention) mekanizması sayesinde paralel işlem yapılabilmiş ve RNN'lerin sıralı işlem sınırlamaları aşılmıştır [4]. GPT-3 ve T5 gibi büyük modeller birçok görevde yüksek başarı sağlasa da, kapalı kitap yapıları nedeniyle eğitim verileri dışındaki güncel bilgilere erişememekte ve bilgi kesilmesi (knowledge cutoff) sorunu yaşamaktadır. Bu durum, güncellik ve uzmanlık gerektiren uygulamalarda güvenilirliği sınırlamaktadır.

A. Erişim Destekli Üretim (Retrieval-Augmented Generation)

RAG yaklaşımı, LLM'lerin bilgi sınırlılığını aşmak için Lewis ve ark. tarafından önerilmiştir [5]. Bu mimari, yanıt üretiminde harici belgeleri kullanarak doğruluğu artırır ve halüsinasyonları azaltır.

RAG iki aşamadan oluşur: geri çağırma ve üretim. İlgili belgelerin bağlam olarak sunulması, özellikle yoğun geri çağırma kullanıldığında açık alan soru-cevap görevlerinde performansı belirgin biçimde yükseltmektedir [4]. Şekil 1'de temel RAG sistemi gösterilmektedir.



Şekil 1. Temel RAG sistemi

Geri çağırma yöntemleri BM25 ve DPR olarak ikiye ayrılır. BM25 kelime tabanlıdır, DPR ise anlamsal eşleştirme sağlar ancak Türkçe için maliyetlidir [3].

Bu çalışmada, düşük kaynaklı dillere uygun Sentence-BERT tabanlı gömme kullanılmıştır. RAG sistemlerinde başarının ana belirleyicisi geri çağırma bileşenidir.

B. Türkçe Doğal Dil İşleme ve Soru-Cevap Çalışmaları

Türkçe'nin eklemeli yapısı anlamsal aramayı zorlaştırsa da BERTurk ve bert-base-turkish-squad gibi modeller bu alanda önemli ilerleme sağlamıştır. Ancak RAG çalışmalarında geri çağırma stratejilerinin etkisi çoğunlukla yeterince karşılaştırılmamıştır; TQuAD bu açıdan temel bir veri setidir. Halüsinasyon, yanlış veya doğrulanamaz bilgi üretimi olup güvenilirliği düşürür. RAG mimarisi, yanıtları harici belgelere dayandırarak bu sorunu azaltır. Bu nedenle değerlendirmede yalnızca üretim metrikleri değil, Recall@k gibi geri çağırma ölçütleri de dikkate alınmalıdır.

3. Yöntem

Bu çalışmada, Türkçe tarih alanına yönelik RAG tabanlı, modüler bir soru-cevap sistemi geliştirilmiştir. Amaç, yalnızca doğru cevaplar üretmek değil, aynı zamanda cevapların kaynaklarını açık biçimde sunmaktır.

A. Veri Kaynakları

Sistemin başarısını artırmak için tek bir veri seti yerine birden fazla kaynak kullanılmıştır. Ana veri kaynağı TQuAD [7] olup, buna ek olarak İslam Ansiklopedisi'nden yaklaşık 120.000

içerik ve Türkçe Vikipedi'den 750 tarih başlığı sisteme dahil edilmiştir. Bu çeşitlilik, gerçekçi kullanım senaryolarını değerlendirmeyi amaçlamaktadır.

B. Metin Parçalama (Chunking)

RAG sistemlerinde parça boyutu kritiktir; literatürde ideal aralık 600–900 karakter olarak belirtilmektedir. Bu çalışmada metinler, cümle yapısı korunarak yaklaşık 800 karakterlik parçalara ayrılmış ve 100 karakterlik örtüşme kullanılarak bağlam sürekliliği sağlanmıştır [5].

C. Gömme (Embedding) ve Temsiller

Bu çalışmada üç temsil yöntemi karşılaştırılmıştır: Türkçe için eğitilmiş Sentence-BERT tabanlı gömme, DPR ve kelime temelli BM25. Sentence-BERT anlamsal benzerlikte en dengeli sonuçları verirken, DPR İngilizce ağırlıklı eğitimi nedeniyle Türkçe'de daha zayıf kalmıştır. BM25 ise klasik ve hâlen güçlü bir temel yöntemdir.

D. Geri Çağırma ve Yeniden Sıralama

Geri çağırma süreci iki aşamalıdır: ilk aşamada yüksek recall hedeflenmiş, ikinci aşamada aday belgeler çapraz-encoder ile yeniden sıralanmıştır. Yoğun gömme benzerliği iç çarpım ile hesaplanmış ve FAISS kullanılarak hızlandırılmıştır [5].

E. Üretim (Generation)

En iyi belgeler, Türkçe BERT tabanlı soru-cevaplama modeli ile işlenmiş; yeniden sıralama ve güven skorları birleştirilerek nihai cevap seçilmiştir. Bu yöntem, halüsinasyon riskini azaltmaktadır [6].

F. Değerlendirme Metrikleri

Sistem, geri çağırma ve üretim aşamaları birlikte değerlendirilmiş; geri çağırma performansı Top-k Accuracy ve Recall@k metrikleriyle ölçülmüştür.

4. Bulgular

Bu bölümde, RAG tabanlı Türkçe tarih soru-cevap sisteminde Gömme Tabanlı, DPR ve BM25 geri çağırma yöntemlerinin karşılaştırmalı sonuçları sunulmaktadır. Değerlendirmeler aynı soru kümesi üzerinde, Bölüm 3'te tanımlanan metriklerle yapılmış; sonuçlar Tablo 1 ve Tablo 2'de verilmiştir.

Tablo 1. Geri Çağırma performanslarının karşılaştırılması

Yöntem	Güven	Recall@1	Recall@3
EBR	0.46	0.45	0.65
DPR	0.26	0.1	0.15
BM25	0.25	0.3	0.35

Tablo 2. Üretim Performanslarının karşılaştırılması

Yöntem	Güven	F1	BLEU	ROUGE-L
EBR	0.46	0.34	0.25	0.35
DPR	0.26	0.025	0.003	0.025
BM25	0.25	0.13	0.082	0.14

Sonuçlar, Gömme Tabanlı Geri Çağırma yönteminin tüm metriklerde en dengeli performansı sunduğunu göstermektedir. F1 ve ROUGE-L üstünlüğü, Türkçe için eğitilmiş Sentence-BERT modellerinin anlamsal benzerlikte daha etkili olduğunu ortaya koymaktadır. BM25, Recall@k değerlerinde rekabetçi sonuçlar üretirken, DPR İngilizce odaklı yapısı nedeniyle daha zayıf kalmıştır. Ayrıca Recall@1 ve Recall@3, doğru belgelerin üst sıralarda getirilmesinde gömme tabanlı yaklaşımın avantajını göstermektedir.

5. Sonuç ve Tartışma

Bu çalışmada, Türkçe tarih alanına yönelik RAG tabanlı bir soru-cevap sistemi geliştirilmiş ve farklı geri çağırma stratejilerinin etkileri karşılaştırılmıştır. Sonuçlar, RAG mimarisinde geri çağırma aşamasının çıktı kalitesini doğrudan belirlediğini göstermektedir.

DPR yaklaşımının düşük performansı, modellerin Türkçe'ye özgü anlamsal temsilleri yeterince yakalayamamasından kaynaklanmaktadır. Buna karşılık, Sentence-BERT tabanlı yoğun gömme yöntemi Türkçe tarih alanı için en dengeli ve güvenilir yapılandırma olarak öne çıkmıştır. BM25 ise düşük hesaplama maliyeti ve yorumlanabilirliği sayesinde güçlü bir alternatif veya tamamlayıcı yöntem olarak değerlendirilebilir.

Kaynaklar

[1] H. Naveed et al., "A Comprehensive Overview of Large Language Models," arXiv preprint arXiv:2307.06435, 2023.

[2] Y. Zhang et al., "A Bibliometric Review of Large Language Model Hallucination," arXiv preprint arXiv:2401.XXXXX, 2024.

[3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781.

[4] A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] S. Yıldırım, "Turkish BERT-Based Question Answering Model," Hugging Face Model Repository, 2020. [Online]. Available: https://huggingface.co/savasy/bert-base-turkish-squad

[7] TQuAD Team, "TQuAD: Turkish Question Answering Dataset v0.1," 2020.

[8] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[9] A. Mallen et al., "When Not to Trust Language Models: Investigating Hallucinations and Their Mitigation," arXiv preprint arXiv:2301.13817, 2023.