T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

METİN ANALİZİ TABANLI SORU CEVAPLAMA SİSTEMLERİ İÇİN MODEL TABANLI DEĞERLENDİRME METRİĞİ

Dilan BAKIR

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Prof. Dr. Mehmet S. AKTAŞ

Eş Danışman

Doç. Dr. Beytullah YILDIZ

Mayıs, 2025

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

METİN ANALİZİ TABANLI SORU CEVAPLAMA SİSTEMLERİ İÇİN MODEL TABANLI DEĞERLENDİRME METRİĞİ

Dilan BAKIR tarafından hazırlanan tez çalışması 27.05.2025 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Prof. Dr. Mehmet S. AKTAŞ Yıldız Teknik Üniversitesi Danışman	Doç. Dr. Beytullah YILDIZ Atılım Üniversitesi Eş-Danışman

Jüri Üyeleri

Prof. Dr. Mehmet S. AKTAŞ, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Oya KALIPSIZ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Selim AKYOKUŞ, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Songül VARLI, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Nizamettin AYDIN, Üye
İstanbul Teknik Üniversitesi _____

Danışmanım Prof. Dr. Mehmet S. AKTAŞ sorumluluğunda tarafımca hazırlanan "Metin Analizi Tabanlı Soru Cevaplama Sistemleri İçin Model Tabanlı Değerlendirme Metriği" başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Dilan BAKIR

İmza

Tezimi yeğenlerim Yiğit Ali ve Melis Alya'a ithaf ediyorum

TEŞEKKÜR

Tezimi tamamlama sürecinde bana destek olan ve yol gösteren herkese en içten teşekkürlerimi sunmak isterim. Öncelikle, tez danışmanım Prof. Dr. Mehmet S. AKTAŞ'a derin minnettarlığımı ifade etmek isterim. Kendisinin akademik rehberliği, bilgi birikimi ve sabrı sayesinde bu çalışmayı şekillendirme ve geliştirme fırsatı buldum. Değerli yönlendirmeleri ve kıymetli zamanını benimle paylaşarak araştırmamın her aşamasında bana ışık tuttuğu için kendisine teşekkür ederim.

Ayrıca, eş danışmanım Doç. Dr. Beytullah YILDIZ'ın da katkılarından dolayı şükranlarımı sunarım. Kendisi, tez sürecinde verdiği yapıcı eleştiriler ve yol gösterici önerileri ile çalışmamın daha sağlam bir temele oturmasına yardımcı olmuştur. Bunun yanı sıra, tez komite üyelerime de araştırmamı titizlikle değerlendirdikleri ve kıymetli geri bildirimleriyle sürecimi daha verimli hale getirdikleri için teşekkür ederim. Tez savunmamda yer alacak değerli hocalarıma da vakit ayırarak çalışmamı değerlendirdikleri ve bana bu süreçte akademik destek sundukları için minnettarım.

Bu süreçte manevi destekleriyle her zaman yanımda olan aileme çok teşekkür ederim. Sabırları, sevgileri ve teşvikleri sayesinde bu süreci daha güçlü ve motive bir şekilde geçirdim. Özellikle zorlu zamanlarda verdikleri destek, bu çalışmayı tamamlamamda büyük bir rol oynadı. Özellikle yeğenlerim Yiğit Ali ve Melis Alya'nın varlığı, akademik hayatım boyunca en büyük motivasyon kaynaklarımdan biri oldu.

Son olarak, bu süreçte birlikte çalıştığım sınıf arkadaşlarıma ve işyerimdeki değerli çalışma arkadaşlarıma teşekkür etmek isterim. Okuldaki arkadaşlarımın paylaştıkları fikirler, birlikte geçirdiğimiz akademik tartışmalar ve dayanışma ruhu, bu süreci benim için daha keyifli hale getirdi. Aynı şekilde, işyerimdeki mesai arkadaşlarımın anlayışı ve destekleri sayesinde akademik çalışmalarımı iş hayatımla dengeli bir şekilde yürütebildim. Hepsine içtenlikle teşekkür ederim.

Dilan BAKIR

İÇİNDEKİLER

KISALTMA LİSTESİ

AS	Araştırma soruları
BA	Bilgi Alma
BDM	Büyük Dil Modelleri
BioASQ	Biomedical Question Answer
BT	Bilgi Tabanı
CÇ	Cevap Çıkarma
CLEF	Cross-Language Evaluation Forum
DDİ	Doğal Dil İşleme
IR	Information Retrieval
KB	Knowledge Base
MOA	Makine Okuduğunu Anlama
MQA-metrik	Mistral Question Answer Metrik
NLP	Natural Language Processing
PMKSB	Popülasyon, Müdahale, Karşılaştırma, Sonuç ve Bağlam
QA	Question Answer
SC	Soru Cevaplama
SDS	Scimago Dergi Sıralaması
SLİ	Sistematik Literatür İncelemeleri
SQuAD	Stanford Question Answering Dataset
TREC	Text Retrieval Conference

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

Metin Analizi Tabanlı Soru Cevaplama Sistemleri İçin Model Tabanlı Değerlendirme Metriği

Dilan BAKIR

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Danışman: Prof. Dr. Mehmet S. AKTAŞ

Eş-Danışman: Doç. Dr. Beytullah YILDIZ

Bu tezde Soru Cevaplama (SC) sistemlerinin değerlendirme süreçlerinde karşılaşılan temel eksiklikler ele alınmaktadır. Geleneksel değerlendirme metrikleri, SC sistemlerinin ürettiği yanıtların semantik doğruluğunu ve bağlamsal uygunluğunu tam olarak yansıtmamaktadır. Mevcut metrikler genellikle yüzeysel benzerliklere odaklanırken, karmaşık metinlerin ve terminolojilerin değerlendirilmesinde yetersiz kalmaktadır. Bu bağlamda, SC sistemlerinin cevaplarının doğruluğunu daha etkin bir şekilde ölçmek için yeni bir değerlendirme yaklaşımına ihtiyaç duyulmaktadır. Tez kapsamında, SC sistemleri için mevcut değerlendirme yöntemleri analiz edilmiştir. Bu tez kapsamında insan yargılarına dayalı veri kümeleri oluşturulmuştur. Elde edilen veri kümeleri sadece yüzeysel benzerliklere dayanmamakta aynı zamanda anlamsal benzerliklere dayalı olarakta oluşturulmuştur. Literatürdeki çalışmalar, mevcut metriklerin yüzeysel yöntemlere dayandığını ortaya koymaktadır. SC sistemlerinde sorulara verilen detaylı veya yazım hatası içeren yanıtların değerlendirmesi yapıldığında yeterli kaliteye ulaşılamadığını ortaya koymaktadır. Bu tez kapsamında bu sorunları ortadan kaldırabilmek için model tabanlı bir değerlendirme metriği önerilmiştir. Önerilen

model tabanlı değerlendirme metriği için bu tez kapsamında, SC sistemlerinin yanıtlarını daha kapsamlı analiz edebilmek amacıyla, bir iş akışı tasarımı sunulmaktadır. Bu metrik ile birlikte, yanıtların semantik bütünlüğü ve bağlamsal uygunluğu değerlendirilerek geleneksel yöntemlerden daha yüksek doğruluk sağlamak hedeflenmektedir. Burdaki doğruluğun sağlandığını ortaya koyabilmek için, model tabanlı değerlendirme metriğinin insan değerlendirmeleriyle yüksek korelasyon gösterip göstermediği tez kapsamında irdelenmiştir. Bu çalışmada, SC sistemleri için geliştirilen değerlendirme metriğinin uçtan uca iş akışı detaylandırılmıştır. Metriklerin oluşturulmasında kullanılan veri kümeleri ve etiketleme süreçleri bu tez kapsamında detaylı olarak açıklanmaktadır. Bu tez kapsamında SC sistemlerinin değerlendirme süreçlerini daha güvenilir hale getirecek yeni bir değerlendirme yaklaşımı sunulmaktadır. Sonuç olarak, bu tez, SC sistemlerinin performansını daha adil ve objektif bir şekilde değerlendirmeye yönelik yeni bir metrik geliştirilmekte ve bu metriğin etkinliğini analiz edilmektedir. Bu çalışmada, SC sistemlerinin kullanım alanlarını genişletmek ve yanıt doğruluğunu artırmak için katkılar sunulmaktadır.

Anahtar Kelimeler: Soru cevap sistemleri, model tabanlı değerlendirme, semantik benzerlik, doğal dil işleme

ABSTRACT

Model Based Evaluation Metric For Text Analysis Based Question Answering Systems

Dilan BAKIR

Department of Computer Engineering

Doctor of Philosophy Thesis

Supervisor: Prof. Dr. Mehmet S. AKTAŞ

Co-supervisor: Assoc. Prof. Dr. Beytullah YILDIZ

This thesis addresses the fundamental deficiencies encountered in the evaluation processes of Question Answering (SC) systems. Traditional evaluation metrics do not fully reflect the semantic accuracy and contextual relevance of the answers produced by SC systems. While existing metrics generally focus on superficial similarities, they are insufficient in evaluating complex texts and terminologies. In this context, a new evaluation approach is needed to measure the accuracy of answers of SC systems more effectively. Within the scope of the thesis, existing evaluation methods for SC systems have been analyzed. Within the scope of this thesis, data sets based on human judgments have been created. The obtained data sets are not only based on superficial similarities, but also on semantic similarities. Studies in the literature reveal that existing metrics are based on superficial methods. It reveals that sufficient quality cannot be achieved when detailed or spelling-error answers given to questions in SC systems are evaluated. Within the scope of this thesis, a model-based evaluation metric is proposed to eliminate these problems. In this thesis, a workflow design is presented for the proposed model-based evaluation metric in order to analyze the responses of SC systems more

comprehensively. With this metric, it is aimed to provide higher accuracy than traditional methods by evaluating the semantic integrity and contextual appropriateness of the responses. In order to demonstrate that the accuracy here is achieved, it is examined in the thesis whether the model-based evaluation metric shows a high correlation with human evaluations. In this study, the end-to-end workflow of the evaluation metric developed for SC systems is detailed. The datasets and labeling processes used in the creation of the metrics are explained in detail in this thesis. A new evaluation approach that will make the evaluation processes of SC systems more reliable is presented in this thesis. As a result, this thesis develops a new metric to evaluate the performance of SC systems in a more fair and objective way and analyzes the effectiveness of this metric. In this study, contributions are made to expand the areas of use of SC systems and increase response accuracy.

Keywords: Question answer systems, model based evaluation, semantic similarity, natural language processing

1
GİRİŞ

Bilgi çağının gelişmesi ile birlikte, teknolojik yenilikler sayesinde birçok alanda devrim yaratılmıştır. Soru Cevaplama (SC) sistemleri, kullanıcıların geniş veri tabanları, dokümanlar arasında gezinmelerine olanak sağlayan önemli araçlar haline gelmiştir. Bu sistemler, araştırma süreçlerini hızlandırmak, temel kaynaklara evrensel erişimi sağlamak ve karmaşık sorunları çözmek için büyük bir potansiyel sağlamaktadır. SC sistemleri, özellikle müşteri odaklı büyük şirketler ve kullanıcı yoğunluğu yüksek platformlar tarafından tercih edilmektedir. Ancak, bu sistemlerin verimliliği ve sağladığı yanıtların doğruluğu, kullanıcı güvenini doğrudan etkileyen önemli bir faktör olmaya devam etmektedir. SC sistemlerinin önerdiği cevapların kalitesi kullanıcı güvenini kazanmak için özellikle büyük şirketler için çok önemlidir. Bu sistemlerinin cevaplarının değerlendirilmesi, genellikle kullanıcı geri bildirimlerine veya sınırlı metriklere dayanmaktadır. Bu durum, önerilen yanıtların kalitesini değerlendirme sürecinde önemli eksikliklere yol açmaktadır. Mevcut değerlendirme metriklerinin birçoğu, özellikle karmaşık metinlerin işlenmesi söz konusu olduğunda, teorik yeterlilikleri ile gerçek dünyadaki performansları arasında bir uyumsuzluk göstermektedir. Bu bağlamda, SC sistemlerinin değerlendirme süreçleri için daha sağlam, objektif ve kapsamlı metriklere ihtiyaç duyulmaktadır. Model tabanlı değerlendirme metrikleri, bu eksiklikleri gidermek için önemli bir potansiyel sağlamaktadır.

1.1 Literatür İncelemesi

Bu bölümde, SC sistemlerinin değerlendirilmesi noktasında bileşenleri, değerlendirme metrikleri ve geliştirme yöntemleri incelenmiş ve bu alandaki önemli çalışmalar özetlenmiştir. Üretilen yanıtların doğruluğunu değerlendirmek oldukça zordur. Bunun temel nedeni, yanıtların serbest formda olmasıdır[1]. N-gram benzerliği gibi mevcut metrikler, tüm tokenlara eşit ağırlık vererek doğru ve yanlış yanıtları yeterince ayırt edememektedir. Bu sorunu ele almak için önerilen KPQA-metrik, kilit ifadeleri öne çıkararak tokenlara farklı ağırlıklar

atamış ve bu sayede yanıtların ana anlamını doğru bir şekilde yakalama yeteneğini artırmıştır. İnsan yargılarına dayalı veri kümelerinde yapılan deneyler, bu yeni metriğin insan değerlendirmeleriyle önemli ölçüde daha yüksek bir korelasyona sahip olduğunu göstermiştir. Özetleme sistemlerinde, kaynak belgeyle tutarlı olmayan içerik üretimi sıkça karşılaşılan bir sorundur. FEQA metriği, bu tür sistemlerin doğruluğunu değerlendirmek için geliştirilmiştir[2]. Özetlerden üretilen soru-yanıt çiftlerini kaynak belgeden yanıtlarla karşılaştırarak uyumsuzlukları tespit etmektedir. İnsan değerlendirmeleriyle yüksek korelasyon gösteren FEQA, özellikle soyutlama düzeyi yüksek özetlerde diğer metriklerden daha etkili sonuçlar vermiştir. Soru üretme sistemlerinin değerlendirilmesinde, referans sorularına olan bağımlılık ve semantik olarak benzer olmayan doğru soruların cezalandırılması gibi sorunlar bulunmaktadır[3]. RQUGE metriği, aday sorunun bağlama uygunluğu değerlendirilerek bu sınırlamaları aşmıştır. İnsan yargılarıyla yüksek korelasyona sahip olan RQUGE, referansa dayalı olmayan bir metrik olarak, sistem performansını değerlendirme süreçlerine yenilikçi bir bakış açısı kazandırmıştır. Chirag ve arkadaşları yanıt kalitesinin değerlendirilmesi için kullanıcı değerlendirmelerine dayalı kapsamlı bir çalışma gerçekleştirmiştir[4]. Yanıtlar, 13 farklı kriter üzerinden insan değerlendiriciler tarafından puanlanmış ve bu değerlendirmeler, kullanıcıların algıladığı kalite ile yüksek uyum göstermiştir. Ayrıca, çeşitli özelliklere dayalı olarak en iyi yanıtı seçmek için makine öğrenimi modelleri eğitilmiştir. Bu araştırma, kullanıcı profillerinin SC sistemlerinde içerik kalitesini değerlendirme ve tahmin etmede kritik bir rol oynayabileceğini ortaya koymuştur. QAEval, bir özetin içerik kalitesini değerlendirmek için referans metinle bilgi örtüşmesini ölçen bir metriktir[5]. Mevcut metriklere kıyasla daha üstün performans sergileyen QAEval, özetlerin değerlendirilmesinde SC tabanlı yaklaşımın faydalarını vurgulamaktadır. Deneysel analizler, QAEval'in mevcut sınırlamaları aştığını ve özetleme sistemlerinin kalitesini değerlendirme açısından daha güvenilir bir araç sunduğunu göstermiştir. SC sistemleri, soru sınıflandırma, bilgi getirme ve yanıt çıkarma aşamalarından oluşur[6]. Bu sistemlerde performans değerlendirmesi, kullanılan metrikler ve uygulama yöntemlerine bağlıdır. Yapılan çalışma, bu temel bileşenlere dair kapsamlı bir inceleme sunmuş ve SC sistemlerinin performansını artırmaya yönelik çeşitli metrik ve teknikleri tanıtmıştır. Lexical bazlı metriklerin, semantik olarak doğru ancak kelime örtüşmesi olmayan yanıtları yanlış değerlendirme sorununu ele alan SAS metriği, anlam benzerliği temelli bir değerlendirme sunmuştur[7]. Transformer modellerinden yararlanan SAS, insan yargılarıyla daha yüksek korelasyona sahiptir ve doğru yanıtların adil bir şekilde değerlendirilmesine olanak tanımaktadır. Son yıllarda SC alanında çok sayıda veri kümesi ve değerlendirme metriği önerilmiştir[8]. Bu çalışmada, SC sistemlerine yönelik 47 benchmark veri

kümesi incelenmiş ve uygulama bazlı yeni bir taksonomi önerilmiştir. Ayrıca, SC görevleri için kullanılan 8 farklı değerlendirme metriği özetlenmiş ve bu alandaki mevcut eğilimler ile gelecekteki çalışmalar için öneriler sunulmuştur.

1.2 Araştırma Probleminin Tanımı

Günümüzde SC sistemleri, kullanıcılara doğru ve anlamlı yanıtlar sağlayarak bilgiye erişim süreçlerini kolaylaştıran kritik bir teknolojik bileşen haline gelmiştir. Bu sistemlerin başarısı, ürettikleri yanıtların kalitesine ve doğruluğuna bağlıdır. Ancak, mevcut değerlendirme yöntemlerinin, SC sistemlerinin verdiği cevapların gerçek dünya senaryolarındaki etkinliğini ölçme konusunda yeterli olmadığı gözlemlenmektedir. Geleneksel metrikler, genellikle yüzeysel kelime eşleşmelerine dayandığından, anlamsal doğruluğu ve bağlamsal uygunluğu değerlendirmede yetersiz kalmaktadır. Bu eksiklikleri gidermek amacıyla model tabanlı değerlendirme metrikleri geliştirilmesi kritik bir gereklilik haline gelmiştir. Model tabanlı metrikler, doğrudan kelime düzeyinde bir eşleşme yerine, derin öğrenme teknikleri ve anlamsal analiz yöntemleri kullanarak SC sistemlerinin çıktılarının kalitesini daha nesnel bir biçimde değerlendirebilmektedir. Bununla birlikte, bu tür bir metriğin geliştirilmesi, veri kümesi tasarımı, etiketleme süreci ve model eğitimi gibi birden fazla bileşeni içeren karmaşık bir süreci gerektirmektedir. Dolayısıyla, model tabanlı bir değerlendirme metriğinin oluşturulma sürecine ilişkin sistematik bir çalışma ihtiyacı doğmaktadır. Bu tez çalışması, SC sistemlerinin yanıtlarını değerlendirmek için kullanılan mevcut değerlendirme metriklerin eksiklikleri üzerine çalışılarak model tabanlı bir değerlendirme metriği geliştirmeyi ve tasarım sürecini detaylandırmayı amaçlamaktadır. Araştırma problemimiz, anlamsal odaklı, insan yargısıyla yüksek korelasyon gösteren yeni bir değerlendirme yaklaşımına ihtiyaç bulunmaktadır. Böyle bir metriğin geliştirilebilmesi için; veri kümesi tasarımı, etiketleme süreçlerinin doğruluğu ve çıktı analizi gibi çok boyutlu unsurların sistematik şekilde ele alınması gerekmektedir. Bu araştırma problemi, söz konusu gereksinimleri karşılayacak bağlam-duyarlı, adil ve nesnel bir değerlendirme metrik yaklaşımının nasıl geliştirilebileceği üzerine odaklanmaktadır. Bu bağlamda, aşağıdaki araştırma soruları ele alınmaktadır:

- AS1: Bağlam-duyarlı, adil ve nesnel değerlendirme metriği geliştirmek için uygun veri kümesi nasıl oluşturmalı ve etiketleme süreci nasıl yapılandırılmalıdır?
- AS2: SC sistemlerinin çıktılarının semantik doğruluğunu değerlendirecek bir metrik geliştirmek için en uygun iş akışı nasıl olmalıdır?

- AS3: Bağlam-duyarlı, adil ve nesnel değerlendirme metrik yapısının, insan yargılarıyla korelasyonu nasıl olmalıdır ve bu yapı geleneksel metriklere kıyasla ne derece etkilidir?

Bu araştırma sorularının yanıtlanması, SC sistemlerinin değerlendirme süreçlerinde çok daha doğru, nesnel ve etkin bir yaklaşım geliştirilmesine katkı sağlamaktadır. Ayrıca, tez çalışması sonucunda elde edilecek bulgular, SC sistemlerinin değerlendirme metotlarının daha güvenilir hale getirilmesine olanak sağlamaktadır. Bu bulgular ayrıca değerlendirme metriklerinin geliştirilmesine yönelik akademik ve sektörel çalışmalar için temel oluşturmaktadır.

1.3 Tezin Amacı

SC sistemlerinin ürettiği cevapları değerlendiren geleneksel metrikler, yanıtların yüzeysel benzerliklerine odaklanarak, metinlerin semantik doğruluğunu ve bağlamsal uygunluğunu göz ardı etmektedir. Bu tezin amacı, SC sistemlerini değerlendiren metriklerin nasıl çalışması gerektiği üzerine iş akışının çıkarılması ve geleneksel metriklerin eksik kalan önemli boşluklarını doldurmaktır. SC sistemlerinin etkin kullanımı, yalnızca doğru sonuçların sağlanmasıyla değil, aynı zamanda bu sistemlerin performanslarının nesnel ve güvenilir bir şekilde değerlendirilmesiyle mümkündür. Eğer bir sistem kaliteli bir cevap üretmezse kullanıcı memnuniyetini olumsuz etkileyebilir. Bunun için cevapları değerlendirme sistemine ihtiyaç duyulmuştur. SC sistemlerinde önerilen cevaplar gerçek cevapların eşanlamlı kelimelerini de önerebilir veya gerçek cevaplarda yazım hataları bulunabilir. Bu durumda geleneksel metrikler eşanlamlı kelimeleri ve yazım hatalarını yakalama noktasında yetersiz kaldıkları için yeni bir değerlendirme sistemlerine ihtiyaç duyulmaktadır. Mevcut değerlendirme yöntemleri genellikle kullanıcı geri bildirimleri ve temel yanıt analizine dayanmakta, ancak bu yöntemler, özellikle karmaşık terminolojiler ve kavramların ele alındığı bağlamlarda yetersiz kalmaktadır. SC sistemleri gerçek cevaptan daha uzun ve detaylı bir cevap önerdiği zaman geleneksel değerlendirme metrikleri gerçek cevap ile tahmin edilen cevabı karşılaştırdığında başarısız sonuç üretmektedir fakat tahmin edilen cevap doğrudur tek farkı daha detaylı oluşudur. Detaylı önerilen cevaplarda geleneksel metodlar yetersiz kalmaktadır. Bu tez çalışmasında SC sistemlerinin ürettiği cevabı değerlendirmek için üretilmesi gerekilen metriğinin uctan uca çalışma mekanizmasını gösteren iş akış topolojisinin çıkarılma gereksinimi ortaya çıkmaktadır.

1.4 Hipotez

Geleneksel değerlendirme metrikleri, SC sistemlerinin ürettiği yanıtların kalitesini ölçmede belirli sınırlılıklar içermektedir. Bu metrikler genellikle yüzeysel kelime eşleşmelerine dayanarak değerlendirme yapmakta ve metinlerin semantik doğruluğunu, bağlamsal bütünlüğünü yeterince göz önünde bulundurmamaktadır. Bu eksiklikler, SC sistemlerinin önerdiği yanıtların gerçek kullanıcı ihtiyaçlarını ne ölçüde karşıladığını doğru bir şekilde belirlemeyi zorlaştırmaktadır. Dolayısıyla, SC sistemlerinin yanıtlarını daha kapsamlı bir şekilde değerlendiren, dilin anlamsal yapısını dikkate alan model tabanlı bir değerlendirme metriğine ihtiyaç duyulmaktadır. Bu çalışma, SC sistemlerinin yanıtlarının değerlendirme sürecini iyileştirmek için model tabanlı metriğin geliştirilmesini önermektedir. Hipotezimiz, SC sistemlerinin önerdiği yanıtların kalitesini artırabilmek için değerlendirme sürecinde kullanılan metriklerin, yalnızca yüzeysel düzeyde değerlendirme yapmak yerine derin öğrenme ve anlamsal analiz tekniklerini kullanarak yeniden yapılandırılması gerektiğidir. Bu bağlamda, geliştirilecek model tabanlı metrik, SC sistemlerinin çıktılarının gerçek dünya senaryolarındaki etkinliğini daha doğru bir şekilde ölçebilecek, böylece sistemlerin iyileştirilmesine yönelik daha sağlam geri bildirimler sunulmaktadır.

Hipotezin doğrulanabilmesi için, model tabanlı değerlendirme metriği oluşturulurken yüksek kaliteli bir veri kümesine ihtiyaç duyulmaktadır. Veri kümesinin, çeşitli soru türlerini kapsayan, farklı dil yapılarını ve kavramsal zenginliği içeren geniş bir yapıya sahip olması gerekmektedir. Ayrıca, veri kümesinin etik ve nesnel bir şekilde etiketlenmesi, metriğin doğruluk ve güvenilirlik açısından güçlü bir temel oluşturmasını sağlamaktadır. Veri kümesinin oluşturulması ve etiketleme süreçlerinin objektif kriterlere dayandırılması, hipotezimizin başarısını destekleyen kritik bileşenlerden biridir. Bu bağlamda, model tabanlı metriğin başarısı, SC sistemlerinin ürettiği yanıtların hem bağlamsal uygunluk hem de anlamsal doğruluk açısından nasıl değerlendirildiğini belirlemek için deneysel analizlerle test edilmektedir. Geliştirilen metriğin, geleneksel metriklerle kıyaslandığında daha yüksek bir değerlendirme doğruluğu sağlayıp sağlamadığı incelenmektedir. Özellikle, SC sistemlerinin eşanlamlı kelimeleri ve anlam bakımından yakın ifadeleri önerme kapasitesi göz önüne alındığında, yeni metriğin bu tür bağlamsal varyasyonları yakalayabilme yeteneği değerlendirilmektedir.

Sonuç olarak, bu hipotez çerçevesinde yürütülecek çalışma, SC sistemlerinin kalite değerlendirme sürecinde önemli bir yenilik sunmayı hedeflemektedir. Model tabanlı değerlendirme metriği geliştirilerek, SC sistemlerinin önerdiği yanıtların

doğruluğunu daha etkin bir şekilde ölçmek ve sistemlerin iyileştirilmesine katkıda bulunmayı amaçlamaktadır. Böylece, SC sistemlerinin daha güvenilir, kullanıcı odaklı ve etkin bir şekilde çalışması sağlanarak, bilgiye erişim süreçlerinin daha verimli hale getirilmesi hedeflenmektedir.

1.5 Tezin Kapsamı

Bu tezin kapsamını tanımlamak için, SC sistemlerinde değerilendirme metriklerinin gerekliliklerini ve uygulama kullanım durumlarını ana hatları ile açıklanmaktadır.

1.5.1 Soru Cevaplama Gereksinimleri

SC sistemleri, büyük ölçekli veri kaynaklarından en uygun yanıtları üretmeyi amaçlayan yapay zeka tabanlı çözümler olarak gelişim göstermektedir. Ancak, bu sistemlerin başarısını ölçmek ve güvenilirliğini artırmak için kullanılan değerlendirme metrikleri, geleneksel yöntemlerle sınırlı kalmaktadır. Mevcut değerlendirme süreçleri, genellikle kelime tabanlı benzerlik ölçümlerine dayalı olup, semantik anlam bütünlüğünü yeterince dikkate almamaktadır. Özellikle, SC sistemlerinin sunduğu yanıtların uzunluk, bağlam uygunluğu ve anlamsal doğruluk gibi faktörlere göre değerlendirilmesi, mevcut metrikler açısından önemli bir zorluk teşkil etmektedir. Bu durum, SC sistemlerinin doğruluğunu ölçmek için daha karmaşık ve bağlamsal farkındalığa sahip yeni metriklerin geliştirilmesini gerekli kılmaktadır.

Bu tez çalışmasında, SC sistemlerinin farklı veri kümeleri ve platformlar üzerindeki performansını değerlendirmek için model tabanlı bir değerlendirme metriği geliştirilmesi amaçlanmaktadır. Bu doğrultuda, öncelikle SC sistemlerinin yanıtlarını ölçmek için kullanılan geleneksel metriklerin sınırlılıkları analiz edilmekte, ardından yeni bir değerlendirme sürecinin nasıl tasarlanması gerektiği ele alınmaktadır. Geliştirilecek olan model tabanlı metriğin etkinliği, özel olarak oluşturulan veri kümeleriyle test edilerek, geleneksel metriklerle karşılaştırılmaktadır. Bu süreçte, hem veri kümesinin oluşturulması hem de yanıtların etiketlenme aşamalarında nesnel ve tutarlı kriterler belirlenmesi gerekmektedir. Böylece, önerilen metriğin SC sistemlerinin ürettiği yanıtların kalitesini daha kapsamlı ve güvenilir bir şekilde değerlendirmesi sağlanmaktadır. SC sistemlerinin değerlendirilmesi yalnızca teknik doğruluk açısından değil, aynı zamanda kullanım senaryolarına uygunluk ve kullanıcı güvenini sağlama açısından da önemlidir. Geleneksel metriklerin eksiklikleri göz önüne alındığında, semantik anlam ilişkilerini ve bağlamsal doğruluğu dikkate alan model tabanlı

değerlendirme metriklerinin gerekliliği ortaya çıkmaktadır. Bu bağlamda, önerilen MQA-metrik modeli, semantik açıdan doğru ancak kelime bazında farklı ifadeler içeren yanıtları daha doğru bir şekilde değerlendirebilme kapasitesine sahip olacak şekilde tasarlanmaktadır. Aynı zamanda, yazım hatalarının ve detaylı cevapların geleneksel metrikler tarafından düşük puanlanmasının önüne geçerek, daha adil bir değerlendirme süreci oluşturulması hedeflenmektedir. Bu tez, SC sistemlerinin teknik gereksinimlerine ek olarak, değerlendirmenin objektifliğini ve güvenilirliğini sağlayacak ölçütleri belirleyerek, model tabanlı metriklerin gerekliliğini ortaya koymaktadır.

1.5.2 Uygulama Kullanım Alanları

SC sistemlerinin ürettiği yanıtları değerlendirmek amacıyla geliştirilen metriklerin, uçtan uca bir süreç çerçevesinde titizlikle tasarlanması gerekmektedir. Değerlendirme metriklerinin geliştirilme süreci, yalnızca teknik doğruluk açısından değil, aynı zamanda hedeflenen uygulama alanlarının gereksinimlerini karşılayacak şekilde şekillendirilmelidir. Bu bağlamda, SC sistemlerinin kullanım alanlarını ve bu sistemler için gerekli olan temel değerlendirme kriterlerini aşağıda özetlenmektedir.

1.5.2.1 Arama Motorları

Arama motorları, kullanıcıların doğrudan sorduğu soruları veya dolaylı ifadelerle verdiği sorguları anlamalıdır. Sorgulardaki eş anlamlıları, mecazları, deyimleri ve bağlamsal anlamları çözümleyebilmelidir. SC sistemi, veritabanları, web sayfaları, dokümanlar ve çizge veritabanları gibi geniş bilgi kaynaklarına erişebilmelidir. Kullanıcının sorusuyla en alakalı bilgileri önceliklendirmelidir. En uygun bilgiyi bulup bu bilgi üzerinden cevabı çıkarabilmelidir. Arama motorları, bilgiye erişimi kolaylaştıran teknolojiler olarak birçok alana entegre edilebilir. E-ticarette ürün arama ve kişiselleştirilmiş öneriler sunarken, sağlık sektöründe tıbbi bilgi erişimi ve hasta verilerinin analizi için kullanılır. Eğitim alanında ders materyali arama ve akademik araştırmalarda etkiliyken, finans sektöründe piyasa analizi ve risk değerlendirmesinde önemli rol oynar. Hukukta içtihat ve mevzuat aramalarında, insan kaynaklarında aday filtreleme ve içerik yönetiminde, medya ve eğlence sektöründe içerik keşfi ve öneri motorlarında yaygın şekilde kullanılır. Bu entegrasyonlar, arama motorlarını sektörlerin bilgi işleme ve kullanıcı deneyimi süreçlerinde vazgeçilmez bir unsur haline getirir.

1.5.2.2 Sohbet robotları

SC sistemlerinin sohbet robotlarındaki kullanım gereksinimleri, kullanıcıların doğal dilde sorularına hızlı, doğru ve bağlamsal olarak uygun yanıtlar sunmayı amaçlar. Sohbet robotları, e-ticaret, sağlık, eğitim, finans, insan kaynakları, turizm, medya, kamu hizmetleri, teknoloji, gayrimenkul ve enerji gibi birçok alana entegre edilebilir ve kullanıcı deneyimini iyileştiren etkili çözümler sunar. E-ticarette müşteri hizmetleri, ürün önerileri ve kampanya yönetiminde kullanılırken, sağlık sektöründe hasta desteği, randevu planlama gibi hizmetler sunar. Eğitimde öğrencilere ders materyali önerileri ve dil öğrenim desteği sağlarken, finans sektöründe müşteri işlemleri, yatırım danışmanlığı ve dolandırıcılık tespiti gibi işlevler görür. İnsan kaynaklarında aday tarama, çalışan destek ve eğitim önerileri sunarken; seyahat ve turizmde rezervasyon işlemleri, gezi önerileri gibi ihtiyaçları karşılar. Medya ve eğlence sektöründe içerik önerileri ve sosyal medya etkileşiminde etkilidir. Sohbet robotları, farklı sektörlerin özel ihtiyaçlarına uygun şekilde özelleştirilerek süreçleri optimize eder ve verimliliği artırır.

1.5.3 Tez Katkıları

SC sistemlerinin ürettiği yanıtların doğruluğunu değerlendirmek için kullanılan geleneksel yöntemler, yazım hatalarının dikkate alınmaması, eş anlamlı kelimelerin tanınamaması ve detaylı yanıtların eksik değerlendirilmesi gibi temel sorunlarla karşı karşıyadır. Bu tez kapsamında, söz konusu eksiklikleri gidermek amacıyla model tabanlı bir değerlendirme metriği geliştirilmiş ve bu metriğin uygulanabilirliğini göstermek için bir prototip iş akış topolojisi oluşturulmuştur. Çalışmanın sağladığı katkılar aşağıda özetlenmektedir:

- SC sistemleri üzerine sistematik bir literatür taraması gerçekleştirilmiştir [9].
- Kullanıcıların sorularına hızlı ve etkili yanıt sağlayan SC sistemleri için kapsamlı bir iş akışı tasarımı önerilmiştir [10].
- Farklı platformlardan gelen ve çeşitli veri kümeleri ile eğitilmiş SC sistemlerini tek bir arayüz üzerinden entegre etmeye olanak tanıyan hibrit bir yazılım mimarisi geliştirilmiştir [11].
- Önerilen hibrit prototip yazılım, bağımsız SC sistemleri ile karşılaştırılmış ve başarımı değerlendirilerek uygulanabilirliği gösterilmiştir [11].
- Çalışma, SC sistemlerinin çoklu platform entegrasyonu konusunda bir yol haritası sunarak gelecekteki araştırmalara katkı sağlamaktadır [11].

- Geleneksel metriklerin eksikliklerini gidermek amacıyla semantik doğruluğa odaklanan yenilikçi bir model tabanlı değerlendirme metriği (MQA-metrik) geliştirilmiştir. Bu metriği eğitmek için, literatürdeki veri eksikliğini gidermek amacıyla squad-qametrik veri kümesi oluşturulmuştur [12].
- İnsan yargısına dayalı yüksek kaliteli veri kümeleri (squad-qametrik ve marco-qametrik) oluşturularak bu veri kümeleri üzerinde kapsamlı deneyler gerçekleştirilmiştir [12].
- Geliştirilen MQA-metrik, semantik anlama ve bağlamsal değerlendirme açısından geleneksel metrikler (BLEU, ROUGE, METEOR) ile karşılaştırıldığında daha yüksek doğruluk ve güvenilirlik sergilemiştir [12].
- Cevap uzunluğu ve gerçek cevap kalitesinin metrik performansına etkisi gibi sorunları ele alarak değerlendirme süreçlerini daha güvenilir hale getirmiştir [12].
- SC sistemlerinin değerlendirilmesine yönelik olarak semantik doğruluğa öncelik veren yeni bir standart önerilmiştir [12].
- Geliştirilen değerlendirme metriği, insan yargısı ile daha yüksek korelasyon göstererek SC sistemlerinin objektif ve güvenilir bir şekilde değerlendirilmesine katkı sunmaktadır [12].

1.6 Tezin Organizasyonu

Bu tez çalışmasının organizasyonu aşağıdaki gibidir. Bölüm 2, SC sistemlerini değerlendirmek için kullanılan metrikler tanıtılmaktadır. Bölüm 3, daha önce yürütülen sistematik literatür taraması çalışmasına, deneysel çalışmalar ve katkılar dahil olmak üzere referans vererek SC hakkında ayrıntılı arka plan bilgisi sağlamaktadır. Bölüm 4, SC sistemleri için iş akış topolojisinin nasıl olacağı ve prototip üzerinde uygulamalı gösterip deneysel sonuçlar gösterilmiştir. Bölüm 5, SC sistemlerine hibrit yaklaşım gösterilmiştir. Bölüm 6, SC sistemlerini değerlendirmek için kaliteli bir verikümesi oluşturup bu verikümesi ile model tabanlı metrik geliştirilmesi prototip üzerinde uygulanıp topolojinin oluşturulması ve deneysel sonuçların çıkartılması sağlanır. Bölüm 7, sonuçları sağlanmaktadır ve belirli katkılar için değerlendirmeler, tartışmalar ilgili bölümler yer almaktadır.

2
TEMEL KONSEPTLER

Bu bölümde, SC sistemlerinin temel kavramlarına dair arka plan bilgisi sunulmaktadır. Öncelikle, SC alanındaki temel yaklaşımlar ve bu yaklaşımların geliştirilmesinde kullanılan yöntemler ele alınmaktadır. Ardından, SC sistemlerinin çıktılarının değerlendirilmesinde kullanılan mevcut metrikler ayrıntılı olarak incelenmektedir. Büyük Dil Modellerinin (BDM), SC sistemlerinde oynadığı rolü ve sağladığı katkıları ele alınmaktadır. Bu kapsamda, BDM yapısal özellikleri, mimari farklılıkları ve SC süreçlerindeki performansları hakkında bilgi verilmektedir. Tez kapsamında yürütülen çalışmaların metodolojik altyapısını anlamak adına, bu modellerin seçim nedenleri ve sağladıkları avantajlar da tartışılmaktadır.

2.1 Soru Cevaplama

SC, uzun yıllardır Doğal Dil İşleme (DDİ) alanında merkezi bir çalışma alanı olmuştur. SC'nın amacı, doğal dilde sorulan soruları doğru bir şekilde cevaplayabilen sistemler tasarlamaktır.Gerekli araştırmalar yapıldığında, Bilgi Alma (BA) ve Cevap Çıkarma (CÇ) dan bir sürec mevcuttur. Doğal dilde verilen bir soru önce analiz aşamasından geçer. SC sistemlerinin en zorlu süreci anlamı doğru çıkarmaktır. Bu süreci iyileştirme üzerine çalışmalar sürmektedir. Geleneksel olarak SC sistemlerinde çoğunlukla kural tabanlı yöntemler tercih edilmiştir. Ancak yapay zekâ alanındaki ilerlemelerle birlikte, bu sistemlerde derin öğrenme yaklaşımlarının kullanımı yaygınlaşmıştır. Bu dönüşümün temel sebebi, kural tabanlı yöntemlerin yazım hatalarını algılamada ve eş anlamlı ifadeleri anlamlandırmada yetersiz kalmasıdır. Derin öğrenme yöntemleri ise bu sınırlamaların üstesinden gelerek önemli başarılar elde etmiş ve günümüzde pek çok çalışma bu yöntemler üzerine yoğunlaşmıştır.

2.2 Soru Cevaplama Sistemleri için Değerlendirme Metrikleri

Geçmişte SC sistemlerinin değerlendirilmesinde, doğrudan bu alana özgü olmayan ancak diğer doğal dil işleme görevlerinde kullanılan çeşitli metriklerden yararlanılmıştır. Bu bölümde, söz konusu metriklere yer verilmiştir. Bleu [13], Rouge [14] ve Meteor [15] gibi ölçütler, öncelikli olarak makine çevirisi ve otomatik özetleme sistemlerinin performansını değerlendirmek amacıyla geliştirilmiştir. Bu metrikler, üretilen cevaplar ile referans metinler arasındaki n-gram benzerliklerine odaklanır ve yüzeysel düzeyde bir kıyaslama sunar. Örneğin, Bleu metriği sistem çıktılarının insanlar tarafından hazırlanmış referans metinlerle benzerliğini ölçerek kalite değerlendirmesi yapar. Özellikle makine çevirisi ve DDİ gibi görevlerde yaygın olarak tercih edilir. Bleu puanı hesaplanmadan önce, sistem çıktıları ile karşılaştırılacak uygun referans metinlerin belirlenmesi gerekir. Sonrasında, bu metinler arasında n-gram düzeyinde bir analiz gerçekleştirilir ve nihai Bleu skoru bu analizlerin birleşimiyle elde edilir. Bleu'nün öne çıkan avantajları arasında otomatik olarak hesaplanabilir olması ve kullanım kolaylığı yer alır. Ancak bazı sınırlılıkları da vardır: örneğin, dil bağımlılığı, yeterli sayıda referans metin bulunmaması durumunda ortaya çıkan doğruluk sorunları, anlamsal uygunluk yerine yüzeysel benzerliklere odaklanması gibi problemler bu metriğin zayıf yönlerindendir. Bu nedenle, Bleu skoru genellikle başka değerlendirme yöntemleri ya da insan yargılarıyla desteklenerek kullanılır.

Rouge'un temel amacı, otomatik olarak oluşturulan metin özetlerini veya çevirilerini referans metinlerle karşılaştırmak ve benzerliklerini ölçmektir. Metin kalitesinin daha kapsamlı bir değerlendirmesini sunmak için iyi bilinen ROUGE metriğine dayanır. ROUGE, metni netliği, kesinliği ve önemi açısından değerlendirir, nihayetinde bir kalite puanı atar ve hataları belirler[14]. ROUGE'un gücü ayrıntılı yaklaşımında olsa da, insan önyargısı, değerlendirmeler için gereken daha fazla zaman ve kaynak ve çeşitli dillerde olası uygulama sorunları gibi zorlukları da sunar. Genel olarak, ROUGE metin oluşturma sistemlerinin daha derinlemesine bir değerlendirmesini sağlar ancak öznel ve dil özgü yönlerinden kaynaklanan dikkatli bir değerlendirme gerektirir. Meteor, özellikle makine çeviri sistemlerinin değerlendirilmesinde, DDİ alanında kullanılan bir değerlendirme metriğidir. Kelime seçimi ve düzenlemesini vurgular, benzer kelime dağarcığı kullanımını ve doğru dizilimi sağlamak için çevrilmiş metinleri referans metinlerle karşılaştırır. Meteor, çevrilen ve referans metinler arasındaki kelime eşleşmelerinin doğruluğunu değerlendirir ve benzerliği 0 ile 1 arasında bir ölçekte ölçer. Kapsamlı bir analiz sağlar ve güvenilirliği artırmak için insan tarafından çevrilmiş referans metinleri kullanır[15]. F1 puanı, hassasiyet ve geri çağırmanın harmonik ortalamasını temsil eden bir performans metriğidir; sınıflandırma problemlerinde

yaygın olarak kullanılır ve özellikle dengesiz veri kümelerinde faydalıdır. F1 puanı, bir modelin doğru pozitif tahminlerinin doğruluğunu (hassasiyet) ve bu doğru pozitif tahminlerin eksiksizliğini (geri çağırma) tek bir değerde birleştirir. Hassasiyet, modelin pozitif olarak sınıflandırdığı örnekler arasında gerçekten pozitif olanların oranını ifade eder. Yani, modelin yaptığı tüm pozitif tahminler içerisinde ne kadarının doğru olduğunu gösteren bir performans ölçüsüdür. Geri çağırma, doğru pozitif tahminlerin toplam gerçek pozitif örneklere oranıdır; başka bir deyişle, modelin tüm pozitif örnekleri ne kadar iyi tanımlayabildiğini gösterir. F1 skoru, kesinlik (precision) ve geri çağırma (recall) değerlerinin harmonik ortalamasıdır. Bu ortalama türü, iki değerden biri çok düşük olduğunda genel skoru da aşağı çeker, böylece dengesizliği ortaya koyar. F1 skoru, yalnızca doğru pozitif tahminleri değil, aynı zamanda yanlış negatifleri ve yanlış pozitifleri de göz önünde bulundurur. Özellikle sınıf dağılımının dengesiz olduğu durumlarda, modelin genel başarımını değerlendirmede etkili bir araçtır çünkü hem hassasiyeti hem de geri çağırmayı dengelemeye çalışır [12]. Tıbbi tanılama sistemleri veya spam tespiti gibi yanlış sınıflandırmanın ciddi sonuçlara yol açabileceği alanlarda F1 skoru kritik önem taşır. Bu metrik, model performansını tek bir değerle özetlese de, bağlama göre farklı önem derecelerine sahip olabilir; örneğin bazı senaryolarda hassasiyetin, diğerlerinde geri çağırmanın daha öncelikli olması mümkündür. Bu nedenle, F1 skorunun anlamı veri setinin yapısına ve uygulama alanına göre değerlendirilmelidir. Kosinüs benzerliği, özellikle bilgi erişimi ve metin madenciliği alanlarında kullanılan popüler bir benzerlik ölçüsüdür. İki vektörün yönleri arasındaki açıyı hesaplayarak, bu vektörlerin birbirine ne kadar benzer olduğunu değerlendirir. Bu özellik sayesinde, vektörlerin uzunluklarından bağımsız olarak çalışır ve özellikle belge karşılaştırmaları için uygundur [12].

Kosinüs benzerliği genellikle kolay yorumlanabilir sonuçlar üretir; 1'e yakın değerler yüksek benzerliği, sıfıra yakın değerler ise düşük benzerliği ifade eder. Ancak bu metrik, kelime sıklığı gibi bazı önemli metin özelliklerini dikkate almaz. Bu nedenle, önemli terimlerin ağırlığına dayanan yöntemlerle birlikte kullanılması önerilir. Belge benzerliğini değerlendirmede etkili olsa da, içerik yapısı karmaşık olan durumlarda sınırlı kalabilir. Jaccard benzerliği, iki küme arasındaki ortak öğelerin sayısının, bu kümelerin birleşimindeki toplam öğe sayısına oranı üzerinden hesaplanır. Genellikle öneri sistemleri ve küme analizleri gibi alanlarda kullanılır [16]. Basit yapısı sayesinde hızlıca uygulanabilir ve farklı büyüklükteki kümeler arasında da doğrudan karşılaştırma yapılmasına imkân tanır. Bu metrik, özellikle kullanıcıların tercih ettiği öğeleri karşılaştırmak gibi görevlerde faydalıdır. Ancak bazı sınırlamaları mevcuttur: örneğin, öğelerin tekrar sıklığını dikkate almaz; dolayısıyla bir öğe birden fazla kez geçse bile etki düzeyi değişmez.

Ayrıca, kümeler arasında çok az ortaklık olduğunda benzerlik oranı oldukça düşük çıkabilir. Büyük veri kümelerinde ise duyarlılığı azalabilir. Bu yüzden, Jaccard benzerliği kullanılmadan önce veri yapısının ve hedef uygulamanın uygunluğu analiz edilmelidir.

2.3 Büyük Dil Modelleri

Dil, insanların iletişim ve kendini ifade etme süreçlerinde temel bir rol oynar. Aynı zamanda, makinelerle etkileşimde önemli bir araçtır. Çeviri, özetleme, bilgi erişimi ve diyalog gibi karmaşık dil görevlerini yerine getirebilecek makineler için genel modellerin gerekliliği her geçen gün artmaktadır. Son yıllarda, transformer mimarileri, artan hesaplama gücü ve büyük ölçekli veri setlerinin kullanılabilirliği sayesinde dil modellerinde önemli ilerlemeler kaydedilmiştir. Bu gelişmeler, BDM insan seviyesine yakın performans göstermesini mümkün kılmıştır[16].

DDİ, istatistiksel yöntemlerden sinirsel dil modellemeye ve ardından önceden eğitilmiş dil modellerinden BDM'lere evrilmiştir[16]. Geleneksel dil modelleri, belirli bir görev için özel olarak eğitilirken, önceden eğitilmiş modeller geniş metin veri setleri üzerinde kendi kendine denetimli bir şekilde eğitilir ve çok sayıda DDİ görevine uyarlanabilir genel temsiller öğrenir. Daha büyük ölçekli önceden eğitilmiş modellerin performans avantajları, model parametrelerinin ve veri seti büyüklüğünün artırılmasıyla BDM'lere dönüşmesine yol açmıştır. BDM'ler, sıfır ve az atışlı (few-shot) öğrenme gibi yeteneklerinin yanı sıra, akıl yürütme, planlama ve karar verme gibi insan benzeri özellikler sergileyebilmektedir. Bu özellikler, büyük ölçekli veri ve model yapılarından kaynaklanmaktadır. BDM'ler, bu özellikleri sayesinde çoklu mod, robotik, araç manipülasyonu, SC sistemleri ve otonom ajanlar gibi çeşitli alanlarda yaygın olarak kullanılmaktadır. Ayrıca, görev odaklı eğitim ve daha etkili yönlendirme teknikleri gibi iyileştirmelerle bu modellerin yetenekleri artırılmaktadır[16].

BDM'lerin eğitim ve çıkarım süreçleri, yavaşlık, yüksek donanım gereksinimleri ve maliyetler gibi sınırlamalar içermektedir. Bu nedenle, daha iyi mimariler ve eğitim stratejileri geliştirme ihtiyacı doğmuştur. Parametre verimli ince ayar, model budama, kuantizasyon, bilgi distilasyonu ve bağlam uzunluğu interpolasyonu gibi yöntemler, BDM'lerin daha verimli kullanılmasını sağlamak için yaygın olarak araştırılmaktadır[16]. Bu yaklaşımlar, BDM sunduğu avantajları optimize ederek daha geniş bir kullanım alanı sunmayı hedeflemektedir. BDM eğitiminde temel adımlardan biri tokenizasyondur. Bu adım, metni parçalanamayan birimler olan tokenlara ayırır. Tokenlar, karakter, alt kelime, sembol veya kelime olabilir. Yaygın

kullanılan tokenizasyon yöntemleri arasında WordPiece, Byte Pair Encoding ve UnigramLM bulunur. Transformer mimarisi giriş dizilerini paralel işler, ancak pozisyon bilgisi doğal olarak yakalanmaz. Bu eksikliği gidermek için pozisyonel kodlama (positional encoding) tanıtılmıştır. Göreceli veya öğrenilen pozisyonel kodlamalar kullanılabilir. Göreceli kodlamalar arasında Alibi ve RoPE öne çıkar [17]. Alibi, yakın tokenlara öncelik vererek dikkat skorlarını konumlar arası mesafeye göre ayarlar. RoPE ise sorgu ve anahtar temsillerini, tokenların mutlak pozisyonlarına bağlı bir açıyla döndürerek mesafeye dayalı bir kodlama sağlar. Transformer mimarisindeki dikkat mekanizması (attention), giriş tokenlarına önem derecelerine göre ağırlık verir. Self-attention aynı bloktan gelen sorgu, anahtar ve değerler ile çalışırken, cross-attention encoder-decoder mimarilerinde kullanılır. Uzun dizilerde zaman karmaşıklığını azaltmak için Sparse Attention ve Flash Attention gibi yöntemler geliştirilmiştir[17].

3
SİSTEMATİK LİTERATÜR TARAMASI

3.1 Genel Yaklaşım

Günümüzde teknolojinin hızla gelişmesiyle birlikte verinin rolü giderek daha kritik hâle gelmiştir. Artan veri hacmiyle birlikte, bu verilerden anlamlı bilgiler elde etmek ve bu bilgileri etkin biçimde işleyebilmek amacıyla Soru-Cevap (SC) sistemleri geliştirilmiştir. SC sistemleri, kullanıcılardan alınan sorgular doğrultusunda, ilgili veriler arasından en uygun cevabı sunmayı hedefler. Bu sistemler; arama motorları, sohbet robotları (chatbotlar) gibi çeşitli uygulama alanlarına sahiptir ve kullanım amaçlarına göre farklılık gösterir. SC sistemleri temelde Bilgi Erişimi (BA), Cevap Çıkarımı (CÇ) ve Doğal Dil İşleme (DDİ) gibi disiplinlerin kesişiminde yer alır. Başlangıçta, arama motorları kullanıcıların doğal dilde yazdıkları belgeleri sorguya uygun bilgi olarak sunmayı amaçlarken, zamanla sistemlerden doğrudan sorulara yanıt vermeleri beklenmeye başlanmıştır. Bu doğrultuda SC alanında pek çok yöntem, yaklaşım ve veri kümesi üzerine çalışmalar gerçekleştirilmiştir. Dolayısıyla, SC sistemlerinin mevcut durumunu bütüncül biçimde ortaya koyan kapsamlı analizlere olan ihtiyaç giderek artmaktadır. Bu literatür taraması, 2000 ile 2022 yılları arasında SC alanında kullanılan araştırma eğilimlerini, veri kümelerini, yöntemlerini ve çerçevelerini belirlemeyi ve analiz etmeyi amaçlamaktadır.

3.2 Araştırma Soruları

Araştırma soruları (AS), çalışmanın odak noktasını belirginleştirmek ve araştırmanın sistematik bir şekilde ilerlemesini sağlamak amacıyla titizlikle oluşturulmuştur. Bu soruların formülasyonu sırasında Popülasyon, Müdahale, Karşılaştırma, Sonuç ve Bağlam (PMKSB) kriterlerinden yararlanılmıştır [18]. Bu yaklaşım, araştırma sorularının yapılandırılmasında metodolojik bir çerçeve sunmaktadır. Tablo 3.1, araştırma sorularının PMKSB yapısına göre düzenlenmiş halini göstermektedir.

Tablo 3.1 PMKSB'nin Özeti

Popülasyon	Soru Cevaplama, Doğal Dil İşleme, Bilgi Alma
Müdahale	Arama Motoru, Cevap Çıkarma, Pasaj Alma, Doküman Alma, Modeller, Metodlar, Teknikler, Veri Kümeleri
Karşılaştırma	n/a
Sonuç	Soru cevaplamanın tahmini doğruluğu, Başarılı soru cevaplama yöntemleri
Bağlam	Sanayi ve akademide çalışmalar, Küçük ve büyük veri kümeleri

Bu literatür taramasıyla ele alınan araştırma soruları ve motivasyonları Tablo 3.2'de gösterilmektedir.

Tablo 3.2 Literatür Taramasında Araştırma Soruları

ID	Araştırma Sorusu	Motivasyon
AS1	En önemli Soru Cevap dergisi hangisidir?	Soru Cevaplama alanındaki en önemli dergileri belirleyin
AS2	Soru Cevaplama alanında en aktif ve etkili araştırmacılar kimlerdir?	Soru Cevaplama araştırma alanında en aktif ve etkili araştırmacıları belirleyin
AS3	Soru Cevaplama alanında araştırmacılar ne tür araştırma konularını seçiyorlar?	Soru Cevaplamada araştırma konularını ve eğilimleri belirleyin
AS4	Soru Cevaplamada en çok hangi tür veri kümeleri kullanılıyor?	Soru Cevaplamada yaygın olarak kullanılan veri kümelerini tanımlayın
AS5	Soru Cevaplama için ne tür yöntemler kullanılıyor?	Soru Cevaplama yöntemi için fırsatları ve eğilimleri belirleyin
AS6	Soru Cevaplama için en sık kullanılan yöntemler nelerdir?	Soru Cevaplama için en çok kullanılan yöntemleri belirleyin
AS7	Soru Cevaplama için hangi yöntem en iyi performansı gösterir?	Soru Cevaplamada en iyi yöntemi belirleyin
AS8	Soru Cevaplama için hangi iyileştirmeler öneriliyor?	Soru Cevaplama da iyileştirmeler için önerilen yöntemleri tanımlayın
AS9	Soru Cevaplama için hangi tür çerçeveler öneriliyor?	Soru Cevaplamada en çok kullanılan çerçeveleri belirleyin

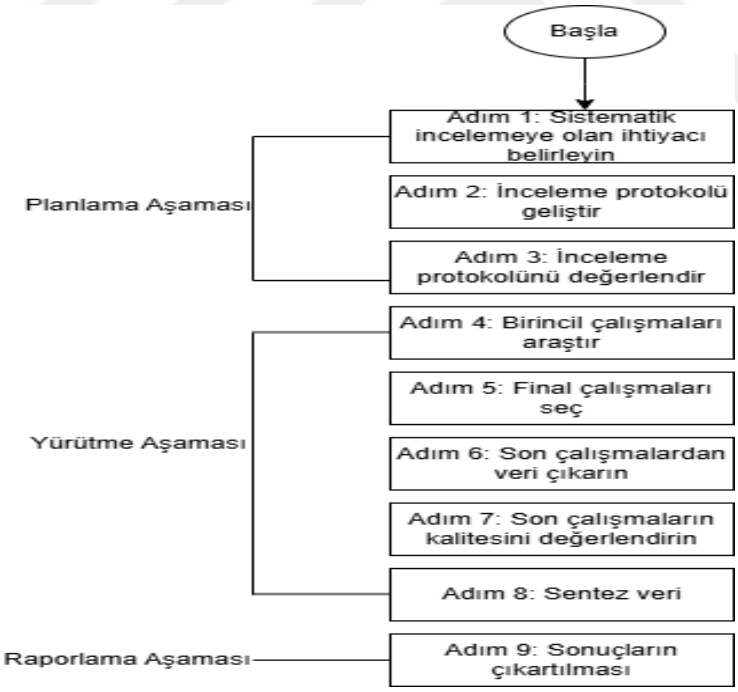
Son çalışmalardan, AS4 ila AS9'u yanıtlamak için SC yöntemleri, çerçeveleri ve veri kümeleri çıkarılır. Daha sonra, SC yöntemleri, çerçeveleri ve veri kümeleri, hangilerinin SC de (AS4 ila AS9) önemli yöntemler, çerçeveler ve veri kümeleri

olduğunu ve hangilerinin olmadığını belirlemek için analiz edilir. AS4 ila AS9 ana araştırma sorularıdır ve kalan sorular (AS1 ila AS3) son çalışmaların bağlamını değerlendirmemize yardımcı olur. AS1 ila AS3 bize SC alanındaki belirli bir araştırma alanının özetini verir.

3.3 Metodoloji

3.3.1 İnceleme Yöntemi

SC sistemleri hakkında literatür taraması yapılırken sistematik bir yaklaşım seçilir. Sistematik literatür incelemeleri (SLİ), SC de iyi bilinen bir inceleme yöntemidir. Sistematik literatür incelemesi (genellikle sistematik inceleme olarak adlandırılır), belirli bir araştırma sorusu, konu alanı veya ilgi konusu hakkındaki tüm mevcut araştırmaları belirleme, değerlendirme ve yorumlama olarak tanımlanır [18]. Bu SLİ, Kitchenham ve Charters (2007) tarafından önerilen orijinal yönergelere dayanmaktadır. Bu bölümdeki bazı çalışmalar ve şekiller (Radjenović, Heričko, Torkar ve Živkovič, 2013) [19], (Unterkalmsteiner ve diğerleri, 2012) [20] ve Wahono [21] tarafından da uyarlanmıştır. Şekil 3.1'de görüldüğü gibi SLİ,



Şekil 3.1 Sistematik Literatür İncelemesi Adımları

literatürün planlanması, yürütülmesi ve raporlanması olmak üzere 3 aşamadan oluşmaktadır. İlk adımda gereksinimler belirlenir. Genel yaklaşım bölümünde gerçekleşme hedeflerinden bahsedilir. Ardından, SC üzerine mevcut SLİ çalışmaları toplanır ve incelenir. Bu çalışmanın amacı, SLİ ikinci aşamasında

araştırmacı kaynaklı önyargıların etkisini en aza indirerek metodolojik tarafsızlığı sağlamaktır (Adım 2). Araştırma sorularını, arama stratejisini, dahil etme ve hariç tutma kriterleriyle çalışma seçim sürecini, kalite değerlendirmesini ve son olarak veri çıkarma ve sentez sürecini tanımlamıştır. İnceleme, inceleme yürütme ve raporlama aşamasında geliştirilir, değerlendirilir ve yinelemeli olarak iyileştirilir.

3.4 Arama Stratejisi

Sistematik literatür taramasının yürütülmesinde arama süreci belirli aşamalardan oluşmaktadır. Bu aşamalar; uygun dijital veri tabanlarının seçilmesi, arama anahtar terimlerinin belirlenmesi, arama sorgularının yapılandırılması ve sorgu sonucunda elde edilen çalışmaların veri tabanlarından çıkarılmasını kapsamaktadır. Literatürde en ilgili çalışmalara ulaşabilmek için öncelikle alanla ilişkili ve güvenilir dijital kaynaklar belirlenmiştir. Çalışma konusunun kapsamını geniş tutmak adına, alanında yaygın kullanılan ve yüksek atıf alan literatür veri tabanları tercih edilmiştir. Bu doğrultuda taramada ACM, IEEE Xplore, ScienceDirect, SpringerLink, Scopus gibi dijital veri tabanları kullanılmıştır. Arama sorgularının oluşturulması belirli metodolojik kriterlere dayandırılmıştır. Bu kriterler aşağıdaki şekilde özetlenebilir:

- Arama terimleri, PMKSB (Popülasyon, Müdahale, Karşılaştırma, Sonuç, Bağlam) çerçevesinde özellikle Popülasyon ve Müdahale bileşenlerine göre türetilmiştir.
- Araştırma soruları doğrultusunda uygun anahtar kelimeler geliştirilmiştir.
- Terimler, ilgili çalışmalarda başlık, özet ve anahtar kelime bölümlerinde taranabilecek şekilde tanımlanmıştır.
- Eşanlamlılar, zıt anlamlılar ve farklı yazım biçimleri göz önünde bulundurularak kapsamlı bir terim listesi oluşturulmuştur.
- Bu terimlerden yararlanılarak mantıksal operatörler (Boolean AND/OR) kullanılarak detaylı bir arama ifadesi inşa edilmiştir.

Kullanılan örnek arama dizgesi şu şekildedir:

("question answering" AND "natural language processing") AND ("information retrieval") AND ("Document Retrieval" OR "Passage Retrieval" OR "Answer Extraction")

Dijital veri tabanlarında yürütülen aramalar; makale başlıkları, özetleri ve anahtar kelimeleri kapsayacak şekilde sınırlandırılmıştır. İnceleme yalnızca 2000–2022 yılları arasında yayımlanmış çalışmaları kapsamış ve değerlendirmeye yalnızca hakemli dergi makaleleri ile konferans bildirileri dahil edilmiştir. Ayrıca, yalnızca İngilizce dilinde yayımlanmış yayınlar dikkate alınmıştır.

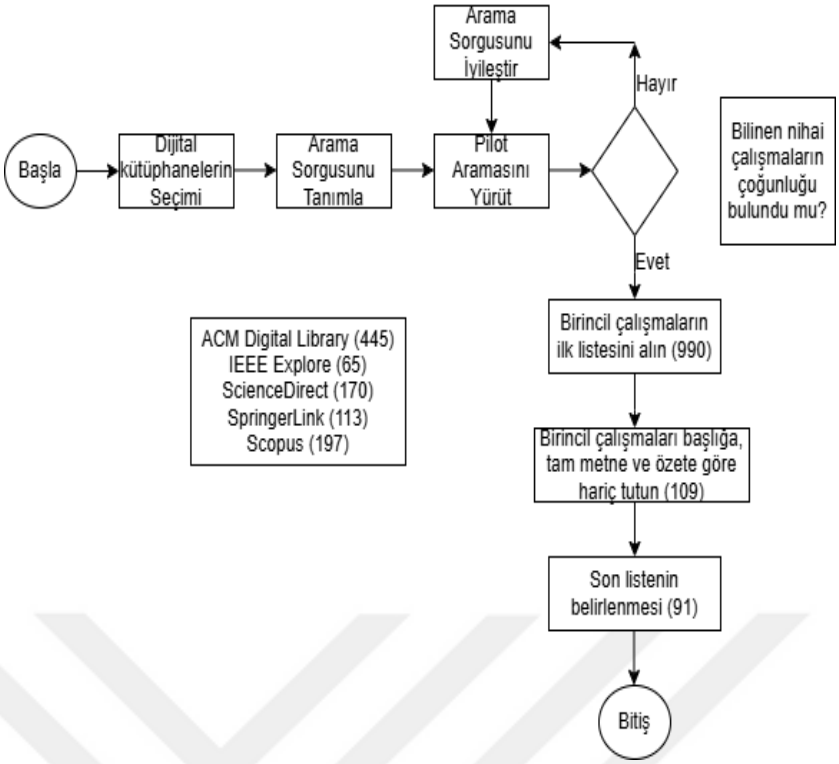
3.4.1 Çalışma Seçimi

Dahil etme ve hariç tutma kriterleri nihai çalışmaları belirlemek için kullanılmıştır. Bu kriterler Tablo 3.3'te gösterilmiştir.

Tablo 3.3 Dahil Etme ve Hariç Tutma Kriterleri

Dahil Etme Kriterleri	Büyük ve küçük ölçekli veri kümelerini kullanan akademik ve endüstriyel çalışmalar
	Soru cevaplama alanında modelleme performansını tartışan ve karşılaştıran çalışmalar
	Hem konferans hem de dergi sürümlerine sahip çalışmalar için yalnızca dergi sürümü dahil edilecektir
	Aynı çalışmanın yinelenen yayınları için yalnızca en yenisi dahil edilecektir
Hariç Tutma Kriterleri	Güçlü bir doğrulaması olmayan veya soru cevaplamanın deneysel sonuçlarını içermeyen çalışmalar
	Soru cevaplama veri kümelerini, yöntemlerini, çerçevelerini soru cevaplamadan farklı bir bağlamda tartışan çalışmalar
	İngilizce yazılmamış çalışmalar

Şekil 3.2, inceleme sürecinin her adımını ve belirlenen sayıyı göstermektedir. Çalışma seçme süreci iki adımda gerçekleştirilmiştir. Başlık, özet ve tam metinli çalışmalar kaldırılmıştır. Literatür çalışmaları ve deneysel sonuçları içermeyen çalışmalar da hariç tutulmuştur. Diğer çalışmalar, kalan çalışmalardan SC ile benzerlik derecesine göre dahil edilmiştir.



Şekil 3.2 Final Çalışmaların Taranması ve Seçilmesi

İncelemenin ilk aşamasında, analiz sürecine dahil edilecek çalışmaların yer aldığı nihai literatür listesi oluşturulmuştur. Bu kapsamda, toplam 91 çalışma seçilerek nihai listeye dahil edilmiştir. Sonrasında, belirlenen bu 91 çalışmanın tamamı tam metinleri üzerinden ayrıntılı olarak değerlendirilmiştir. Bu inceleme sürecinde, çalışmaların araştırma sorularıyla olan ilgisi, içerik benzerlikleri ve metodolojik yeterlilikleri göz önünde bulundurularak önceden belirlenen dahil etme ve hariç tutma kriterleri doğrultusunda seçim yapılmıştır.

3.4.2 Veri Çıkarımı

Son çalışmaları çıkarmanın amacı, araştırma sorularını ele almaya katkıda bulunan verileri toplamaktır. Seçilen 91 çalışma için veri çıkarma formu her birine ayrı ayrı uygulanmıştır (Adım 6). Veri çıkarma formu, araştırma sorularını yanıtlamak için gereken son çalışmalardan veri toplamak üzere tasarlanmaktadır. Özellikler, tanıtmak istediğimiz araştırma soruları ve analizler aracılığıyla belirlenmektedir. Tablo 3.4'te gösterilen araştırma sorularını yanıtlamak için altı özellik kullanılmaktadır. Veri çıkarma yinelemeli olarak gerçekleştirilir.

Tablo 3.4 Araştırma Sorularına Eşlenen Veri Çıkarma Özellikleri

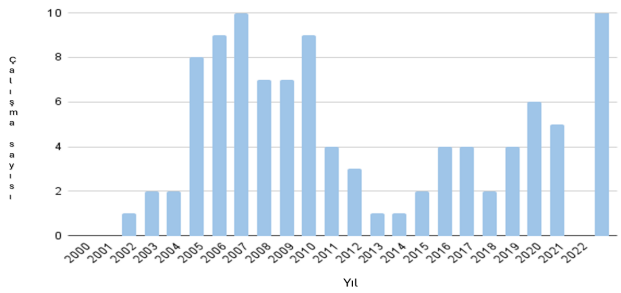
Özellik	Araştırma Soruları
Araştırmacılar ve Yayınlar	AS1, AS2
Araştırma Trendleri ve Konuları	AS3
Soru Cevaplama Veri Kümeleri	AS4
Soru Cevaplama Ölçütleri	AS4
Soru Cevaplama Yöntemleri	AS5, AS6, AS7, AS8
Soru Cevaplama Çerçeveleri	AS9

3.4.3 Geçerliliğe Yönelik Tehditler

Bu derlemede, soruları cevaplamada kullanılan tekniklere dayalı olarak yapılan çalışmaların incelenmesi amaçlanmıştır. Literatür taranırken, dergilerde yayınlanan tüm makalelerin başlıklarını manuel olarak incelemek zor olduğundan bazı konferans bildirileri ve bazı soru-cevap makaleleri dergilerden çıkarılmıştır.

3.4.4 Önemli Dergi Yayınları

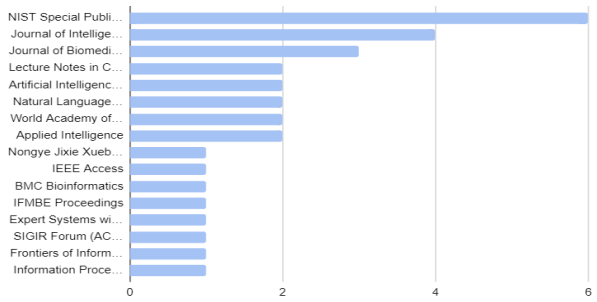
Bu literatür incelemesinde, SC alanında toplamda 91 çalışma yer almaktadır. Bu çalışmalara dayanarak, SC konusundaki araştırmaların yıllar içindeki sayısal değişimini analiz ettik. Buradaki amaç, alanın zaman içindeki evrimini gözlemlemektir. Yıllara dayalı gözlemler, Şekil 3.3'te sunulmuştur. Elde edilen bulgular, SC alanına olan ilginin 2005 yılı itibarıyla önemli ölçüde arttığını ve gerçekleştirilen çalışmaların daha güncel bir düzeye geldiğini göstermektedir.



Şekil 3.3 Seçilen Çalışmaların Yıllara Göre Dağılımı

Seçilen son çalışmalara göre, en önemli soru cevap dergileri Şekil 3.4'te gösterilmektedir.

Tablo 3.5, en önemli soru cevap dergilerinin Scimago Dergi Sıralaması (SDS) değerini ve Q kategorilerini (Q1-Q4) göstermektedir. Dergi yayınları SDS değerlerine göre sıralanmıştır.



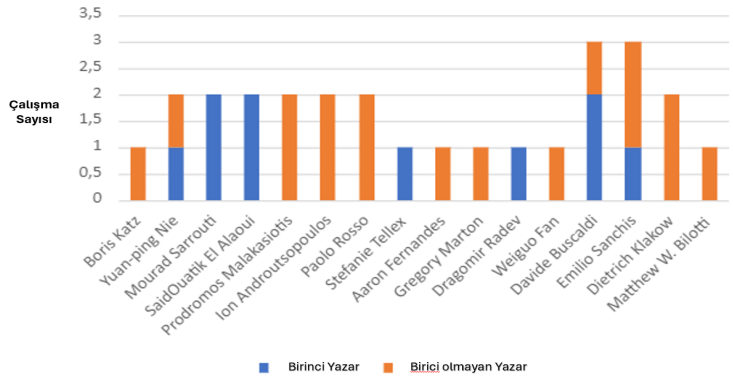
Şekil 3.4 Dergi Yayınları ve Seçili Çalışmaların Dağılımı

Tablo 3.5 Seçilen Dergilerin Scimago Dergi Sıralaması (SDS)

Dergi Yayınları	SDS
BMC Biyoinformatik	1,567
Uygulamalı Uzman Sistemler	1,368
SIGIR Forum (ACM Bilgi Alma Özel İlgi Grubu)	1,337
Bilgi İşleme Yönetimi	1,061
Biyomedikal Bilişim Dergisi	1,057
Tıpta Yapay Zeka	0,98
Uygulamalı Zeka	0,791
IEEE Erişim	0,587
Nongye Jixie Xuebao/Çin Tarım Makineleri Derneği İşlemleri	0,461
Akıllı Bilgi Sistemleri Dergisi	0,424
Bilgi Teknolojisinin Sınırları, Elektronik Mühendisliği	0,406
Doğal Dil Mühendisliği	0,29
Bilgisayar Biliminde Ders Notları	0,249
NIST Özel Yayını	0,202
IFMBE Bildirileri	0,152
Dünya Bilim Akademisi	0,137

3.4.5 En Aktif ve Etkili Araştırmacılar

Seçilen son çalışmalar arasında, SC alanına önemli katkılarda bulunmuş ve aktif olarak çalışan araştırmacılar araştırılmış ve belirlenmiştir. Şekil 3.5'da en aktif ve etkili araştırmacılar çalışma sayısına göre listelenmiştir. Boris Katz, Yuan-ping Nie, Mourad Sarrouti, SaidOuatik El Alaoui, Prodromos Malakasiotis, Ion Androutsopoulos, Paolo Rosso, Stefanie Tellex, Aaron Fernandes, Gregory Marton, Dragomir Radev, Weiguo Fan, Davide Buscaldi, Emilio Sanchis, Dietrich Klakow, Matthew W. Bilottiare SC alanındaki en aktif araştırmacılardır.



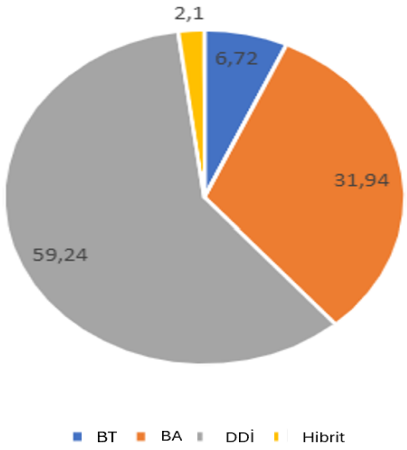
Şekil 3.5 Etkili Araştırmacılar ve Çalışma Sayısı

3.4.6 Soru Cevaplama Alanındaki Araştırma Konuları

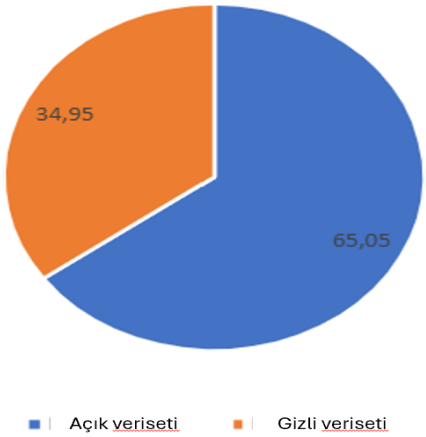
Bu soruyu cevaplamak için Yao'nun sınıflandırma makalesini ele alınmıştır. Seçilen son çalışmaların analizi, mevcut SC araştırmasının dört konuya odaklandığını ortaya koymaktadır [21]:

- DDİ Tabanlı: Bu yaklaşımlar, cevapların çıkarılmasında makine öğrenimi ve doğal dil işleme tekniklerini kullanır.
- BA Tabanlı: Arama motoru teknolojileriyle ilgili olarak, bu yöntemler cevabın ve ilgili belgelerin veya pasajların alınması ve sıralanması süreçlerine odaklanır.
- BT Tabanlı (Bilgi Teknolojileri): Bu tür yaklaşımlar, yapılandırılmış veriler üzerinden cevap bulmaya yönelir. Kelime tabanlı aramalar yerine, standart veritabanı sorguları kullanılır [22]
- Hibrit Tabanlı: Hibrit model, BA, DDİ ve BT tekniklerinin bir birleşimini ifade eder.

Şekil 3.6, 2000 ile 2022 yılları arasında SC konusundaki araştırma alanlarının dağılımını göstermektedir. Bu bağlamda, 91 çalışmanın %6,72'si BT paradigması, %31,94'ü DDİ paradigması, %59,24'ü BA paradigması ve %2,1'i Hibrit yaklaşımı kullanmıştır. DDİ uygulamalarının biraz daha fazla kullanıldığını gözlemlemekteyiz. SC, araştırmacılarının çoğunun araştırma konuları olarak DDİ'yi seçtiği sonucuna varılabilir. Araştırmacıların bu konuya odaklanmasının nedenleri olarak, arama motoru aracılığıyla bilgi edinme üzerine yapılan çalışmalar artmaktadır. Yapılandırılmamış verilerden en doğru cevabı çıkarmak için birçok metin DDİ ve makine öğrenmesi tekniği uygulanmaya çalışılmıştır.



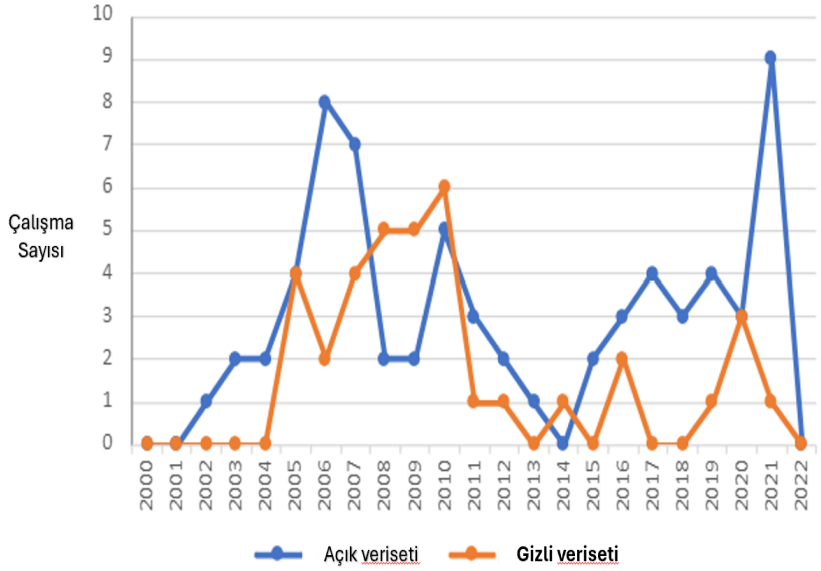
Şekil 3.6 Araştırma Konularının Dağılımı



Şekil 3.7 Veri Kümelerinin Toplam Dağılımı

3.4.7 Soru Cevaplamada Kullanılan Veri Kümeleri

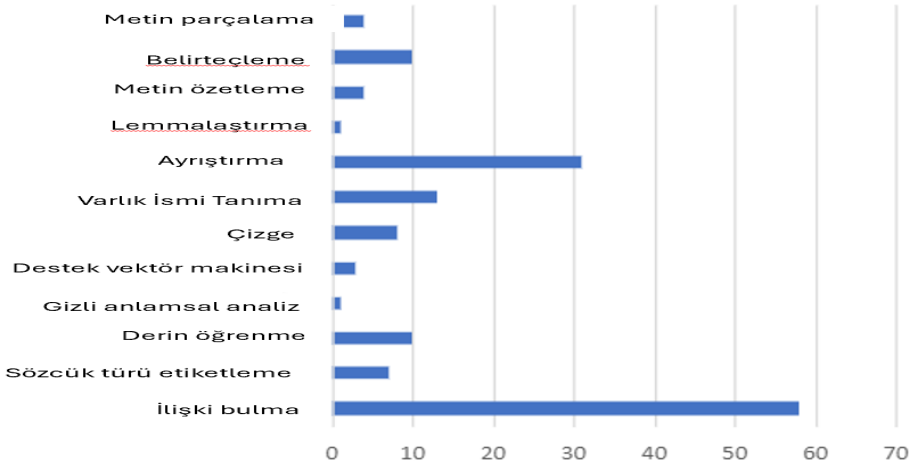
Veri seti, makine öğrenmesinin uygulandığı bir veri koleksiyonudur [22]. Eğitim seti, öğrenme sistemine verilerek modelin eğitildiği veridir. Test seti ya da değerlendirme seti, modelin performansını eğitim verisi üzerinde geliştirildikten sonra ölçmek amacıyla kullanılan veri kümesidir. Son çalışmada yer alan 91 çalışmada kullanılan verilerin veri seti tipleri Şekil 3.7'de gösterilmiştir. Veri setleri incelendiğinde, 2000-2022 yılları arasındaki çalışmaların %65,05'i kamuya açık veri iken, %34,95'i özel veridir. En çok kullanılan kamuya açık veriler arasında Text Retrieval Conference (TREC) [23], Cross-Language Evaluation Forum (CLEF) [24] ve Biomedical SC (BioASQ) [25] verilerini görmekteyiz. Özel veriler özel şirketlere ait olduğundan kamuya açık veri setleri olarak dağıtılamaz. Şekil 3.8'da veri kümelerinin yıllara göre dağılımı sunulmuştur. Çalışmaların %35,95'i özel veri kümeleridir. Bu veri kümeleri kamuya açık olmadığından, çalışmaların sonuçları önerilen modellerin sonuçlarıyla karşılaştırılamaz. Dağılımlara bakıldığında, kamuya açık verilerin kullanımına ilişkin artan bir farkındalık vardır.



Şekil 3.8 Açık ve Gizli Veri Setlerinin Dağılımı

3.4.8 Soru Cevaplamada Kullanılan Yöntemler

Şekil 3.9'da görüldüğü gibi 2000 yılından bu yana SC alanında kullanılan ve önerilen on dört yöntem belirlenmiştir. Belirlenen bu yöntemler Şekil 3.9 gösterilmiştir.



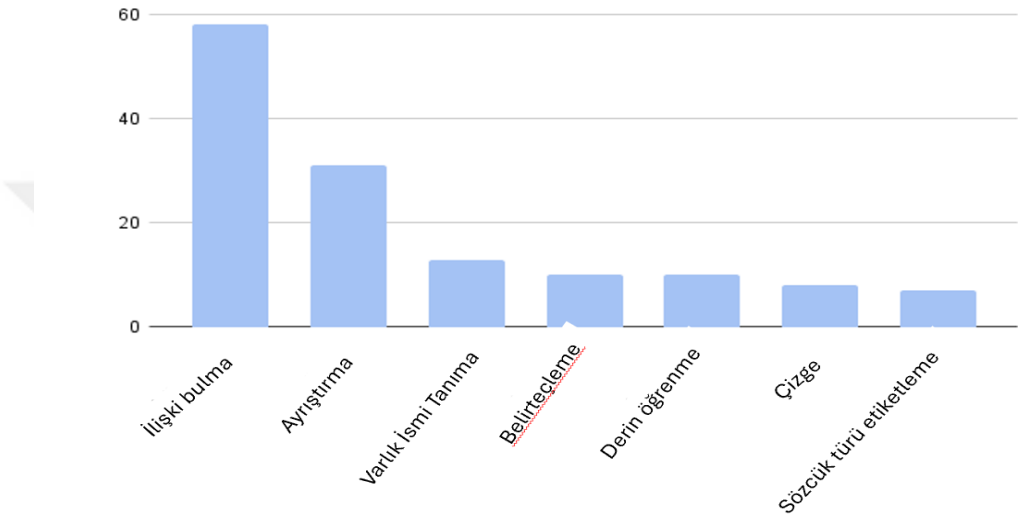
Şekil 3.9 Soru Cevaplamada Kullanılan Yöntemler

3.4.9 Soru Cevaplamada En Çok Kullanılan Yöntemler

Şekil 3.10'de gösterildiği gibi, en çok uygulanan on dört sınıflandırma yönteminden yedisi gösterilmiştir. Bunlar:

- İlişki bulma

- Ayrıştırma
- Belirteçleme
- Derin öğrenme
- Grafik
- Sözcük türü etiketleme

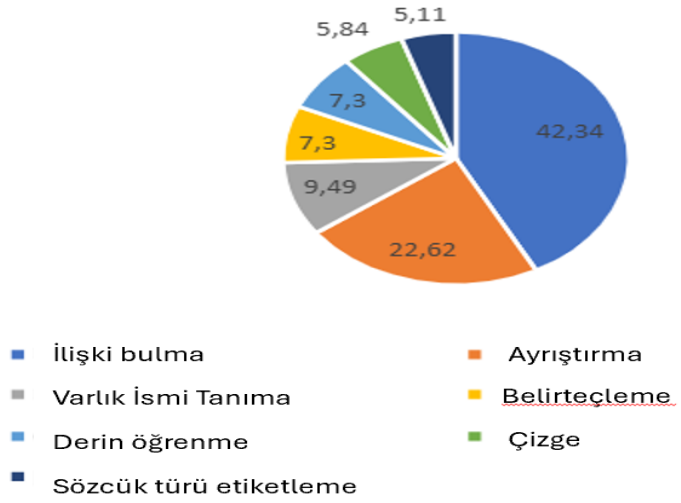


Şekil 3.10 Soru Cevaplamada En Çok Kullanılan Yöntemler

İlişki bulma, ayrıştırma, varlık ismi tanıma, belirteçleme en çok kullanılan dört yöntemdir. Şekil 3.11'de gösterildiği gibi, seçilen çalışmaların %81,82'si tarafından benimsenmiştir.

3.4.10 Soru Cevaplamada Sıklıkla Kullanılan Yöntemler

SC alanında birçok çalışma yapılmıştır. Yapılan literatür incelemesi, DDİ, BA ve CÇ'den oluşan bir işlem hattı sürecini ortaya koymaktadır. Doğal dilde formüle edilmiş bir soru, ilk olarak analiz aşamasına tabi tutulur. Yani bir sonraki adım olan belge almayı kolaylaştırmak için arama sorguları oluşturulur. Literatür incelendiğinde ilk çalışmalarda alma aşamasında çoğunlukla tf-idf, bm25 [26–28] gibi klasik yöntemlerin kullanıldığı görülmektedir. Burada kullanıcı tarafından girdi olarak alınan sorguya benzer kelimeler aranarak alma sağlanır. Yapılan diğer çalışmalarda, en yaygın kullanılan yöntemlerden biri varlık ismi tanıma (VİT) ve sözcük türü etiketleme teknikleridir. Bu yaklaşımlar sayesinde,



Şekil 3.11 Çalışmaların Yöntem Türlerine Göre Dağılımı

semantik rol etiketlemesi uygulandığında, alma aşamasında başarıda belirgin bir artış gözlemlenmiştir [29–31]. Ayrıca, klasik sınıflandırma yöntemlerinden biri olan SVM (Destek Vektör Makineleri) kullanımı da yaygındır. Burada, sorgunun ait olduğu kategori, o kategoriye ait belgelerin alınmasında kullanılan sınıflandırıcıdır. Semantik anlamda iyileştirme, SVM ile sağlanmıştır [27, 32]. Bununla birlikte, klasik yöntemlerin bazı sınırlamaları bulunmaktadır; özellikle, yanlış yazılmış kelimeleri veya anlam bakımından benzer terimleri doğru şekilde tanımlayamama gibi sorunlarla karşılaşılabilmektedir. Literatür taraması yapıldığında, son yıllarda derin öğrenme temelli çalışmaların önemli ölçüde arttığı dikkat çekmektedir. Derin öğrenme yöntemleri üzerine yapılan incelemelerde, klasik yöntemlere kıyasla daha üstün sonuçlar elde edildiği tespit edilmiştir [33–36]. Derin öğrenmenin temel avantajı, kelimelerin anlamını ve yanlış yazımları doğru bir şekilde yakalayabilmesidir. Bu nedenle, son yıllarda SC alanındaki çoğu çalışma derin öğrenme tekniklerine dayanmaktadır.

3.4.11 Soru Cevaplama İçin Önerilen Yöntem İyileştirmeleri

Araştırmacılar son zamanlarda SC de en çok kullanılan derin öğrenme, çizge ve klasik yöntemler altında öneriler sunmuşlardır. Bu bölümde önerilen teknikler analiz edilmiştir.

3.4.11.1 Derin Öğrenme

Klasik yöntemlerin başaramadığı anlamsal yakalama ve kelime eşleştirme durumundan kaçınılarak, derin öğrenme ile devrim atlanmıştır. Derin öğrenme teknikleri birçok alanda önemli ilerlemeler kaydettiğinden, SC sistemlerinin birçok aşamasına entegre edilmiştir. Soruların sınıflandırılmasında, belgelerin CNN, LSTM, BERT modelleri kullanılmıştır (Chen Y.)(Nie P.)(Kratzwald, B.)(Lin, H) [33, 36, 37]. Literatür incelendiğinde, bu çalışmalarda Makine Okuduğunu Anlama (MOA) yöntemi benimsenmiş ve SC sistemlerinde bir dönem atlanmıştır. MOA yönteminde modern alıcı-okuyucu mimarisi sunulmuştur. Alıcı, soru girdisine göre ilgili belgeleri getirir. Öte yandan, okuyucu, alıcıdan döndürülen belgelerden cevabı çıkaran aşamadır (Pappas, D. ve diğerleri) (Yao Cong) [34, 38]. Alınan belgeleri okuyucu aşamasına vermeden önce, belgelerin yeniden sıralanması gibi çalışmalar yürütülmüştür (Pappas, D. ve diğerleri) [34]. Bazı çalışmalarda, cümledeki herhangi bir kelimenin diğer kelimelerle ilişkisini öz-dikkat ekleyerek ortaya çıkarmak için çalışmalar yürütülmüştür (Nguyen, T.M) [39].

3.4.11.2 Çizge

Grafik ağ metinleri grafiğe dönüştürürken, kelimeler düğümler olarak belirlenir. İki kelimeyi bir kenarla birbirine bağlar ve aralarındaki ilişkiyi gösterir. Literatürü incelediğimizde, grafik tekniğinin kullanıldığı ve anlamsal çıkarımda başarılı sonuçlar elde edildiği görülmüştür (Yuyu Zhang)(Guo, Q.-1.)(Grau, B.) [36, 40, 41]. Yuyu Zhang, grafik tabanlı yeniden sıralama konusundaki çalışmasında son teknolojiye göre daha başarılı olduğunu belirtmiştir (Yuyu Zhang) [36].

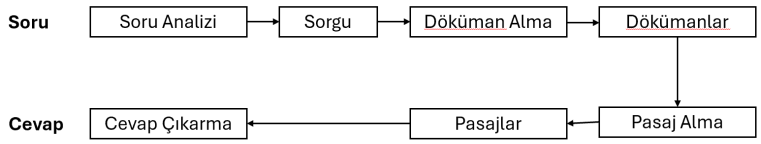
3.4.12 Çok Atıf Alan ve Dolayısıyla Etkili Üç Mimari

SC alanında çok atıf alan ve dolayısıyla etkili üç mimari: Tellex mimari, Cao. Y mimari, Radev D. mimari

3.4.12.1 Tellex S. Mimari

Tellex, bu çalışmada, belge alma çalışmalarına kıyasla pasaj alma algoritmalarının çok fazla incelenmemiş olması nedeniyle pasaj alma algoritmalarına odaklanıldığı belirtilmektedir [28]. Geçiş alıcı algoritmalarının SC üzerindeki etkilerini nicel olarak inceler. Telex, farklı geçiş alıcı algoritmalarının performansını karşılaştırmak için modüler bir test ortamı geliştirdi [28]. Önerilen altyapı dört modülden oluşur.

- Soru analiz modülleri girilen soruları sorguya dönüştürür ve soruları analiz eder. Problemin türü belirlenmektedir. Örneğin, beklenen yanıt türü "Albert



Şekil 3.12 Tellex S. Mimari

Einstein ne zaman doğdu?" tarih türüdür. Bu bilginin yanıt çıkarma sürecini kısaltmaya yardımcı olacağı belirtilmektedir.

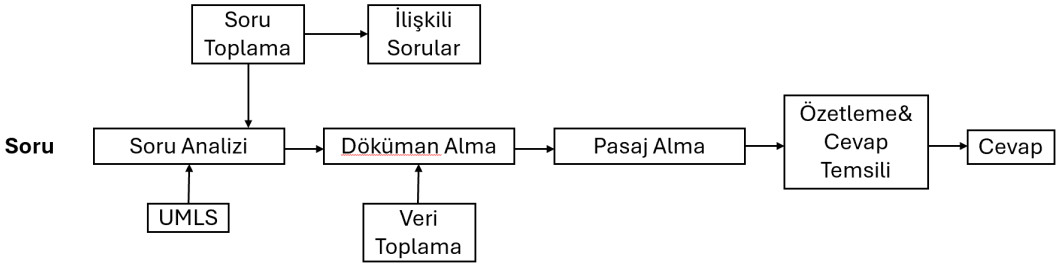
- Belge alma modülü, ilgili belgeyi korpustan vermektedir. Ek bir işlem olarak belge kümesine indirgendiği belirtilmektedir.
- Geçiş alma modülü, döndürülen belgeleri işler ve sorgu terimlerine göre puanlanmış pasajların sıralı listelerini vermektedir.
- Cevap çıkarma modülü, kullanıcının sorusuna nihai yanıtı vermek için döndürülen pasajlardan yanıtı aramaktadır. Beklenen yanıt türüyle eşleşen vit teknolojisi kullanılmaktadır.

Tellex 8 farklı pasaj alma algoritması uygulamaktadır. Veri kümesi olarak TREC-10 seçilip, uygulanan algoritma koleksiyonunun sonuçlarını birleştirerek yeni bir pasaj alma algoritması tasarlanmaktadır. Telex, bu Oylama algoritmasını BM, ISI ve SiteQ kullanarak çalıştırmıştır. Oylama, algoritmanın başarı oranının diğerlerinden daha yüksek olduğunu belirtmektedir [28].

3.4.12.2 Cao. Y Mimari

Cao, klinik sorular üzerinde sağlam bir semantik analiz gerçekleştirdi ve yanıt olarak soru odaklı çıkarımsal özetler üretmek için AskHERMES adlı bir klinik SC sistemi oluşturmuştur [27]. Şekil 3.13, oluşturdukları çerçeveyi göstermektedir. Giriş olarak doğal dil klinik sorusu alan AskHERMES'in sistem mimarisini göstermektedir. Soru Analizi, sorudan bilgi gereksinimlerini otomatik olarak çıkarır ve bir sorgu terimleri listesi oluşturmaktadır. Sorgu terimi genişletmesi için UMLS bilgi kaynağı kullanılmaktadır. Extract Related Questions modülü benzer soruların bir listesini sağlar. Get Info, yerel olarak indekslenmiş ilgili belgeleri vermektedir. Bilgi çıkarma, ilgili pasajları belirlemiştir. Summarizing and Presenting Answers, cevap pasajlarını toplar, gereksiz bilgileri kaldırır, otomatik olarak yapılandırılmış özetler oluşturur ve özetleri soruyu soran kullanıcıya sunmuştur [27]. AskHERMES daha erken aşamalarındadır ve bilgi kaynağı Google veya UpToDate ile karşılaştırıldığında sınırlıdır. Ancak, karmaşık sorularda

hem Google hem de UpToDate sistemlerinden daha iyi performans gösterme potansiyeline sahip olduğu belirtilmiştir [27].



Şekil 3.13 Cao. Y. Mimari

3.4.12.3 Radev D. Mimari

Radev, arama motorlarını güçlendiren bir mimari geliştirdiklerini belirtmektedir. Beş aşamadan oluşan bir mimari gerçekleştirmişlerdir [42]. Bunlar sorgu modülasyonu, belge alma, pasaj çıkarma, ifade çıkarma ve cevap sıralamasıdır. Önerdikleri yöntem, bu aşamaların son üçüne yönelik bazı olasılıksal yaklaşımları tanımlar. Veri seti olarak TREC kullanılmıştır [42].

3.5 Tartışma

Bu literatür taraması, 2000-2022 yılları arasında SC araştırmalarında kullanılan eğilimleri, veri kümelerini, yöntemleri ve çerçeveleri belirlemeyi ve analiz etmeyi amaçlamıştır. Dahil etme ve hariç tutma kriterlerine bağlı olarak, finalde 91 SC çalışması bir araya getirilmiştir. Bu literatür taraması sistematik bir literatür taraması olarak yürütülmüştür. Sistematik literatür taraması, belirli araştırma sorularına yanıtlar sağlamak için mevcut araştırma kanıtlarını belirleme, değerlendirme ve yorumlama süreci olarak tanımlanmaktadır. Literatür taraması yapıldığında, gürültülü veriler, model performansı ve başarı oranları gibi zorlukların ele alındığı görülmekte ve bu konular hâlâ araştırmaya açık alanlar olarak kalmaktadır. Seçilen nihai çalışmaların incelenmesi sonucunda, mevcut SC araştırmalarının dört ana konu etrafında yoğunlaştığı anlaşılmaktadır: BT,BA,DDİ ve Hibrit yaklaşımlar. SC yöntemlerinin toplam dağılımı aşağıdaki gibidir. Araştırma çalışmalarının %6,72'si BT yöntemlerine, %31,94'ü BA yöntemlerine, %59,24'ü DDİ yöntemlerine ve %2,10'u Hibrit temele odaklanmıştır. Ayrıca analiz edilen çalışmaların %65,05'inde açık erişimli veri kümeleri tercih edilmişken, %34,95'inde özel veri kümeleri kullanılmıştır. SC sistemlerinin geliştirilmesinde toplamda on dört farklı yöntem uygulanmış olup, bunlar arasında en yaygın şekilde kullanılan yedi temel yöntem öne çıkmaktadır. Bunlar ilişki bulma (benzerlik

mesafesi), ayrıştırma, varlık ismi tanıma , belirteçleme, derin öğrenme, sözcük türü etketleme, grafiktir. Bu tekniklerden bazılarını kullanarak, araştırmacılar SC alanında doğruluğu artırmak için bazı teknikler önermişlerdir. Ayrıca, bu çalışmalar sonucunda en çok atıfta bulunulan ve etkili üç mimari belirlenmiştir. Bunlar Cao Y. mimari, Telex S. mimari, Radev D. mimarisidir. (Cao) AskHERMES SC sistemini uyguladılar. Bu sistem belge tabanlı bir çalışma olduğundan, anahtar kelime hakkında bilgi çıkarımının eksik olduğunu belirtmektedir. (Radev),(Tellex) Çalışmalarında ölçeklenebilirlik sorununa değinmişlerdir. Geliştirdikleri gerçek zamanlı SC sisteminin birçok ön işleme adımının yanıt süresini uzattığını ve performansının iyileştirilmesi gerektiğini belirtmektedirler. Son çalışmaların listesi Tablo 3.6'da sunulmaktadır. Bu liste yıl, son çalışmalar, yayınlar, veri kümeleri, yöntemler ve konular olmak üzere altı nitelikten oluşmaktadır. 91 son çalışma yayın yılına göre sıralanmıştır. Tablo 3.6, SC alanında sistematik bir literatür taramasının sonuçlarını sunan zihin haritasını göstermektedir. Bu zihin haritasıyla, gelecekteki çalışmalar için bir fikir vermek üzere büyük bir resim olarak sunulmaktadır. Büyük resimde kullanılan yöntemleri ve seçimleri görmek, olaylara yeni bir bakış açısı getirmektedir.

Tablo 3.6 Literatür Taramasındaki Ayrıntılı Tablo Analizi

Yıl	Final Çalışma	Veri kümesi	Method	Konu
2003	[43]	Açık	Metin Parçalama	DDİ
2004	[44]	Açık	İlişki Bulma	BA,
2021	[45]	Gizli	Derin Öğrenme	BA
2021	[36]	Açık	Çizge	BA, DDİ
2021	[46]	Açık	Derin Öğrenme	BA, DDİ
2021	[34]	Açık	Derin Öğrenme	BA,DDİ
2020	[33]	Gizli	Derin Öğrenme	Hibrit
2020	[47]	Açık	Çizge + Derin Öğrenme	BA, DDİ
2020	[48]	Açık	İlişki Bulma, Metin Özetleme	DDİ
2020	[37]	Gizli	Derin Öğrenme	BA,DDİ
2020	[49]	Gizli	Belirteçleme, İlişki Bulma, Metin Özetleme, Ayrıştırma	BA, DDİ
2019	[35]	Gizli	Derin Öğrenme, İlişki Bulma	DDİ
2019	[50]	Açık	İlişki Bulma	DDİ

Tablo 3.6 Literatür Taramasındaki Ayrıntılı Tablo Analizi (Devamı)

Yıl	Final Çalışma	Veri kümesi	Method	Konu
2019	[51]	Açık	Vit, sözcük türü etiketleme, Ayrıştırma	BA, DDİ
2018	[52]	Açık	Vit, İlişki Bulma	DDİ
2018	[53]	Açık	İlişki Bulma	DDİ
2017	[54]	Açık	İlişki Bulma	DDİ, BA
2017	[55]	Açık	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ, BA
2016	[56]	Açık	İlişki Bulma, Vit, Ayrıştırma	DDİ
2016	[57]	Gizli	Metin özetleme, İlişki Bulma	DDİ, BA
2016	[58]	Açık	Vit	DDİ, BT
2016	[59]	Açık	İlişki Bulma	DDİ, BA
2015	[60]	Açık	İlişki Bulma, Metin Özetleme	DDİ
2015	[61]	Açık	Ayrıştırma	DDİ, BT
2014	[62]	Gizli	Vit, Ayrıştırma, sözcük türü etiketleme	DDİ, BA
2013	[63]	Açık	İlişki Bulma	BA,DDİ
2012	[64]	NTCIR	İlişki Bulma	DDİ
2012	[62]	Açık	Belirteçleme, İlişki Bulma	DDİ
2011	[65]	Gizli	İlişki Bulma	DDİ, BA
2011	[66]	Açık	Çizge, Ayrıştırma	BA,DDİ
2011	[27]	Açık	Ayrıştırma, İlişki Bulma	DDİ
2010	[67]	Gizli	Çizge, Ayrıştırma, İlişki Bulma	DDİ, BT
2010	[68]	Gizli	İlişki Bulma	DDİ, BA
2010	[69]	Açık	İlişki Bulma	BA,DDİ
2010	[29]	Gizli	Vit, Ayrıştırma	Hibrit
2010	[70]	Gizli	İlişki Bulma	DDİ, BT
2010	[71]	Gizli	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ, BT
2010	[72]	Gizli	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ, BT

Tablo 3.6 Literatür Taramasındaki Ayrıntılı Tablo Analizi (Devamı)

Yıl	Final Çalışma	Veri kümesi	Method	Konu
2010	[73]	Açık	İlişki Bulma, Ayrıştırma, Belirteçleme	BA,DDİ
2009	[74]	Açık	İlişki Bulma	DDİ
2009	[75]	Gizli	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ, BT
2009	[76]	Gizli	Çizge, Ayrıştırma	DDİ
2009	[77]	Gizli	İlişki Bulma	DDİ
2009	[78]	Gizli	İlişki Bulma	DDİ,BT
2009	[79]	Açık	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ
2008	[80]	Gizli	fuzzy logic	BA
2008	[81]	Açık	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ, BA
2008	[82]	Gizli	Vit, sözcük türü etiketleme, Ayrıştırma	DDİ, BA
2008	[83]	Gizli	İlişki Bulma,Vit	DDİ, BA
2008	[84]	Gizli	Vit	DDİ, BA
2008	[85]	Gizli	İlişki Bulma	DDİ
2007	[86]	Açık	İlişki Bulma,Vit, Ayrıştırma	DDİ
2007	[87]	Açık	İlişki Bulma, Ayrıştırma, Belirteçleme	DDİ
2007	[88]	Açık	İlişki Bulma,Vit	BA,DDİ
2007	[89]	Gizli	İlişki Bulma	DDİ
2007	[90]	Açık	İlişki Bulma	DDİ, BA
2007	[41]	Gizli	Çizge	DDİ,
2007	[91]	Gizli	İlişki Bulma, Ayrıştırma	DDİ
2007	[92]	Açık	Destek vektör makinesi, İlişki Bulma	DDİ
2007	[93]	Açık	İlişki Bulma, Ayrıştırma	DDİ, BA
2007	[94]	Gizli	Destek vektör makinesi, İlişki Bulma	BA, DDİ
2006	[95]	Açık	İlişki Bulma	BA, DDİ

Tablo 3.6 Literatür Taramasındaki Ayrıntılı Tablo Analizi (Devamı)

Yıl	Final Çalışma	Veri kümesi	Method	Konu
2006	[96]	Gizli	Vit, Metin Parçalama, stemming	BA, DDİ
2006	[97]	Açık	Ayrıştırma, Metin Parçalama	DDİ
2006	[98]	Açık	İlişki Bulma	BA,DDİ
2006	[73]	Açık	İlişki Bulma	BA, DDİ
2006	[99]	Gizli	Ayrıştırma, Metin Parçalama	BA, DDİ
2006	[100]	Açık	İlişki Bulma	DDİ
2006	[101]	Açık	İlişki Bulma	BA, DDİ
2005	[102]	Gizli	İlişki Bulma	BA, DDİ
2005	[103]	Gizli	İlişki Bulma, Ayrıştırma	DDİ
2005	[104]	Açık	Ayrıştırma, İlişki Bulma	DDİ
2005	[105]	Gizli	Çizge	DDİ
2005	[106]	Açık	sözcük türü etiketleme, Ayrıştırma	BA,DDİ
2005	[107]	Gizli	İlişki Bulma,	BA, DDİ
2005	[108]	Açık	sözcük türü etiketleme, Ayrıştırma	DDİ
2005	[109]	Açık	sözcük türü etiketleme, Ayrıştırma, lemma	BA,DDİ
2004	[44]	Açık	İlişki Bulma,	DDİ
2003	[28]	Açık	İlişki Bulma,	BA,DDİ
2002	[42]	Açık	İlişki Bulma, Ayrıştırma	BA,DDİ

4
WİKİPEDİA VERİ SETİNDE AÇIK ALAN SORU CEVAPLAMA OKUYUCU SİSTEMLERİ SAĞLAMAK İÇİN BİR İŞ AKIŞI

4.1 Genel Yaklaşım

Önemli belgeler arttıkça, bunlardan kritik bilgileri çıkarma ihtiyacı da artmaktadır. Elbette, verideki artış istenen bilgiyi bulmayı zorlaştırmaktadır. Bunu önlemek için arama motorları gibi birçok SC sistemi geliştirilmektedir. Kullanıcı etkileşiminin yoğun olduğu kurumlar, bilgiye hızlı ve doğru erişim sağlamak amacıyla genellikle açık alan soru-cevap (SC) sistemlerine yönelmektedir. Bu çalışma, belge tabanlı bilgi erişim süreçlerini desteklemek üzere, literatürde tanımlanmış açık alan okuma sistemlerini temel alarak bir iş akışı modeli önermektedir. Önerilen iş akışı, Wikipedia tabanlı bir veri kümesi üzerinde uygulanmakta olup, Wikipedia'nın sürekli güncellenen, topluluk kaynaklı ve geniş kapsamlı yapısı, açık alan SC sistemleri için ideal bir test ortamı sunmaktadır. Bu kapsamda, sistemin değerlendirilmesi için SQuAD veri kümesi tercih edilmiştir. Açık alan SC sistemleri, artan bilgi ihtiyacına paralel olarak son yıllarda önemli bir araştırma alanı haline gelmiştir. Bu sistemler, Makine Okuduğunu Anlama (MOA) süreci kapsamında, bir sorguya en uygun cevabın belgeler üzerinden çıkarılması amacıyla birden fazla aşamadan geçer. Bu çalışmada, her bir aşama ayrı modüller halinde ele alınmakta, kullanılan teknikler, yöntemler ve değerlendirme kriterleri ayrıntılı olarak sunulmaktadır. Başlangıçta kural tabanlı yöntemlerle geliştirilen SC sistemleri, günümüzde büyük ölçüde derin öğrenme tabanlı yaklaşımlara evrilmiştir. Bunun temel nedeni, kural tabanlı sistemlerin yazım hataları ve eşanlamlı kelimeler gibi dilin doğal yapısından kaynaklanan karmaşıklıkları yakalamakta yetersiz kalmasıdır. Derin öğrenme modelleri ise bu eksiklikleri gidermekte ve çok daha yüksek başarı oranları sunmaktadır. Bu nedenle önerilen iş akışı, derin öğrenme temelli bir okuma sisteminin tasarımını esas almaktadır. Literatürdeki güncel çalışmalar, derin öğrenme mimarilerinin

özellikle dikkat (attention) mekanizmaları aracılığıyla metni daha etkili bir şekilde anladığını ortaya koymaktadır. Bu bağlamda, SC sistemlerinde sinirsel dikkat mekanizmaları kullanılarak metnin anlamlı bölümleri belirlenmekte ve böylece başarı oranı artırılmaktadır [ör. [110–112]]. MOA alanında yapılan diğer araştırmalarda ise kodlayıcı-kod çözücü yapılarının kullanıldığı ve bu yaklaşımların etkili sonuçlar ürettiği raporlanmıştır [ör. [113–116]]. Bu tür modeller, giriş olarak verilen soruyu ve bağlamı birlikte işleyerek doğrudan cevabı üretmektedir. SC okuyucu bileşenleri, genellikle çıkarımsal ve üretken olmak üzere iki gruba ayrılmaktadır. Çıkarımsal okuyucular, cevabın bağlam içindeki başlangıç ve bitiş konumlarını belirlerken; üretken okuyucular, cevabı doğrudan üretir. Bu çalışmada her iki yaklaşım da değerlendirilmiştir: çıkarımsal okuyucu olarak çift yönlü dikkat mekanizması içeren modeller, üretken okuyucu olarak ise T5 mimarisi tercih edilmiştir. Çalışmamızda, T5 tabanlı üretken modeller ile çift yönlü dikkat mekanizması kullanan çıkarımsal modeller karşılaştırmalı olarak ele alınmakta; bu modellerin uygulanması sırasında takip edilmesi gereken ayrıntılı iş akışı adım adım sunulmaktadır. İş akışı içerisinde kullanılan modüller, gerekli veri türleri, yöntem seçimi ve işlem sıraları ayrıntılı biçimde açıklanmakta; elde edilen sonuçlar ise SQuAD veri kümesi üzerinden analiz edilerek paylaşılmaktadır.

4.2 Araştırma Soruları

Açık alan SC sistemleri, yapılandırılmamış metin belgeleri üzerinden sorulara en uygun yanıtları otomatik olarak üretmeyi amaçlamaktadır. Bu belgeler, dar (küçük) veya geniş (büyük) bağlamlara sahip olabilir. Dar bağlamlardan bilgi çıkarmak genellikle daha az karmaşıklık içerdiğinden, kural tabanlı yöntemlerle etkili sonuçlar elde edilebilmektedir. Ancak bağlam genişledikçe, bu yöntemlerin hem doğruluk hem de performans açısından yetersiz kaldığı görülmektedir. Geniş kapsamlı metinlerdeki anlamı yakalayabilme konusunda derin öğrenme tabanlı yaklaşımlar, önemli avantajlar sunarak başarı oranlarını belirgin biçimde artırmıştır. Bu bağlamda çalışmamız, açık alan SC sistemlerinde diziden-diziye (sequence-to-sequence) mimarilere dayalı okuyucu sistemlerin tasarımı ve performansını değerlendirmeyi amaçlamaktadır. Araştırma sürecine yön veren temel sorular aşağıda sıralanmıştır:

- AS1: Diziden-diziye mimariler temel alınarak etkili bir SC okuyucu sistemi iş akışı nasıl yapılandırılabilir?
- AS2: Soru-cevap okuyucu sistemlerinde hangi tür diziden-diziye modülleri iş akışına entegre edilebilir?

- AS3: Okuyucu iş akışında çift yönlü dikkat mekanizması kullanan diziden-diziye modellerin performans düzeyi nedir?
- AS4: T5 tabanlı transformatör modelleri, artan veri kümesi boyutları karşısında nasıl bir davranış sergilemektedir?
- AS5: Tasarlanan diziden-diziye iş akışının başarı oranı hangi ölçütlerle ve nasıl değerlendirilebilir?

4.3 Gereksinimler ve Kullanım Durumu

Bu çalışmada, açık alan SC okuyucu sistemleri için bir iş akışı öneriyoruz. Bu iş akışının başarılı bir şekilde uygulanabilmesi için belirli gereksinimlerin karşılanması ve kullanım durumlarının tanımlanması gerekmektedir. Bu bölümde, sistemin gereksinimlerini ve bu gereksinimlerin hangi kullanım durumlarında karşılanacağını detaylandırıyoruz. Veri kümesi gereksinimleri olarak sistemin eğitimi ve değerlendirilmesi için SQuAD gibi etiketlenmiş soru-cevap çiftlerine ihtiyaç duyulmaktadır. Bu veri kümesi, modelin performansını ölçmek için kullanılmaktadır. Sistem, T5 ve çift yönlü dikkat mekanizması gibi derin öğrenme modellerini kullanmaktadır. Bu modellerin eğitimi için yüksek performanslı hesaplama kaynaklarına (GPU/TPU) ihtiyaç vardır. Python programlama dili ve TensorFlow, PyTorch gibi derin öğrenme kütüphaneleri kullanılarak gerçekleştirilmektedir. Model eğitimi ve testi için yüksek performanslı GPU veya TPU'lar gereklidir. Bu, büyük veri kümeleri üzerinde hızlı ve etkili bir şekilde çalışmayı sağlar. Veri kümeleri ve eğitilmiş modeller için yeterli depolama alanı gereklidir. Wikipedia gibi büyük bir bilgi kaynağı üzerinde belirli bir soru sorar. Sistem, bu soruya en uygun cevabı bulmak için veri kümesini tarar ve cevabı kullanıcıya sunar. Bu senaryo, sistemin büyük bir veri kümesi üzerinde hızlı ve doğru bir şekilde arama yapabilmesini gerektirir. Ayrıca, kullanıcıya anlaşılır ve doğru cevaplar sunabilmesi için yüksek doğruluk oranına sahip olmalıdır. Araştırmacılar veya geliştiriciler, sistemin performansını artırmak için farklı modelleri eğitir ve değerlendirir. Bu süreçte, SQuAD gibi etiketlenmiş veri kümeleri kullanılır. Sistemin farklı modeller üzerinde eğitim ve test yapabilmesi için esnek bir yapıya sahip olmasını gerektirir. Ayrıca, eğitim sürecinde yüksek hesaplama gücüne ihtiyaç duyulur. Sistem, belirli bir alana (örneğin, tıp, hukuk) özgü belgeler üzerinde SC görevini gerçekleştirir. Bu tür belgeler, genel içerikli belgelerden farklı bir dil ve terminolojiye sahip olabilir. Sistemin özel alanlara özgü terminolojiyi anlayabilmesi ve bu tür belgeler üzerinde etkili bir şekilde çalışabilmesi için özelleştirilmiş modellere ihtiyaç duyar. Kullanıcı, gerçek zamanlı olarak sorular sorar ve sistem anında cevaplar üretir. Bu tür bir senaryo, özellikle

chatbot'lar veya sanal asistanlar için kullanışlıdır. Sistemin düşük gecikme süresiyle çalışabilmesini ve kullanıcıya hızlı bir şekilde cevap verebilmesini gerektirir. Ayrıca, sistemin yüksek doğruluk oranına sahip olması önemlidir. Bu gereksinimler ve kullanım durumları, önerilen iş akışının başarılı bir şekilde uygulanabilmesi için temel oluşturmaktadır.

4.4 Temel Kavramlar

Transformatör mimarisi temel alınarak geliştirilen önceden eğitilmiş dil modelleri genellikle iki farklı yapıda sınıflandırılmaktadır: Autoencoder tabanlı ve diziden-diziye (sequence-to-sequence) yapılar. Autoencoder modeller yalnızca kodlayıcı (encoder) bileşenine sahiptir ve bu kategoriye BERT modeli örnek olarak gösterilebilir. Buna karşılık, T5 ve BART gibi diziden-diziye modeller hem kodlayıcı hem de kod çözücü (decoder) bileşenleri içerir ve genellikle Wikipedia gibi büyük ölçekli metin koleksiyonlarıyla ön eğitimden geçirilir. Üretken okuyucu modeller, bağlam içerisinden doğrudan cevap konumunu belirlemek yerine, girdiden doğal dilde bir cevap üretmeye odaklanır. Bu tür sistemlerde cevap, metin içerisinde birebir geçmek zorunda değildir. T5 ve BART gibi diziden-diziye temelli modeller, bu bağlamda sıklıkla kullanılmaktadır. Üretken yapıdaki okuyucular, klasik çıkarımsal sistemlerden farklı olarak doğrudan yanıt metnini üretmeye yönelir. Çıkarımsal okuyucular ise yanıtın bağlam içinde yer aldığı varsayımıyla çalışır ve doğru cevabın bulunduğu metin parçasını belirlemeye çalışır. Örneğin, BERT tabanlı DPR modeli [117], bağlamlar arasında tarama yaparak en uygun cevap aralığını tespit eder.

Bazı sistemlerde ise grafik tabanlı yapılardan yararlanılarak, bilgi grafikleri üzerinden yanıt üretilmeye çalışılır. Bu sistemlerde, grafik evrişimsel ağlar aracılığıyla pasaj temsilleri öğrenilir ve ardından bu temsillerden cevap çıkarılır. Diğer bazı modeller, tek bir belgeye değil, çok sayıda dokümana dayalı olarak cevap aralığı belirler. Örneğin, DrQA [112] sistemi belgeleri paragraflara ayırır ve her paragraftan ad-özne tanıma (NER), sözcük türü etiketleme (POS) ve terim sıklığı (TF) gibi özellikler çıkarır. Bu özellikler daha sonra Bi-LSTM katmanı aracılığıyla işlenerek yanıtın bulunduğu metin parçası tahmin edilir. BERTSerini çalışması, sistemlerin yanıt aralıklarına verdikleri olasılık puanlarının doğrudan karşılaştırılmasının yanıltıcı olabileceğini öne sürerek bir normalizasyon mekanizması önermiştir. Bu yaklaşım daha sonra BidAF ve BERT gibi birçok açık alan soru-cevap sisteminde benimsenmiştir ve okuyucu modüllerinin geliştirilmesinde temel bir yöntem haline gelmiştir. 2015 yılından itibaren dikkat (attention) mekanizmalarının önemi giderek artmış ve farklı doğal dil işleme

görevlerinde etkili biçimde kullanılmaya başlanmıştır. Bu kapsamda dikkat temelli bellek yapıları geliştirilmiş ve öz-dikkat (self-attention) mekanizması LSTM tabanlı ağlarla birleştirilerek LSTMN mimarisi ortaya çıkarılmıştır [118]. Öz-dikkat sayesinde, farklı konumlardaki bilgilerin birbirleriyle ilişkisi daha etkin biçimde modellenebilmiştir [119]. Dikkat tabanlı yaklaşımlar, görüntü altyazılama [120], metin özetleme [121], konuşma tanıma [122], video açıklama üretimi [123], sinirsel makine çevirisi [124], görsel soru-cevap sistemleri [125] ve metinsel çıkarım tanıma [126] gibi birçok farklı alanda başarılı biçimde uygulanmıştır. Ayrıca, cümle temsilleri elde etmek için dikkat mekanizmasına dayalı evrişimsel sinir ağları da önerilmiştir [127]. 2017 yılında Vaswani ve arkadaşları tarafından sunulan Transformatör modeli, tamamen dikkat mekanizmasına dayalı bir sinir ağı yapısı olarak geliştirilmiş ve geleneksel tekrarlayıcı ya da evrişimsel yapıların yerini almıştır [119]. Bu modelde, hem kodlayıcı hem de kod çözücü yığınları yer almakta ve her yığın çoklu dikkat katmanları ile ileri beslemeli katmanlardan oluşmaktadır. Konumsal bilgi, bu yığınlara başlangıçta eklenmektedir. RNN tabanlı diziden-diziye modeller, bir girdi dizisini alarak karşılık gelen bir çıktı dizisi üretir. Bu yapılar, soru-cevap sistemlerinde sıkça tercih edilmektedir; çünkü bir soru cümlesine karşılık doğal dilde bir cevap cümlesi oluşturulması mümkündür. LSTM tabanlı modeller, girdiyi gizli durum ve hücre durumu şeklinde özetleyerek kodlayıcının son durumlarını kod çözücünün başlangıç durumlarına aktarır. Bu başlangıç vektörleri aracılığıyla kod çözücü, sırasıyla yanıtı oluşturmaya başlar. T5 modeli, tüm doğal dil işleme görevlerini "metin girişi–metin çıkışı" çerçevesi içinde tanımlar. Bu özelliği sayesinde çok amaçlı bir yapı sunar. T5, BERT ile benzer şekilde maskeli dil modeli (MLM) ilkesiyle eğitilmiş olmakla birlikte, önemli bir farklılık taşır: BERT, her bir maskelenmiş kelimeyi temsil etmek için ayrı ayrı [MASK] belirteçleri kullanırken, T5 birden fazla ardışık kelimeyi tek bir maske sembolüyle temsil edebilir. Bu yönüyle T5, daha esnek bir yapı sunar ve çok çeşitli dil işleme görevlerinde başarıyla kullanılabilir.

4.5 Literatür İncelemesi

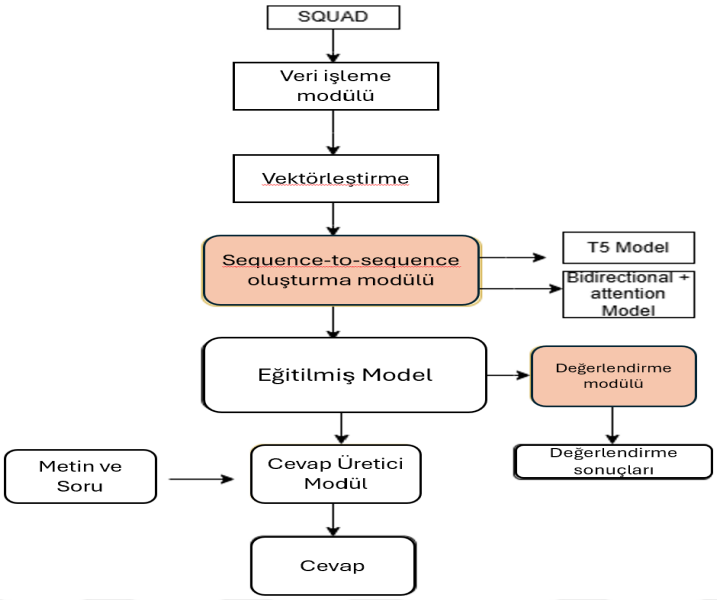
Açık alan soru-cevap (SC) sistemleri, genellikle yapılandırılmamış metin koleksiyonlarından doğru bilgiyi çıkararak soruları yanıtlamayı amaçlayan sistemler olarak tanımlanmaktadır. Bu alan, başlangıçta bilgi tabanları (KB) üzerine inşa edilmiş ve Freebase gibi kaynaklardan yararlanılarak WebQuestions ve SimpleQuestions gibi veri kümeleri geliştirilmiştir. Ancak, bilgi tabanlarının içerdiği sınırlı kapsam ve yapılandırılmış veri ihtiyacı nedeniyle, araştırmacılar doğrudan metin üzerinden yanıt üretmeye odaklanmaya başlamıştır. Bu yönelimin temelinde, sistemlerin doğal dili anlaması ve yorumlaması gerekliliği yatmaktadır.

Özellikle derin öğrenme yaklaşımlarının yaygınlaşması ve dikkat (attention) tabanlı modellemelerin gelişmesi, SC alanında önemli ilerlemelere katkı sağlamıştır. Başlangıçta kullanılan ön eğitimli dil modelleri arasında GPT yalnızca sol bağlamı modelleyerek sınırlı bir görüş alanına sahipken, Elmo modeli hem sol hem de sağ bağlam temsillerini birleştirerek daha dengeli bir anlayış sunmuştur. Ancak bu modeller, bağlamın iki yönlü etkileşimini tam anlamıyla yakalayamamıştır. BERT'in ortaya çıkışıyla birlikte, sol ve sağ bağlamlar arasındaki karşılıklı ilişki maskeli dil modelleme (MLM) yöntemiyle etkili bir biçimde öğrenilmeye başlanmıştır. Bu yaklaşım sayesinde, eksik kelimeler bağlamdan tahmin edilerek daha başarılı dil temsilleri elde edilmiştir. Makine okuma anlayışı (MOA) alanındaki araştırmalarda, dikkat mekanizmalarının çeşitli biçimlerde uygulandığı gözlemlenmektedir. İlk örneklerden biri olan Bahdanau'nun dinamik dikkat modeli [128], sorgu, bağlam ve geçmiş dikkat skorları üzerinden ağırlıkları güncellemeyi önermiştir. Daha sonra Chen ve arkadaşları bu yapıyı detaylandırmış [129], Wang ve Jiang ise RNN'lerin ilerlemesiyle birlikte sorguyu dikkate alan yeni bir model geliştirmiştir [130]. Bu gelişmeleri takiben, BiDAF (Çift Yönlü + Dikkat) modeli önerilmiş ve hafızasız dikkat yapısıyla dikkat çeken bir yapı haline gelmiştir [116].

SC sistemlerinde kullanılan okuyucu modeller, bağlamdan doğrudan yanıt üreten ya da yanıtın yer aldığı konumu belirleyen iki ana sınıfa ayrılır. İlki olan çıkarımsal okuyucular, yanıtın bağlam içerisinde açıkça bulunduğunu varsayarak belirli bir metin aralığını tahmin eder. Bu çalışmada geliştirilen çift yönlü dikkat modeli bu gruba örnek teşkil eder. Diğer tür ise üretken okuyucular dır; bu sistemler, bağlamı yorumlayarak doğal dilde yeni bir yanıt üretir. Son dönemde üretken okuyucu modellerine yönelik çalışmalar, BART [114] ve T5 [115] gibi transformatör temelli mimariler kullanılarak yoğunlaşmıştır. Ancak bu modeller, zaman zaman dil bilgisel hatalar, tutarsızlıklar ve mantıksal sorunlar gibi eksiklikler göstermekte ve bu nedenle daha fazla araştırma gerektirmektedir [131]. Önceki dönemlerde SC sistemlerinde ağırlıklı olarak çıkarımsal okuyucular kullanılmıştır [116, 117, 132–139], ancak son yıllarda üretken okuyucu modellerine yönelim artmıştır [140, 141]. SC sistemlerine yönelik daha geniş kapsamlı bir değerlendirme çalışması [142] tarafından sunulmuş olup, bu çalışmada sistemlerin bileşenlerine ilişkin bütüncül bir iş akışı önerilmiştir. Literatürde çeşitli dağıtık mimarilerle tasarlanmış iş akışı tabanlı sistem örnekleri mevcuttur [143–147]. Ancak bu araştırmadan farklı olarak, söz konusu çalışmada özel olarak SC sistemlerine özgü iş akışlarının tasarımı ve uygulanmasına odaklanılmıştır. Ayrıca, kendi kendini optimize eden sistemlerde meta veri kullanımına dayalı çözümler de önerilmiştir [148–150]. Bu çalışmada ise, SC sistemlerinin yetkinliğini artırmak amacıyla diziden-diziye (seq2seq) modellerin sunduğu üretken yapılar ön plana çıkarılmıştır.

4.6 Önerilen Metodoloji

Bu bölümde, Okuyucu sistemi için baştan sona bir iş akışının sistem mimarisini, bir bağlam ve bir soru girildiğinde ve bu bağlamdan cevabı çıkarmak için ne tür bir iş akışının geçmesi gerektiğini öneriyoruz. Bu iş akışı, araştırma sorularımız AS1 ve AS2'nin çalışmamızdır. Önerilen bu iş akışı Şekil 4.1'de gösterilmiştir ve aşağıda ayrıntılı olarak açıklanmıştır. Açık alan soru-cevap sistemlerinde

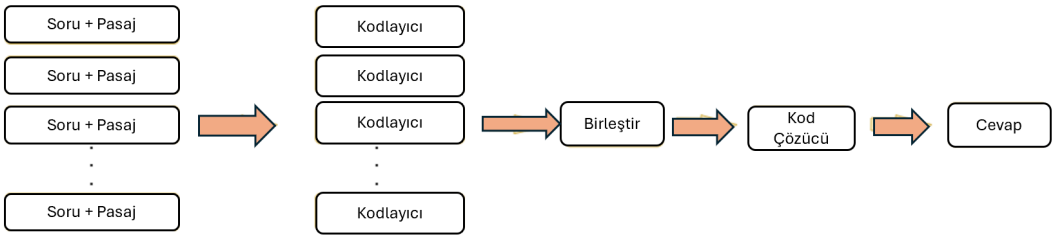


Şekil 4.1 Açık Alan Soru Cevaplama Okuyucu Sistemleri İçin İş Akışı

kullanılan okuyucu modeller, bağlam ve soru çiftlerini girdi olarak alarak bu bilgilere dayanarak yanıt üretir. Bu kapsamda, iki farklı ön işleme yaklaşımı uygulanmaktadır. T5 tabanlı modellerde, veri işleme aşamasında her bir pasaj ve başlığa sırasıyla "bağlam", "başlık" ve "soru" gibi açıklayıcı önekler eklenir. Ardından, bu metin parçaları birleştirilerek belirteçleyiciden geçirilir ve modelin işleyebileceği biçimde vektörel temsillere dönüştürülür [115]. Çift yönlü dikkat mekanizması kullanan modellerde ise, bağlam ve soruya ek olarak doğrudan doğru cevabın kendisi de giriş olarak verilmektedir [116]. Bu modeller, yanıtın bağlam içerisindeki başlangıç ve bitiş konumlarını tespit etmeye çalışır. Metinler, hem kelime düzeyinde hem de karakter düzeyinde belirteçlenerek daha ayrıntılı bir şekilde işlenir. Kelime gömme işlemiyle her sözcük, yüksek boyutlu sürekli bir vektörle temsil edilir. Bu amaçla genellikle önceden eğitilmiş GloVe vektörlerinden yararlanılır. Bu yöntem, kelimelere anlamsal yakınlık kazandırarak modelin bağlamı daha iyi öğrenmesini sağlar. Çift yönlü dikkat tabanlı modellerde, kelime gömme ile birlikte karakter gömme de uygulanır ve bu sayede dilin daha ince yapısal özellikleri modele kazandırılır.

Kodlayıcı, giriş dizisindeki her belirteci işleyerek bu diziyi sabit uzunlukta bir

bağlam vektörüne dönüştürür. Elde edilen bu vektör daha sonra kod çözücü birime iletilir. Bu yapı sayesinde model, verilen soru ve bağlamdan doğru cevabı üretme yeteneği kazanır. Şekil 4.2, bu süreci gerçekleştiren T5 mimarisine sahip modellerdeki kodlayıcı ve kod çözücü bileşenlerini göstermektedir.



Şekil 4.2 Açık Alan Soru Cevaplama Okuyucu Sistemleri İçin Encoder-Decoder Modeli

T5 modeli: T5, metinden metne yaklaşımını kullanan bir dönüştürücü modeldir. Çeviri, Sınıflandırma, SC, metni girdi olarak alıp besleme ve bazı hedef metinler oluşturma gibi bazı görevler için eğitim verilir. çift yönlü dikkat mekanizması: çift yönlü dikkat mekanizması birçok aşamadan ve altı katmandan oluşur.

- Karakter Gömme Katmanı: Bu katman, her kelimeyi karakter düzeyindeki evrişimsel sinir ağları (CNN) kullanarak bir vektörle temsil eder. Bu yöntem, kelimelerin karakter bazındaki yapısal özelliklerini yakalamayı amaçlar.
- Kelime Gömme Katmanı: Önceden eğitilmiş kelime gömme modellerinden faydalanarak, her kelimeye anlam taşıyan bir vektör atar. Bu katman, kelimelerin semantik ilişkilerini daha iyi temsil edebilmek için kelime düzeyinde bir temsille işlem yapar.
- Bağlamsal Gömme Katmanı: Kelimelerin etrafındaki kelimelerden elde edilen bağlamsal ipuçlarını kullanarak, kelimelerin gömme vektörlerini geliştirir. Bu işlem, hem soru hem de bağlam metinleri üzerinde uygulanarak, daha derin bağlamsal temsiller elde edilmesini sağlar.
- Dikkat Akışı Katmanı: Bu katman, her kelime için bir özellik vektörü üretmek amacıyla, soru ve bağlam vektörlerini birleştirir. Bu sayede, önemli bilgiler daha güçlü bir şekilde işlenir ve kelimeler arası ilişkiler daha iyi anlaşılır.
- Modelleme Katmanı: Bağlamı analiz etmek için Tekrarlayan Sinir Ağları (RNN) kullanılır. Bu katman, metin içerisindeki zaman ve sıralama ilişkilerini modelleyerek anlamlı çıkarımlar yapmayı sağlar.

- Çıktı Katmanı: Son olarak, sorgu için en uygun cevabın üretildiği katmandır. Bu aşama, önceki katmanlarda işlenen verilerden yararlanarak sonuca ulaşır.

Modelin performansını değerlendirmek amacıyla, kesinlik (precision), geri çağırma (recall) ve F1 puanı hesaplanmıştır. Bu bağlamda, araştırma sorularından AS5 ile ilişkili olarak başarı oranı analiz edilmiştir.

- Kesinlik, modelin ürettiği tahmin ile gerçek cevap arasında örtüşen belirteç (token) sayısının, toplam tahmin belirteci sayısına oranı olarak tanımlanır.
- Geri çağırma ise, modelin tahmininde yer alan ve yer gerçeği ile örtüşen belirteç sayısının, yer gerçeğindeki toplam belirteç sayısına oranıdır.
- F1 puanı: (2 * kesinlik * geri çağırma) / (kesinlik + geri çağırma)

4.7 Önerilen Metodolojinin Prototip Uygulaması ve Analizi

Squad veri kümesiyle çift yönlü+dikkat ve T5 tabanlı çalışmalar yaptık. Squad, büyük bir Wikipedia makaleleri kümesindeki makine anlama veri kümesidir. Squad verilerinden (soru-cevap) 2500, 5000, 7500, 10000, 12500, 15000, 17500 ve 20000 parçalık veri çiftleri aldık ve hem Çift yönlü+dikkat hem de T5 tabanlı için eğitim sonuçlarını kademeli olarak çıkardık. Bu bölümde, AS3 ve AS4 sorularına yanıt arıyoruz. Train yaparken, Ram 51 GB diskli Google Colaboratory Pro+. 166 GB alan kullanıldı. T5 tabanlıda, hiperparametreler 100 adım, optimize edici Adam, toplu iş boyutu 2, öğrenme oranı 1e-4 olarak belirlendi. Çift yönlü+dikkat, hiperparametreler 25 epoch, optimize edici Adam, öğrenme oranı 0.0005. Tablo 4.1 ve Tablo 4.2'de gösterilmiştir.

Tablo 4.1 T5 tabanlı Değerlendirme sonuçları

	2500	5000	7500	10000	12500	15000	17500	20000
f1	0,40	0,43	0,42	0,55	0,48	0,48	0,59	0,53
hassasiyet	0,22	0,24	0,24	0,32	0,27	0,26	0,38	0,30
geri çağırma	0,62	0,60	0,49	0,68	0,64	0,65	0,62	0,62

4.8 Tartışma

Bu tezde, açık alan Soru-Cevap (SC) okuyucu sistemleri için bir iş akışı önerilmektedir. Önerilen iş akışının etkinliğini ve uygulanabilirliğini gösterebilmek adına bir prototip uygulaması da geliştirilmiştir. Bu prototip, özellikle Wikipedia

Tablo 4.2 Çift Yönlü Dikkat Mekanizması Değerlendirme sonuçları

	2500	5000	7500	10000	12500	15000	17500	20000
f1	0,21	0,14	0,14	0,25	0,19	0,21	0,28	0,22
hassasiyet	0,21	0,19	0,17	0,26	0,27	0,16	0,32	0,24
geri çağırma	0,33	0,25	0,22	0,35	0,42	0,24	0,41	0,59

veri kümeleri üzerinde test edilmiştir. Çalışmada, genel içerikli metinlerden bilgi çıkarma sürecine odaklanılmıştır, ancak geliştirilen metodolojinin, belirli belgelere de uygulanabilmesi mümkündür. SC okuyucu sistemi iş akışı sağlamak amacıyla farklı derin öğrenme yaklaşımları kullanılmıştır. Bunlardan biri T5 modeline dayanırken, diğeri çift yönlü dikkat mekanizması modelini kullanmaktadır. Bu iki yaklaşım, aynı veri kümeleri üzerinde farklı metriklerle karşılaştırılarak prototipin etkinliği test edilmiştir.

Bu çalışmanın amacı, SC okuyucu sistemleri aracılığıyla bilgiye nasıl erişileceğini ve bu bilgilere nasıl yönlendirilip gezileceğini gösterirken, belirtilen araştırma sorularına yanıtlar sunmaktır. Bilgi tabanlı araştırmalar geliştirmek oldukça zaman alıcı ve çaba gerektiren bir süreç olduğundan, son yıllarda derin öğrenme gibi modern yöntemlerin kullanıldığı çalışmalarda bir artış gözlemlenmektedir. Çalışma kapsamında, her iki okuyucu sisteminin uygulanmasına ilişkin test sonuçları analiz edilmiş ve mantıksız ya da tutarsız yanıtların üretilebileceği tespit edilmiştir. Bu bulgu, bu alandaki araştırma faaliyetlerinin daha da derinleştirilmesi gerektiğini ortaya koymaktadır. Gelecekte, önerilen iş akışının üzerinde yapılan çalışmalarla, SC işlevselliğinin kalitesinin artırılması hedeflenmektedir.

5
MERKEZİ OLMAYAN HİBRİT SORU CEVAPLAMA SİSTEMLERİNE BİR YAKLAŞIM

5.1 Genel Yaklaşım

Günümüzde, internet üzerindeki büyük miktarda bilgi, doğru yanıtların elde edilmesini her geçen gün daha da zorlaştırmaktadır. Bu zorlukları aşmak amacıyla, araştırmacılar, doğal dildeki sorguları anlayabilen ve uygun yanıtlar sunabilen SC sistemleri geliştirmiştir. Ancak, bu tür sistemler genellikle farklı organizasyonlar tarafından geliştirilmekte ve çeşitli platformlarda çalıştırılmakta, bu da kullanıcıların doğru ve zamanında yanıt almalarını zorlaştırmaktadır. Bu sorunların çözülmesine yönelik olarak, hibrit SC sistemlerinin geliştirilmesi önerilmektedir. Hibrit SC sistemleri, kullanıcıların çeşitli veri kümeleri üzerinde eğitilmiş ve farklı platformlarda çalışan birden fazla SC sistemine tek bir merkezi arayüz üzerinden erişmesini mümkün kılmaktadır. Bu model, büyük veri setlerinden bilgi çıkarma sürecinin doğruluğunu ve etkinliğini önemli ölçüde artırma potansiyeline sahiptir. Bu çalışma, hibrit soru-cevaplama sistemi çerçevesinin geliştirilmesine yönelik uygulanabilirliği araştıran bir yaklaşıma odaklanmaktadır. Önerilen yazılım mimarisi, farklı veri kümeleri üzerinde eğitilmiş ve çeşitli platformlarda çalıştırılmış soru-cevaplama sistemlerine tek bir arayüz üzerinden erişim imkanı sunarak, bu sistemlerin şeffaf ve bütünleşik bir biçimde kullanılmasını sağlamaktadır. Özellikle, önerilen yazılım mimarisi, kullanıcıların birden fazla SC sistemine tek noktadan erişim sağlamalarına olanak tanımakta ve böylece sistemlerin birlikte çalışabilirliğini artırmaktadır. Bu yaklaşım, hem kullanıcı deneyimini iyileştirmekte hem de SC sistemlerinin kullanım süreçlerini daha verimli hale getirmektedir. Önerilen çerçevenin kullanılabilirliğini ve etkinliğini göstermek amacıyla bir prototip yazılım geliştirmiştir. Gerçekleştirilen deneysel çalışmalar, söz konusu prototipin pratik uygulamalarda faydalı ve uygulanabilir olduğunu ortaya koymuştur.

Bu çalışmanın genel bulguları, hibrit SC sistemlerinin geleceği için önemli sonuçlar

sunmaktadır. Özellikle, hibrit sistemlerin geliştirilmesi, büyük hacimli metin verilerinden bilgi edinme sürecinin erişilebilirliğini ve doğruluğunu önemli ölçüde artırma potansiyeline sahiptir. Çalışmamızda, hibrit SC sistemlerinde kullanılan T5 ve BERT modellerinin performanslarını karşılaştırmaktayız. Hibrit SC sistemlerini entegre ederek, bu alanı ileriye taşıma ve doğal dil işleme modellerinin doğruluk oranlarını iyileştirme hedefindeyiz.

5.2 Araştırma Soruları

Bu araştırmanın temel amacı, geniş veri kümesinden bilgilerin etkili ve doğru bir şekilde elde edilmesini sağlamak için hibrit bir soru-cevap sisteminin uygulanabilirliğini incelemektir. Bu bağlamda, çalışmada şu araştırma soruları ele alınacaktır:

- AS1: Tek bir arayüz aracılığıyla, farklı veri kümesi üzerinde eğitilmiş ve çeşitli platformlarda çalıştırılan çoklu sistemlere erişim sağlanabilen bir hibrit soru-cevaplama sistemi çerçevesinin oluşturulabilmesi için gereken teknik gereksinimler ve yazılım yapısı nedir?
- AS2: Birden fazla soru-cevaplama sistemine tek bir arayüz üzerinden erişim sağlanabilmesi için hibrit bir soru-cevaplama sistemi çerçevesi nasıl tasarlanabilir ve uygulanabilir?
- AS3: Farklı veri kümeleri üzerinde eğitilmiş ve farklı platformlarda çalışan soru-cevaplama sistemlerini hibrit bir çerçevede entegre etmek için göz önünde bulundurulması gereken temel unsurlar nelerdir?
- AS4: Hibrit soru-cevaplama sistemlerinin etkinliği nasıl değerlendirilebilir?

Bu araştırma soruları kapsamında, çalışma hibrit soru-cevaplama sistemlerinin geliştirilmesi, bilgiye erişilebilirliğin artırılması ve yanıtların doğruluğunun iyileştirilmesi konularında önemli çıkarımlar yapmayı amaçlamaktadır. Elde edilen bulguların, soru-cevaplama sistemlerinin gelecekteki gelişimi ve uygulama alanları açısından kayda değer katkılar sağlaması beklenmektedir.

5.3 Gereksinimler ve Kullanım Durumu

Bu bölümde, önerilen hibrit soru-cevaplama sisteminin gereksinimleri ve kullanım durumları ayrıntılı bir şekilde ele alınmaktadır. Sistemin etkin ve verimli bir şekilde işlev gösterebilmesi için belirli teknik ve işlevsel gereksinimlerin

karşılanması büyük önem taşımaktadır. Bu gereksinimlerin doğru bir şekilde belirlenmesi ve uygulanması, sistemin performansını ve güvenilirliğini doğrudan etkileyecektir. Ayrıca, hibrit SC sisteminin kullanım alanları ve olası senaryoları da bu bölümde açıklanmaktadır. Bu bağlamda, sistemin hangi durumlarda ve nasıl kullanılabileceği konusunda kapsamlı bir bakış açısı sunulmaktadır. Önerilen hibrit soru-cevaplama sisteminin gereksinimleri, teknik ve işlevsel özellikler temelinde aşağıdaki şekilde sınıflandırılmaktadır.

5.3.0.1 İşlevsel Gereksinimler

- Çoklu Sistem Entegrasyonu: Sistem, farklı veri kümeleri üzerinde eğitilmiş ve farklı platformlarda çalıştırılan birden fazla SC sistemini tek bir arayüz üzerinden entegre edebilmelidir.
- Dağıtıcı Modülü: Gelen sorguları ilgili alt SC sistemlerine yönlendirebilmeli ve bu sistemlerden gelen yanıtları birleştirebilmelidir.
- Veri Son İşleme Modülü: Alt sistemlerden gelen yanıtları işleyebilmeli ve bu yanıtları birleştirerek kullanıcıya sunabilmelidir. Bu işlemler arasında durdurma sözcüklerinin kaldırılması, köklendirme ve benzeri doğal dil işleme teknikleri yer almalıdır.
- Federatör Modülü: Birden fazla SC sisteminden gelen yanıtları birleştirme veya kesişim alma gibi federasyon yöntemlerini uygulayabilmelidir.
- Federatör Modülü: Birden fazla SC sisteminden gelen yanıtları birleştirme veya kesişim alma gibi federasyon yöntemlerini uygulayabilmelidir.
- Kullanıcı Arayüzü: Kullanıcıların sorgularını girebileceği ve sonuçları görebileceği kullanıcı dostu bir arayüz sağlanmalıdır.

5.3.0.2 Teknik Gereksinimler

- Platform Bağımsızlığı: Sistem, farklı platformlarda çalışan alt SC sistemleriyle uyumlu olmalıdır. Bu, sistemin platform bağımsız bir mimariye sahip olmasını gerektirir.
- Ölçeklenebilirlik: Sistem, yeni SC sistemlerinin eklenmesine izin verecek şekilde ölçeklenebilir olmalıdır.
- Performans: Sistem, kullanıcı sorgularına hızlı ve doğru yanıtlar verebilmeli ve yürütme süresi performans metriği açısından kabul edilebilir bir performans sergilemelidir.

- Güvenilirlik: Sistem, sürekli olarak çalışabilmeli ve hata durumlarında kullanıcıya bilgi verebilmeli ve hataları yönetebilmelidir.
- Veri Güvenliği: Sistem, kullanıcı verilerinin güvenliğini sağlamalı ve veri ihlallerine karşı koruma sağlamalıdır.

5.3.1 Kullanım Durumu

Önerilen hibrit SC sistemi, aşağıdaki kullanım durumlarında etkili bir şekilde kullanılabilir:

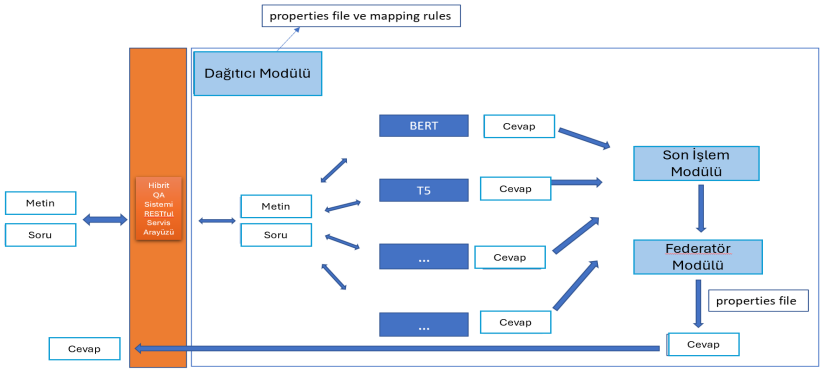
- Bilgi Alma: Kullanıcılar, farklı veri kümeleri üzerinde eğitilmiş birden fazla SC sistemine tek bir arayüz üzerinden erişerek, büyük miktardaki metin verilerinden bilgi alabilirler.
- Araştırma ve Geliştirme: Araştırmacılar, farklı doğal dil işleme modellerinin performanslarını karşılaştırmak ve hibrit sistemlerin etkinliğini değerlendirmek için bu sistemi kullanabilirler.
- Çok Dilli Uygulamalar: Sistem, farklı dillerdeki veri kümeleri üzerinde eğitilmiş SC sistemlerini entegre ederek, çok dilli bilgi alma uygulamalarında kullanılabilir.

Bu gereksinimler ve kullanım durumları, önerilen hibrit SC sisteminin hem teknik hem de işlevsel açıdan nasıl bir çözüm sunabileceğini ortaya koymaktadır. Sistem, bu gereksinimleri karşılayarak, kullanıcıların bilgiye erişimini kolaylaştıracak ve doğal dil işleme alanındaki araştırmaları ilerletebilecek bir potansiyele sahiptir.

5.4 Önerilen Metodoloji

Bu araştırmada, Şekil 5.1'de sunulan hibrit bir soru-cevaplama (SC) sistemi önerilmektedir. Bu önerilen yapı, kullanıcıların tek bir arayüz üzerinden şeffaf bir şekilde birden fazla SC sistemini kullanabilmesine olanak sağlar. Mimari, çeşitli alt SC sistemlerinden elde edilen yanıtları birleştiren modülleri içermektedir.

Sistemin bileşenleri arasında; Dağıtıcı Modülü, Veri Son İşleme Modülü ve Federatör Modülü bulunmaktadır. Her bir modülün işlevi aşağıda detaylı olarak açıklanmıştır. Dağıtıcı Modülü: Bu modül, gelen talepleri doğru alt SC sistemlerine iletmekle sorumludur. Eşleme kuralı dosyasını kullanarak, bu modül her bir alt sistem için yolları ve işlev adlarını eşler. Ayrıca, her bir SC alt sisteminin tahmin



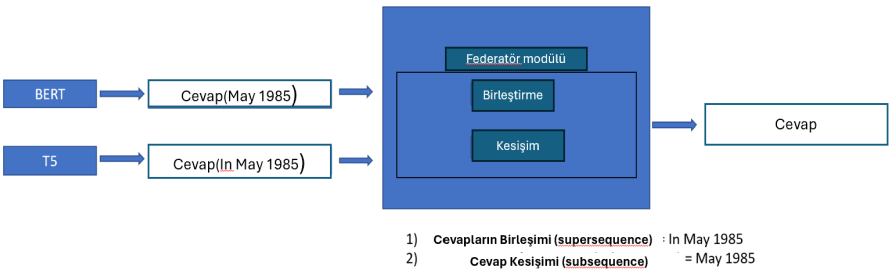
Şekil 5.1 Hibrit soru cevaplama mimarisine yaklaşımımız

işlevleriyle ilişkilendirilen sınıf ve işlev adlarını da içerir. Bu modül, önerilen hibrit SC sisteminin genel yapılandırma parametrelerini içeren bir özellik dosyasını da kullanır. Dağıtıcı modülünün tasarımı, alt sistemlerin teknolojik yapılarından bağımsızdır ve alt sistemlerin değişmesi, bu modülün işleyişini etkilemez. Veri-Son İşleme Modülü: Önerilen yapı sadece hibrit tahmin işlevlerini değil, aynı zamanda birleşik tahmin işlevlerini de aktif hale getirecektir. Eğer federatör tabanlı tahmin işlevi son kullanıcı tarafından talep edilirse, son işleme modülü, her bir alt sistemin soru ve yanıt sonuçlarını federatör kapasitesine uygun şekilde işler. Bu işlem, gereksiz kelimelerin çıkarılması, kök haline getirme gibi işlemleri de içerebilir.

Federatör Modülü: Bu modül, Şekil 5.2'de gösterilmiştir. Her biri bir federatif yöntemden sorumlu olan iki alt modül içerir. Burada, yalnızca birleşim ve kesişim tabanlı sonuç federasyonunu uyguladık. Bu yöntemlerin ayrıntılarını aşağıda tartışıyoruz. Bir dizi s'yi $< a_1, a_2 \ldots, a_r >$ ile gösteriyoruz, burada a_i bir kelime kümesidir ve aynı zamanda s'nin bir öğesi olarak da adlandırılır. Bir dizinin bir elemanını (veya bir öğe kümesini) $x_1, x_2, \ldots, x_k$ ile gösteririz, burada x_i bir öğedir. Uzunluğu k olan bir diziye k-dizisi denir. Soru-cevap sistemini bir dizi olarak tasvir edersek, bunu s dizisi <hello world, explore the world> olarak gösterebiliriz. a_1= hello world, a_2= explore the world a_1= {hello, world} a_2= {explore, the, world}

Birleşim: İki dizinin birleşimi, her iki dizideki tüm elemanları, yinelemeler olmadan içerir. Örneğin, eğer dizi A = [hello, world] ve dizi B = [explore, the, world] ise, bunların birleşimi A ∪ B = [hello, explore, the, world].

Kesişim: İki dizinin kesişimi yalnızca her iki dizide de görünen öğeleri içerir. Örneğin, dizi A = [hello, world] ve dizi B = [explore, the, world] ise, kesişimleri A ∩ B = [world].



Şekil 5.2 Dağıtıcı Modül

5.5 Prototip Uygulaması ve Deneysel Çalışma

Prototip: Önerilen yazılım mimarisini Python programlama dilini kullanarak Pytorch kütüphanesini (versiyon 1.13.1) kullanarak uyguladık. Dönüştürücü model olarak T5 ve BERT modellerini, belirteçleyici olarak T5tokenizer ve BertTokenizer'ı kullandık.

Veri Seti: Bu çalışmada, BERT ve T5 dönüştürücü modelleri kullanılarak hibrit SC sistemleri çalışması yapıldı. Bu çalışmada TQUAD veri seti kullanıldı. TQUAD (Türkçe Soru Cevaplama Veri Seti), Türkçe doğal dil işleme alanında kullanılmak üzere oluşturulmuş bir soru-cevap veri setidir. Bu veri seti, Türkçe doğal dil işleme alanında kullanılmak üzere Bilkent Üniversitesi araştırmacıları tarafından hazırlanmıştır. TQUAD toplam 27.503 soru-cevap çifti içerir ve bu çiftler otomatik olarak Türkçe Wikipedia sayfalarından çıkarılmıştır. Bu veri seti, Türkçe doğal dil işleme alanındaki araştırmaların ilerlemesine katkıda bulunabilir.

Sonuçlar Üzerine Deneysel Çalışma ve Tartışma: Bu çalışmada, hibrit bir soru-cevap sistemi tasarlanmış olup, farklı transformatör modelleri ile hibrit sistemin performans sonuçları incelenmiştir. Bu sonuçlar, T5 ve BERT modelleri kullanılarak elde edilen verilerle kıyaslanmış ve sıralı işlemler (birleşme ve kesişme gibi) sonucunda elde edilen yürütme süreleri ve standart sapmalar Tablo 5.1'de sunulmuştur.

BERT modelinden alınan her yanıtın yürütme süresi ve standart sapması, T5 ve hibrit yanıtlarla birlikte hesaplanmış ve otuz ardışık istek gönderilerek Tablo 5.1'te gösterilmiştir. Bu tablo, hibrit sistemin ek yük getirmeden daha hızlı yürütme süreleri sağladığını ve sistemin işlem süresinin ihmal edilebilir düzeyde olduğunu ortaya koymaktadır. Çalışmamızın amacı, birden fazla transformatör modelini hibrit bir yapıya entegre ederek daha verimli ve kullanışlı bir sistem tasarlamaktır. İncelenen sonuçlar, hibrit soru-cevap sisteminin düşük işlem yüküyle rekabetçi bir yürütme performansı sunduğunu göstermektedir, böylece hibrit sistemlerin etkinliğini ve başarısını doğrulamaktadır.

Tablo 5.1 Hibrit Soru Cevaplama için Yürütme Performansı

	Hibrit T5 (ms)	T5 (ms)	Hibrit BERT (ms)	BERT (ms)
avg	699.58	692.64	92.19	86.71
std	24.45	17.84	7.89	5.44

6
SORU CEVAPLAMA SİSTEMLERİ İÇİN MODEL TABANLI BİR DEĞERLENDİRME METRİĞİ

6.1 Genel Yaklaşım

SC sistemlerinde önerilen yanıtların kalitesine yönelik sağlam bir değerlendirme yapısının bulunmaması, SC sistemlerinin güvenilirliğine ilişkin önemli bir sorun teşkil etmektedir. Mevcut değerlendirme yöntemleri ağırlıklı olarak kullanıcı geri bildirimlerine ve temel yanıt analizine dayanmaktadır. Ancak bu yöntemler, yanıtların karmaşıklık açısından yeterince kapsamlı bir şekilde değerlendirilmesini sağlamamaktadır. Bu durum, SC sistemlerinin karmaşık terminolojileri ve kavramları anlama ve işleme yeteneğinin belirsiz kalmasına neden olmakta; teorik yetenekler ile pratik performans arasında bir uyumsuzluk ortaya çıkarmaktadır. Geleneksel değerlendirme metrikleri büyük ölçüde benzerlik temelli karşılaştırmalara dayanmakta ve bu nedenle yanıtların yorumlama ve çıkarımlarla dolu olduğu bağlamlarda yetersiz kalmaktadır. Bu tür metrikler, yüzeysel karşılaştırmalara ağırlık vererek kelime çarpıtmaları, eş anlamlı kullanımlar ve benzer yapılar gibi dilsel nüansları göz ardı etmektedir. Bu eksiklik, yanıtların semantik doğruluğu ve bağlamsal uygunluğunun yeterince değerlendirilememesiyle sonuçlanmakta ve SC sistemlerinin ürettiği yanıtların doğruluğunu olumsuz etkilemektedir. Sonuç olarak, geleneksel değerlendirme yöntemleri, karmaşık ve anlam yüklü yanıtları değerlendirirken yetersiz kalmakta, bu da SC sistemlerinin değerlendirilmesinde güvenilirlik sorunlarına yol açmaktadır. Dolayısıyla, SC sistemlerinin yanıt kalitesini daha doğru ve nesnel bir şekilde değerlendirebilecek yenilikçi metriklerin geliştirilmesi gerekliliği ortaya çıkmaktadır.

SC sistemlerinde, BLEU [13], ROUGE [14] ve METEOR [15] gibi klasik değerlendirme metrikleri, tahmin edilen yanıtın doğruluğunu değerlendirmek amacıyla kullanılmaktadır. Ancak bu metriklerin SC sistemlerinde kullanımının eleştirel bir bakış açısıyla incelenmesi gerekmektedir, çünkü söz konusu metrikler çeviri sistemleri için geliştirilmiştir ve aday çeviri metni ile referans metnin

uzunluklarının yaklaşık olarak aynı olduğu varsayımına dayanmaktadır. Bu metriklerde, referans metni üretildikten sonra aday metnin uzunluğu daha kısa olduğunda veya uzunluk dengesi sağlanmadığında (örneğin bir örnekte olduğu gibi), puan ceza oranı kadar azaltılmaktadır. Ayrıca, kesişim kelimelerinin sayısı, referans veya tahmin edilen metnin uzunluğuna bölünerek hesaplanmaktadır.

Buna karşılık, günümüzde üretilen üretken modeller, SC sistemlerinde daha ayrıntılı ve uzun tahmin edilen yanıtlar üretmekte olup, bu yanıtların gerçek cevaptan daha uzun olması yanlış oldukları anlamına gelmemektedir. Dahası, bu geleneksel metriklerin hesaplama formülleri kesişen kelime sayısına dayalı olduğundan, eş anlamlı kelimeler üretilmesi durumunda bu metrikler başarısız olmakta ve sistemin değerlendirilmesinde hatalı sonuçlar vermektedir. Bu bağlamda, çeviri sistemleri için geliştirilen geleneksel metriklerin SC sistemlerini değerlendirmek amacıyla kullanılması anlamlı değildir.

SC sistemlerinde daha doğru bir değerlendirme yapabilmek için, model tabanlı yöntemlere dayanan gelişmiş metriklerin kullanımı gerekmektedir. Büyük dil modellerindeki son gelişmelerden yararlanan model tabanlı yaklaşımlar, bu karmaşık değerlendirme gereksinimlerine daha iyi uyum sağlama potansiyeline sahiptir. Bu yaklaşımlar, yalnızca tahmin edilen yanıtın yüzeysel benzerliğini değil, aynı zamanda içeriğin anlamlılığı, uygulanabilirliği gibi çok boyutlu değerlendirme kriterlerini de dikkate almaktadır. Model tabanlı çalışma yapabilmek için yüksek kaliteli bir veri kümesine ihtiyaç vardır. Bu çalışmada, güçlü bir squad-qametrik veri kümesi sağlayarak metriklerin üretilmesine katkıda bulunuyoruz. Bu veri kümesini kullanarak oluşturduğumuz MQA-metrik (Mistral Question Answer) modeli, geleneksel değerlendirme yöntemlerine kıyasla üstün performans göstermektedir. Bu tezde, ağırlıklı olarak sözdizimi ve n-gram benzerliğine dayanan SC sistemleri için geleneksel değerlendirme metriklerinin sınırlamalarını ele almayı amaçlıyoruz. Bu geleneksel metrikler, SC sistemlerini doğru bir şekilde değerlendirmek için gereken anlamsal doğruluğu yakalamada ve eşit uzunlukta olmayan yanıtların değerlendirilmesinde genellikle başarısız olur. Bu sınırlamaların üstesinden gelmek için, sözdizimsel benzerlikten ziyade anlamsal değerlendirmeye odaklanmak üzere tasarlanmış yeni bir model tabanlı değerlendirme metriği olan MQA-metriği öneriyoruz. Bu araştırma, veri kümesi etiketlemesinde nesnelliği sağlama, gerçek ve tahmin edilen yanıtlar arasında sözdizimi benzerliği olmadığında metriklerin etkinliğini değerlendirme ve tahmin edilen yanıtın uzunluğunun metrik performansı üzerindeki etkisini değerlendirme gibi birkaç temel zorluğu ele alıyor. Ek olarak, gerçek yanıtların kalitesinin metrik sonuçlarını nasıl etkilediğini araştırıyoruz. Bu zorluklar, MQA-metriğinin geliştirilmesine ve doğrulanmasına rehberlik ediyor ve önerilen yaklaşımımızın

temelini oluşturuyor. Bu çalışmada, özellikle SC sistemleri için tasarlanmış, MQA-metriği model tabanlı metriği tanıtıyoruz. Öncelikli olarak sözdizimsel karşılaştırmalara dayanan mevcut metriklerin aksine, MQA-metriği yanıtların doğruluğunu değerlendirmek için anlamsal anlayıştan yararlanır. Bu yaklaşım, büyük dil modellerinin (BDM) alanındaki son gelişmelerden yararlanarak, yüzeysel benzerliklerin ötesine geçen karmaşık yanıtların daha derinlemesine değerlendirilmesine olanak tanır. Eş anlamlı kelimeleri tanıma, bağlamsal doğruluk ve yanıt uzunluğundaki değişkenlik gibi zorlukları ele alarak önerilen MQA-metriği, SC sistemleri için uyarlanmış daha sağlam bir değerlendirme çerçevesi sağlar. Bu metrik, anlamsal doğruluğu ölçmek için değerlendirme sürecini iyileştirerek SC değerlendirmesi alanında ileriye doğru umut verici bir adım haline getirir.

6.2 Araştırma Soruları

Bu çalışma, SC sistemlerinin değerlendirilmesinde kullanılan geleneksel metriklerin sözdizimi ve n-gram benzerliğine odaklanmakta ve SC sistemlerinin doğru değerlendirilmesi için gerekli olan anlamsal doğruluğu ve eşit olmayan uzunlukta olan yanıtlarda yetersiz kalması gibi sınırlamalarını ele almayı amaçlamaktadır. Geleneksel metriklerin bu eksikliklerini gidermek amacıyla, MQA-metrik olarak adlandırılan yeni bir model tabanlı değerlendirme metriği geliştirilmiştir. Bu yaklaşımın geçerliliğini test etmek için ise insan yargısına dayalı iki veri kümesi, "squad-qametrik" ve "marco-qametrik", oluşturulmuştur. Bu çalışmada ele alınan temel zorlukları veri kümesi etiketlemede nesnellik sağlama, gerçek ve tahmin edilen yanıtlar arasında sözdizimsel benzerlik bulunmadığında metrik etkinliğinin korunması, tahmin edilen yanıt uzunluğunun metrik performansı üzerindeki etkisinin belirlenmesi ve gerçek yanıt kalitesinin metrik sonuçları üzerindeki etkisinin değerlendirilmesidir. Bu araştırma sorunlarını çözmek amacıyla geliştirilen araştırma soruları aşağıda sunulmuştur:

- AS1: Oluşturulan squad-qametrik ve marco-qametrik veri kümeleri nasıl oluşturulmuştur?
- AS2: Soru-Cevap Sistemlerinin sonuçlarının doğruluğunu değerlendirmek için model tabanlı bir değerlendirme kriteri oluşturma sürecinde hangi iş akışı tasarımı benimsenmiştir?
- AS3: Veri kümesi etiketleme işlemi için oluşturulan arayüzler nasıl tasarlanmıştır?

- AS4: Soru-cevap sistemlerinin yanıt kalitesini değerlendirmek amacıyla geliştirilen MQA-metrik modelinin etkinliği nasıl ölçülebilir?
- AS5: Klasik metriklere kıyasla MQA-metrik modelinin sağladığı avantajlar nelerdir?
- AS6: İnsan puanları ile klasik metrikler arasındaki korelasyon düzeyi nedir?
- AS7: İnsan puanları ile MQA-metrik arasındaki korelasyon düzeyi nedir?

6.3 Gereksinimler ve Kullanım Durumu

Bu bölümde, önerilen model tabanlı değerlendirme metriği olan MQA-metriğinin gereksinimleri ve kullanım durumları detaylandırılmaktadır. Sistemin başarılı bir şekilde işlev görebilmesi için belirli teknik ve işlevsel gereksinimlerin karşılanması gerekmektedir. Ayrıca, bu metriğin hangi senaryolarda ve nasıl kullanılabileceği de bu bölümde açıklanmaktadır. Anlamsal Değerlendirme: Metrik, yanıtların sözdizimsel benzerliğinden ziyade anlamsal doğruluğunu değerlendirebilmelidir. Bu, eş anlamlılar ve farklı ifadeler kullanıldığında bile yanıtların doğruluğunu ölçebilme yeteneğini gerektirir. Uzun ve Ayrıntılı Yanıtların Değerlendirilmesi: Metrik, üretken modeller tarafından üretilen uzun ve ayrıntılı yanıtları değerlendirebilmelidir. Geleneksel metriklerin aksine, yanıt uzunluğundan bağımsız olarak doğruluğu ölçebilmelidir. İnsan Yargısıyla Uyum: Metrik, insan yargısıyla yüksek korelasyon göstermelidir. Bu, metrik sonuçlarının insan değerlendirmelerine yakın olmasını gerektirir. Önerilen MQA-metriği, aşağıdaki kullanım durumlarında etkili bir şekilde kullanılabilir:

SC Sistemlerinin Değerlendirilmesi: MQA-metriği, farklı SC sistemlerinin performansını değerlendirmek için kullanılabilir. Bu, özellikle üretken modeller tarafından üretilen uzun, ayrıntılı veya eşanlamlı yanıtların doğruluğunu ölçmek için idealdir. Araştırma ve Geliştirme: Araştırmacılar, farklı doğal dil işleme modellerinin performanslarını karşılaştırmak ve hibrit sistemlerin etkinliğini değerlendirmek için bu metriği kullanabilirler. Uygulamalar: Hukuk, tıp gibi özel alanlarda kullanılan SC sistemlerinin değerlendirilmesinde MQA-metriği, alanın terminolojileri ve bağlamları dikkate alarak daha doğru sonuçlar üretebilir. Bu gereksinimler ve kullanım durumları, önerilen MQA-metriğinin hem teknik hem de işlevsel açıdan nasıl bir çözüm sunabileceğini ortaya koymaktadır. Metrik, bu gereksinimleri karşılayarak, kullanıcıların bilgiye erişimini kolaylaştıracak ve doğal dil işleme alanındaki araştırmaları ilerletebilecek bir potansiyele sahiptir.

6.4 Temel Kavramlar

ROUGE'un temel amacı, otomatik olarak oluşturulan metin özetlerini veya çevirilerini referans metinlerle karşılaştırmak ve benzerliklerini ölçmektir. Metin kalitesinin daha kapsamlı bir değerlendirmesini sunmak için iyi bilinen ROUGE metriğine dayanır. ROUGE, metni netliği, kesinliği ve önemi açısından değerlendirir, nihayetinde bir kalite puanı atar ve hataları belirler. ROUGE'un gücü ayrıntılı yaklaşımında olsa da, insan önyargısı, değerlendirmeler için gereken daha fazla zaman ve kaynak ve çeşitli dillerde olası uygulama sorunları gibi zorlukları da sunar. Genel olarak, ROUGE metin oluşturma sistemlerinin daha derinlemesine bir değerlendirmesini sağlar ancak öznel ve dil özgü yönlerinden kaynaklanan dikkatli bir değerlendirme gerektirir [14].

ROUGE-L, hem doğruluk (precision) hem de kapsayıcılık (recall) değerlerini göz önünde bulunduran F1 tabanlı bir formülle hesaplanmaktadır. Aşağıda gösterildiği üzere, LCS(X,Y) ile model çıktısı X ve referans cevap Y arasındaki en uzun ortak alt dizi uzunluğu hesaplanır. Bu değer, ayrı ayrı doğruluk ve kapsayıcılık oranlarına dönüştürülerek aşağıdaki denklemde 6.1, 6.2, 6.3 birleştirilir:

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + \beta^2 \cdot \text{Precision}} \tag{6.1}$$

$$\text{Precision} = \frac{\text{LCS}(X,Y)}{|X|}, \tag{6.2}$$

$$\text{Recall} = \frac{\text{LCS}(X,Y)}{|Y|} \tag{6.3}$$

BLEU (Bilingual Evaluation Understudy), makine çevirisi çıktılarının referans çevirilerle n-gram tabanlı benzerliklerini ölçen klasik ve yaygın kullanılan bir değerlendirme metriğidir. Özellikle otomatik çeviri ve metin üretim sistemlerinde, modelin çıktısının kelime düzeyinde referansla ne kadar örtüştüğünü ölçmek amacıyla geliştirilmiştir. BLEU metriği denklem 6.4, 6.5 de görüldüğü gibi, model çıktısı ve referans cümleler arasındaki n-gram eşleşmeleri üzerinden doğruluk (precision) hesaplaması yapar. Bununla birlikte, modelin çok kısa cevaplar üretmesini engellemek için ayrıca bir penalizasyon katsayısı (brevity penalty) da içerir. Bu sayede hem n-gram benzerliği yüksek olan hem de uzunluk bakımından referansa yakın cevaplara daha yüksek puan verilir.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{6.4}$$

$$\text{BP} = \begin{cases} 1 & \text{eğer } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{eğer } c \leq r \end{cases} \tag{6.5}$$

Tanımlar:

- p_n: n-gram düzeyinde doğruluk (precision)
- w_n: n-gram ağırlıkları (genellikle eşit, $w_n = \frac{1}{N}$)
- c: model çıktısının uzunluğu
- r: referans cevabın uzunluğu

METEOR, özellikle makine çeviri sistemlerinin değerlendirilmesinde, doğal dil işleme alanında kullanılan bir değerlendirme metriğidir. Kelime seçimi ve düzenlemesini vurgular, benzer kelime dağarcığı kullanımını ve doğru dizilimi sağlamak için çevrilmiş metinleri referans metinlerle karşılaştırır. METEOR, denkleminde 6.6, 6.7, 6.8 görüldüğü gibi çevrilen ve referans metinler arasındaki kelime eşleşmelerinin doğruluğunu değerlendirir ve benzerliği 0 ile 1 arasında bir ölçekte ölçer (yüksek puan önemli benzerliği gösterir). Kapsamlı bir analiz sağlar ve güvenilirliği artırmak için insan tarafından çevrilmiş referans metinleri kullanır [15]. F1 puanı, hassasiyet ve geri çağırmanın harmonik ortalamasını temsil eden bir performans metriğidir; sınıflandırma problemlerinde yaygın olarak kullanılır ve özellikle dengesiz veri kümelerinde faydalıdır. F1 puanı, bir modelin doğru pozitif tahminlerinin doğruluğunu (hassasiyet) ve bu doğru pozitif tahminlerin eksiksizliğini (geri çağırma) tek bir değerde birleştirir. Hassasiyet, doğru pozitif tahminlerin model tarafından yapılan toplam pozitif tahminlere oranıdır; başka bir deyişle, model tarafından pozitif olarak tahmin edilen örneklerin kaçının gerçekten pozitif olduğunu gösterir. Geri çağırma, doğru pozitif tahminlerin toplam gerçek pozitif örneklere oranıdır; başka bir deyişle, modelin tüm pozitif örnekleri ne kadar iyi tanımlayabildiğini gösterir.

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{6.6}$$

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \tag{6.7}$$

$$\text{Penalty} = \gamma \left(\frac{ch}{m}\right)^{\theta} \tag{6.8}$$

Tanımlar:

- P: Doğruluk (Precision)

- R: Kapsayıcılık (Recall)
- ch: Sıralı olmayan eşleşme parçacıklarının (chunk) sayısı
- m: Eşleşen toplam kelime sayısı
- γ, θ: Penalty katsayıları (genellikle deneysel olarak belirlenir)

F1 skoru, bilgi erişimi, sınıflandırma ve metin üretimi gibi görevlerde kullanılan bir değerlendirme metriğidir. Denklem 6.9 de görüldüğü üzere Precision (kesinlik) ve Recall (duyarlılık) değerlerinin harmonik ortalamasını alarak hesaplanır. Bu sayede, sadece doğru pozitifleri değil, aynı zamanda kaçırılan (false negative) örnekleri de dikkate alır. F1 skoru özellikle dengesiz veri kümelerinde (örneğin pozitif örneklerin az olduğu durumlarda) daha anlamlı bir değerlendirme sağlar. F1 skoru, precision ve recall değerlerinin harmonik ortalaması olarak aşağıdaki formülle hesaplanır:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{6.9}$$

Tanımlar:

- $P = \frac{TP}{TP+FP}$: Doğruluk (Precision)
- $R = \frac{TP}{TP+FN}$: Duyarlılık (Recall)
- TP: Doğru pozitif sayısı
- FP: Yanlış pozitif sayısı
- FN: Yanlış negatif sayısı

Kosinüs benzerliği, özellikle metin madenciliği, bilgi alma ve doğal dil işleme gibi alanlarda yaygın olarak kullanılan iki vektör arasındaki benzerliği değerlendirmek için kullanılan bir ölçüdür. Kosinüs benzerliği denklem 6.10 de görüldüğü gibi, iki vektör arasındaki açının kosinüsünü hesaplar; bu, iki vektör arasındaki yön benzerliğini temsil eder ve -1 ile 1 arasında değişir.

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \tag{6.10}$$

Tanımlar:

- $\vec{A}, \vec{B}$: Karşılaştırılan metin vektörleri
- $\vec{A} \cdot \vec{B}$: Nokta çarpımı

- $\|\vec{A}\|$: Vektör A'nın normu
- $\|\vec{B}\|$: Vektör B'nin normu

Kendall's Tau denklem 6.11 da görüldüğü gibi, iki sıralı değişken (ordinal veri) arasındaki ilişkiyi ölçen bir istatistiksel metriktir. Değer aralığı –1 ile +1 arasındadır: +1 → sıralamalar tamamen uyumlu 0 → ilişki yok –1 → sıralamalar tamamen zıt Temel mantığı, tüm sıralı çiftler arasında uyumlu (concordant) ve zıt (discordant) olanların farkına dayanır. Özellikle küçük veri setleri ve bağımsız sıralama sistemleri arasındaki tutarlılığı incelemede etkilidir. Kendall korelasyon katsayısı, sıralı iki değişken arasındaki ilişkinin gücünü ölçer:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{6.11}$$

Tanımlar:

- C: Uyumlu çift (concordant pair) sayısı
- D: Zıt çift (discordant pair) sayısı
- n: Gözlem sayısı

Spearman korelasyonu denklem 6.12 de görüldüğü gibi, iki değişkenin sıraları (rank'leri) arasındaki monotonik ilişkiyi ölçer. Pearson'dan farklı olarak değerlerin kendisi değil, sıraları kullanılır. Özellikle: Dağılımı bilinmeyen, Aykırı değerlere duyarlı olmayan, Sıralı veri (ordinal) içeren durumlar için uygundur.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{6.12}$$

6.5 Literatür İncelemesi

Son yıllarda, açıklamaların, metin oluşturma süreçlerinin, özetlerin ve doğal dil oluşturma dahil diğer ayrıntıların değerlendirilmesinde otomatik ölçümlerin incelenmesine önemli bir önem verilmiştir. N-gram tabanlı benzerlik ölçümlerini hesaplayan geleneksel metrikler geliştirilmiştir ve önceki bilim insanları bu metriklerin dil görevleriyle düşük korelasyonlara sahip olduğunu gösteren çalışmalar üretmişlerdir. Bu ölçümlerin kullanılmasına rağmen eleştirilmiştir (örneğin, [151–153]). BERTScore [154] üretilen metni değerlendirmek için bağlamsal yerleştirmeyle ölçüm değeri sağlar. Haber özetlerinin değerlendirilmesinde yaygın olarak kullanılan ölçütlerin çoğu yüksek ROUGE

puanlarına odaklanır [155]. Ancak yaklaşımımız, n-gram örtüşmesine veya yüzeysel benzerliğe güvenmek yerine daha derin anlamsal anlamı yakalamaya odaklanarak önceki çalışmalardan ayrılır. Sözdizimsel yapıyı vurgulayan BERTScore ve ROUGE'un aksine, MQA metriği bağlamsal doğruluğu ve nüansı değerlendirmek için büyük dil modellerinden yararlanır ve geleneksel metriklerin kelime farklılıkları nedeniyle gözden kaçırabileceği anlamsal olarak doğru yanıtları değerlendirmesine olanak tanır.

APES'in tanıtımı, okunan metnin anlaşılmasındaki ilerlemelerden yararlanarak bir paradigma değişimine yol açmıştır. APES, özetin ana fikirleriyle ilgili olarak dikkatlice seçilmiş soruları yanıtlama yeteneğini değerlendirerek daha kapsamlı bir değerlendirme sağlar. Piramit yöntemi, bu yöntemi otomatikleştirme çabalarını içerir ve insan tarafından yazılmış referans özetlerinin temel konuşma birimlerinden (EDU'lar) çıkarımsal özetlere dönüştürülmesini önerir. Bu yöntem, tanınan veri kümelerinin çeşitli manuel taramaları arasında güçlü bir bağlantı olduğunu göstermiştir. Jones ve Galliers, kaliteyi dahili ve harici olarak ölçen özet sistemlerinin değerlendirilmesini tartışmaktadır [156]. Otomatik olarak oluşturulan özetler yararlı olmayabilir. TIPSTER, özetlerin soruyla alakalı olup olmadığını belirleme, kategoriyi belirleme vb. görevler yoluyla özetleri değerlendirmeye odaklanır ve ardından bunu insan tarafından oluşturulan model özetiyle karşılaştırır. Jing ve diğerleri, farklı sistemlerin farklı uzunluklarda iyi performans gösterdiğini buldukları bir pilot deney yürüttüler. Diğer sistemler, paragrafı bir özet birimi olarak kullanır. Birden fazla cümleden oluşan bir paragraf, bu tür bir değerlendirme için uygun değildir. Makine tarafından oluşturulan özetler için, cümle taşıyıcının benzerliği, bağlamdaki her cümle için yerleştirmeyi hesaplar. Bağlamlar, ELMo ile Wikipedia belgesinden alınan özetlerdir. HEQ, konuşma SC veri kümesi QuAC tarafından önerilen insan tabanlı puandır. Tahminlerin F1 puanının insan-F1 puanını aştığı veya onunla eşleştiği durumların yüzdesini ölçer. Mutabakat Tabanlı Kod Özetleme Değerlendirmesi, n-gramlara farklı ağırlıklar vermek için bir algoritma kullanarak bir ölçüm sağlar. Chakraborty ve diğerleri tarafından yapılan bir çalışmada [157], iki metnin benzerliği Bert Embedding vektörü kullanılarak incelenir. Ancak, yaklaşımı yüzeysel özet veya n-gram örtüşmeleri yerine anlamsal derinliğe ve terimlere odaklanarak önceki çalışmalardan farklıdır. Çıkarımsal özetler için EDU dönüşümüne dayanan piramit yaklaşımı gibi yöntemlerin aksine, MQA metriği bağlamsal olarak zengin ve eşanlamlılarla dolu yanıtları değerlendirmek için büyük dil modellerinden yararlanır ve SC sistemlerinde anlamsal tutarlılık ve içerik hizalamasının daha doğru bir değerlendirmesini sunar.

PyrEval [158], daha fazla cümle benzerliği hesaplayarak SCU'yu hesaplar; metin benzerliğine dayanan diğer metriklerden daha az başarılı görülmüştür. Metrikler

üzerine yapılan çalışmaların çoğu özetleri değerlendirir ve SC sistemleri burada da kullanılabilir [159]. Eyal ve diğerleri [155], metrikleri hesaplarken APES'i kullanır ve özeti değerlendirmek için SC'yı APES ile kullanır. Ad varlıkları kaldırılır ve boşluklar yerine boşluk doldurma soruları yerleştirilir ve görev hangi varlığın kaldırıldığını tahmin etmektir. FEQA [160], özeti girdi belgeleriyle karşılaştırır ve özetin doğru bilgi içerip içermediğini test eder. Ancak, varlık kaldırma veya özetlerde olgusal hizalama gibi belirli görevlere odaklanan APES veya FEQA gibi önceki yaklaşımların aksine, MQA-metriğimiz yanıtların bağlamsal derinliğini yakalayarak SC sistemlerindeki yanıtların anlamsal tutarlılığını değerlendirmek üzere tasarlanmıştır. Bu model tabanlı metrik, cümle düzeyindeki benzerliğin sınırlarını aşarak daha sağlam bir değerlendirme elde etmek için büyük dil modellerinden yararlanır ve eş anlamlıların anlaşılmasına öncelik verir.

Araştırma çalışmaları, bilgi grafikleri yerine SC sistemlerini, metin özetlemeyi ve varlık bağlantısını iyileştirmeye odaklanır. Bir makale, yapılandırılmamış web metinlerinden gelen özel bilgileri entegre eden ve yanıt seçimini geliştirmek için seçici dikkati kullanarak en son teknoloji sonuçları elde eden iki aşamalı bir model önermektedir [161]. Başka bir çalışma, değiştirilmiş BM25'i kelime yerleştirmeleriyle birleştiren ve DUC veri kümeleri [162] üzerinde gösterilen daha etkili cümle düzeyinde temsil ve sıralamaya yol açan bir belge özetleme tekniğini tanıtmaktadır. Benzer şekilde, "Temizle ve Öğren" modeli, aday yanıtları filtreleyerek API SC'daki sahte çözümleri ele alır ve APIQASet [163] üzerinde yüksek doğruluk elde eder. Ek olarak, KG-SimpleQA için geliştirilmiş varlık bağlantısı önerilmiştir, varlık belirsizliği ve sözcük dağarcığı sorunlarının üstesinden gelinmiştir ve bu da daha iyi genel performansla sonuçlanmıştır [164]. Ancak, bu yaklaşım sözdizimsel benzerlikten ziyade anlamsal anlayışı önceliklendiren model tabanlı bir değerlendirme metriği olan MQA metriği geliştirerek önceki çalışmalardan ayrılıyor. N-gram eşleştirmesine veya yüzeysel karşılaştırmalara büyük ölçüde dayanan geleneksel metriklerin aksine, metriğimiz daha derin bağlamsal doğruluğu ve eşanlamlı tanımayı değerlendirmek için büyük dil modellerinden yararlanır ve SC sistemlerinde daha ayrıntılı ve güvenilir değerlendirmelere olanak tanır.

Diğer çalışmalar, hizmet kalitesinin değerlendirilmesine ve hiyerarşik yanıt özetleme süreçlerine vurgu yapmaktadır. Örneğin, SC sitelerinde bir sınıflandırma modeli, BERT ve DistilBERT sınıflandırıcılarını kullanarak tahmin edilen etiketler ile yanıt kalitesi arasında güçlü bir korelasyon olduğunu göstermektedir [165]. Emeklilik hizmetlerinin kalitesini değerlendirmeye yönelik bir diğer çalışma ise, ifade, duygu ve kelime özelliklerini içeren çok boyutlu bir dikkat evrişimli sinir ağı modeli kullanmakta ve bu yaklaşımın geleneksel yöntemlere kıyasla daha

nesnel olduğu kanıtlanmaktadır [166]. Topluluk tabanlı SC sistemlerinde ise, yanıtların özetlenmesi için önerilen hiyerarşik yapı, anket sorularının eksiksizliğini ele almakta ve hatırlama, kesinlik ve özlü olma açısından önemli iyileştirmeler sağlamaktadır [167].

Ancak, bu çalışmadan farklı olarak, önerilen yaklaşım belirli model çıktılarının (örneğin hizmet kalitesi veya özetleme) iyileştirilmesine odaklanmak yerine, büyük dil modellerini kullanarak anlamsal doğruluğu değerlendiren bir metrik geliştirmeye yönelmektedir. Bu model tabanlı değerlendirme metriği, bağlamsal ifadeleri ve eşanlamlı kelimeleri yakalamakta olup, SC sistemlerinin dinamik değerlendirme gereksinimlerine daha iyi uyum sağlamaktadır. Çalışmalar ayrıca bilgi erişimi, web içeriği madenciliği ve hizmet benzerliği ölçümündeki ilerlemeleri de araştırıyor. Toshiba'nın araştırması, BRIDJE sisteminin başarılarını vurgulayarak diller arası bilgi alma konusundaki yenilikleri özetliyor [168]. Başka bir çalışma, daha iyi özellik çıkarma için adlandırılmış varlık tanıma ve metin segmentasyonunu entegre ederek geliştirilmiş anlamsal segmentasyon yoluyla web içeriği madenciliğini geliştiriyor [169]. Hizmet odaklı geliştirmelerde hizmet benzerliğinin ölçümü, daha iyi analiz ve yeniden kullanılabilirliği kolaylaştıran XL-BPMN modelleri kullanılarak resmileştirilir [170]. Son olarak, deneysel bir çalışma, GitHub Copilot gibi AI araçlarının yazılım mühendisliğindeki etkinliğini inceleyerek AI destekli programlamanın görev tamamlamayı ve özellik uygulamasını iyileştirdiğini gösteriyor [171]. Ancak, alma veya içerik segmentasyonunu iyileştirmeye odaklanan önceki çalışmalardan farklı olarak, bu yaklaşım SC sistemleri içinde anlamsal doğruluğu yakalamak için özel olarak tasarlanmış bir model tabanlı değerlendirme metriği sunuyor. Büyük dil modellerinden yararlanarak, yalnızca sözdizimini değil, aynı zamanda bağlamsal ifadeleri da değerlendiriyor ve insan yargısıyla daha iyi korelasyon sağlayan ve yüzeysel benzerlik ölçümlerinin ötesine geçen daha derin bir değerlendirmeye olanak tanıyor.

Wang ve diğerleri, veri kümeleri olarak EVOUNA'yı kullanarak açık alan SC sistemlerinde doğruluğu değerlendirmek için BDM'lere dayalı QA-Eval değerlendirme metriğini geliştirdiler. Bu çalışma, BDM'lerle geliştirilen QA-Eval'in etkinliğini göstermektedir. BDM'lerin bu metrikteki başarısı böylece literatürde belirlenmiş ve kabul edilmiştir. Ek olarak, KPQA-metrik [172], SC sistemlerinin değerlendirme metriğini ölçmek için anahtar kelimelere ağırlıklar verir.Bu ağırlıklarla oluşturulan yanıtın referans yanıta yakınlığını ölçer. Çalışmamızda yaptığımız gibi, bu çalışmada yanıtlar BERT ile çıkarıldı. Tahmin edilen yanıt ile gerçek yanıt arasındaki benzerliği değerlendirmek için 10 çalışana Likert ölçeğinde 1 ile 5 arasında değerler verilir. Buna karşılık, çalışmamız

BDM'leri basitçe dahil etmenin ötesine geçerek, metrik üretimi için BDM'lerin anlamsal anlama yeteneklerini kullanan bir teknik uygular. N-gram örtüşmesine veya anahtar kelime tabanlı ölçümlere dayanan geleneksel metriklerin aksine, bu yaklaşım BDM'lerin bağlamsal anlama ve eşanlamlı tanıma kapasitesini kullanır. Bu, farklı ifadeler veya eşanlamlılar kullanıldığında bile tahmin edilen ve referans cevaplar arasındaki anlamsal yakınlığı doğru bir şekilde değerlendirebilen daha ayrıntılı metrik üretimine olanak tanır.

Çok sayıda çalışma, çeşitli alanlarda hesaplamalı ve metrik tabanlı değerlendirme çerçevelerini araştırmıştır. Aydin ve ark. [173] ve Nacar ve ark. [174], altyapı optimizasyonuna vurgu yaparak, coğrafi bilgi sistemlerinin ve işbirlikçi Grid hizmetlerinin mekansal ve hesaplamalı verileri etkili bir şekilde yönetmek ve işlemek için kullanımını araştırmıştır. Fox ve ark. [175] ve Aktas ve ark. [176], jeofizik veri işleme araçlarını deprem araştırmalarındaki belirli uygulamalarla entegre ederek Grid hesaplamanın uygulamasını ilerletti, QuakeSim gibi platformlar aracılığıyla [177]. Ayrık ancak alakalı bir bağlamda, Kapdan ve diğerleri [178] ve Sahinoglu ve diğerleri [179] kod metriklerine ve doğrulama çerçevelerine odaklandı ve yazılım ve mobil uygulama geliştirmede sistematik değerlendirmenin kritik rolünü vurguladı. Bu çalışmalar öncelikli olarak altyapı veya yapısal değerlendirmeyi vurgularken, mevcut çalışma, geleneksel sözdizimi odaklı yaklaşımlara göre bağlamsal ve anlamsal doğruluğu önceliklendirerek SC sistemlerini değerlendirmek için büyük dil modellerinden yararlanan bir anlamsal değerlendirme metriği olan MQA-metrik'i tanıtıyor.

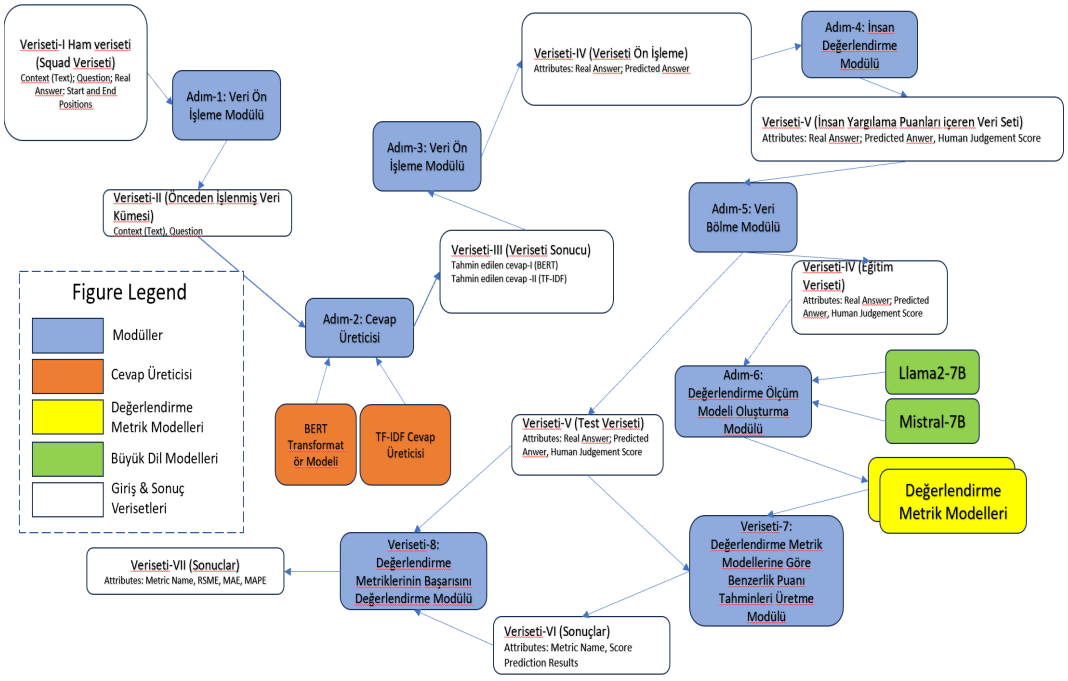
Briman ve Yıldız [180], ROUGE gibi geleneksel metriklerin sınırlamalarını aşmak için benzerlik, gereklilik ve kabul edilebilirlikten yararlanarak soyutlayıcı özetleme için kapsamlı bir değerlendirme metriği tanıttı. Benzer şekilde, önceki çalışmamız [16]'de, özellikle hukuki SC sistemleri için bir değerlendirme metriği geliştirdik. Bu metrik, hukuki dilin ve muhakemenin benzersiz taleplerini ele alarak, yanıtları alaka ve bağlamsal uyuma göre değerlendirdi. Bu yaklaşımlar, özetlerin kalitesini değerlendirmek için bağlamsal özellikleri entegre ederek anlamsal derinliği vurguladı. Bu çalışmalar, n-gram tabanlı yöntemlerin ötesine geçerek metin özetleme ve SC değerlendirmesi alanını ilerletirken, bizim çalışmamız çeşitli bağlamlarda anlamsal tutarlılığı değerlendiren SC sistemleri için genelleştirilmiş bir metriğe odaklanarak farklılaşıyor.

Kullanıcı arayüzü testi ve clickstream veri analizinin kesişim noktasında, Uygun ve arkadaşları [181] kullanıcı arayüzü testini iyileştirmek için büyük veri bilimi projelerinde büyük ölçekli grafik veri işlemenin kullanımını araştırdı. Çalışmaları, bu tür verilerin karmaşıklığını ele almak için ölçeklenebilir ve verimli çerçevelere

olan ihtiyacı vurguladı. Paralel olarak, Olmezoğulları ve Aktaş, kullanıcı gezinme davranışını öğrenmeye odaklanan tıklama akışı veri dizileri için bir temsil tekniği olan Pattern2Vec [182]'i önerdi. Ek olarak, ilgili çalışmaları [183] tıklama akışı verileri için yerleştirme tabanlı temsilleri araştırdı ve bu yöntemlerin kullanıcı davranışını anlamak ve tahmin etmek için uygulanabilirliğini gösterdi. Bu çalışmalar kullanıcı etkileşimi verilerinin temsili ve analizine odaklanırken, bizim çalışmamız SC sistemlerinin anlamsal değerlendirmesine yoğunlaşarak farklılık göstermektedir. Davranışsal kalıpları ve yapısal veri gösterimlerini vurgulayan yukarıdaki yöntemlerin aksine, MQA ölçümümüz, bağlamsal ve anlamsal hizalamaları yakalamak için büyük dil modellerinden yararlanır ve SC sistemleri için daha sağlam ve genelleştirilmiş bir değerlendirme çerçevesi sağlar.

6.6 Önerilen Metodoloji

Bu metedoloji bölümünde Şekil 6.1'te gösterildiği gibi metrik oluşturma iş akışı sunulmaktadır. SC sistemlerinin doğruluk ve kalite açısından sürekli olarak geliştirilmesine rağmen, bu sistemlerin değerlendirilmesinde kullanılan ölçütlerin de aynı oranda iyileştirilmesi gerekmektedir.



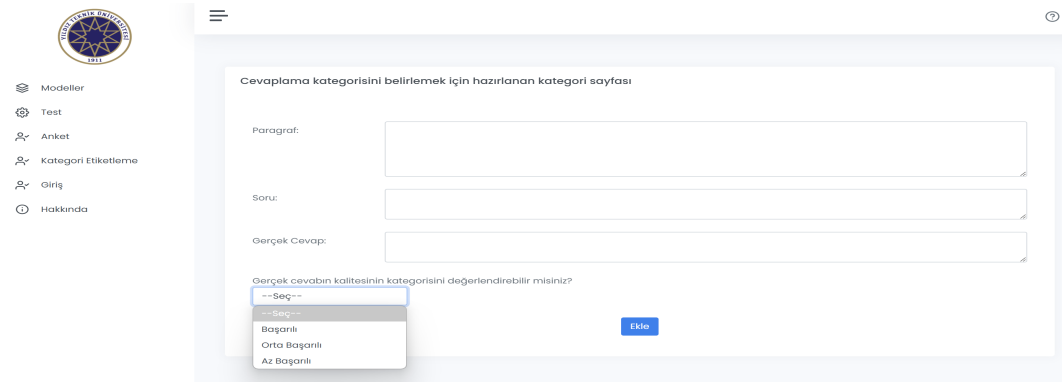
Şekil 6.1 Soru Cevaplama İçin Metrik Oluşturma İş Akışı

Bu bağlamda, model tabanlı bir değerlendirme ölçütü geliştirmek amacıyla çalışmamızda başlangıçta kullanılabilir bir veri seti bulunmamaktaydı. Bu eksikliği gidermek için, beş bağımsız hakeme sekiz adet anket sorusu yöneltilerek insan

yargısına dayalı bir değerlendirme veri seti oluşturulmuştur. Elde edilen bu veri seti, üretken bir model tabanlı değerlendirme metriği oluşturmak üzere temel olarak kullanılmıştır. Bu metodolojinin uygulanması, her biri detaylı olarak açıklanan çeşitli adımları içermektedir. Tüm adımların kodları açık erişim sağlanarak GitHub platformuna yüklenmiştir [184]. Önerilen iş akışının genel yapısı Şekil 6.1'te sunulmaktadır.

Veri Ön İşleme Modülü ve Soru Cevap Tahmin Modülü: Bu çalışmanın girişinde ana hatlarını çizdiğimiz sorunları ele almak için veri setimizi dikkatlice seçtik. SQuAD v2.0 eğitim veri seti kullanılarak BERT tabanlı bir SC modeli eğitilmiş ve ardından SQuAD veri setinin 10.000 örnekten oluşan test bölümü üzerinde modelin tahmin ettiği yanıtlar üretilmiştir. Elde edilen tahminler incelenmiş, BERT'in yanıt üretemediği ya da yanıtların bağlam açısından tutarsız olduğu örnekler veri setinden çıkarılmıştır. Daha sonra, tahmin edilen yanıtların referans (gerçek) yanıtlarla olan benzerliğini değerlendirmek amacıyla BLEU, METEOR ve ROUGE metrikleri uygulanmıştır. Bu metrikler, sözdizimsel benzerlik ya da ayrıntılı yanıt içerikleri açısından kısıtlı performans sergileyebildiğinden, yalnızca 0,5'in altında puan alan örnekler seçilerek daha anlamlı karşılaştırmaların yapılabileceği bir veri alt kümesi oluşturulmuştur. Bu doğrultuda, toplam 600 yüksek kaliteli örnek içeren bir squad-metrik veri seti elde edilmiştir [185]. Ayrıca, çeşitliliği artırmak amacıyla squad-qametrik veri seti [186] kapsamında 0,5'in altında puan almış 500 eğitim örneği ve 0,5'in üzerinde puan almış 100 test örneği seçilmiştir. Test veri setini zenginleştirmek amacıyla, MARCO veri setinden alınan 100 gerçek yanıt, eşanlamlı kelimelerle değiştirilerek marco-qametrik [187] test veri seti oluşturulmuştur.

Kategori Belirleme Modülü: Bu adımın temel amacı, insan yargısına dayalı bir SQuAD v2.0 veri kümesinden oluşan bir veri kümesi oluşturmaktır. Gerçek cevapların incelenmesi sonucunda, bazı örneklerde eksik ya da yetersiz cevapların bulunduğu gözlemlenmiştir. Bu durum, tahmin edilen cevap doğru olsa bile, değerlendirme metriğinin bu iki cevabı karşılaştırırken hatalı sonuç üretmesine yol açabilmektedir. Bu sorunu aşmak adına, squad-qametrik ve marco-qametrik veri kümelerindeki gerçek cevaplar belirli kategorilere ayrılmıştır. Kategorilendirme süreci için özel bir kullanıcı arayüzü tasarlanmıştır [184]. Kullanıcılar, giriş sayfası üzerinden sisteme eriştikten sonra, Şekil 6.2'de gösterilen kategori etiketleme ekranına yönlendirilmiştir. Bu arayüzde, kullanıcılara paragraf, ilgili soru ve gerçek cevap sunularak, söz konusu cevabın kalitesini değerlendirmeleri istenmiştir. Değerlendirme, şu üç kategoriden birinin seçilmesi yoluyla gerçekleştirilmiştir: "1. Kategori: Başarılı", "2. Kategori: Orta Başarılı" ve "3. Kategori: Az Başarılı".



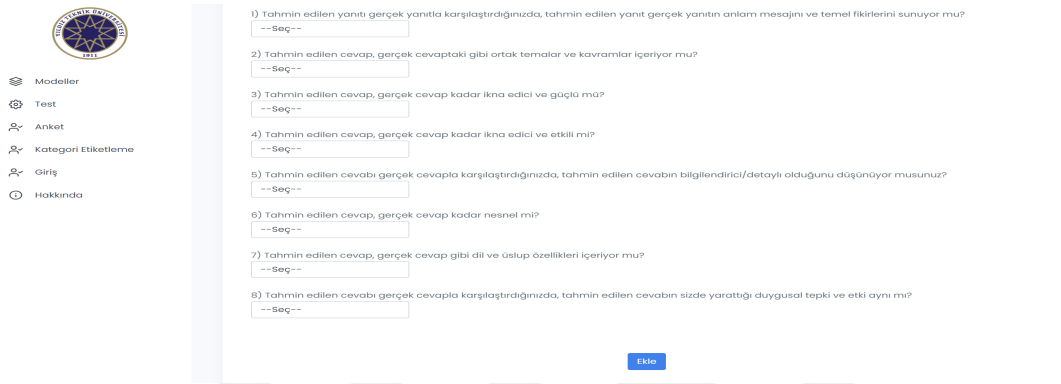
Şekil 6.2 Kategori Etiketleme için Arayüz

Anket Modülü: Veri kümeleri önceki adımda SQuAD MARCO veri kümesinden çıkarıldı ve bu adımda, kullanıcı arayüzünde beş farklı hakeme sekiz anket sorusu sorduk, tahmin edilen cevaba ve gerçek cevaba baktık. Soruları Likert ölçeğinde bir ile beş arasında puanlamalarını istedik ve puanlar şu şekilde etiketlendi; 1: Katılmıyorum, 2: Kararsızım, 3: Az Katılıyorum, 4: Çok Katılıyorum, 5: Kesinlikle Katılıyorum. Her hakemden her soru için sekiz ayrı puan aldık. Beş hakemden her anket sorusu için beş puan aldık; bu beş puanın ortalamasını aldık, böylece her anket sorusu için bir puan elde ettik. Ardından, sekiz anket sorusu için her puanı alıp ve ortalamalarını aldık, böylece her gerçek cevap ve her tahmin edilen cevap için insan puanımızı elde ettik. Sekiz anket sorusu şunlardı:

- Tahmin edilen yanıtı gerçek yanıtla karşılaştırdığınızda, tahmin edilen yanıt gerçek yanıtın anlam mesajını ve temel fikirlerini sunuyor mu?
- Tahmin edilen cevap, gerçek cevaptaki gibi ortak temalar ve kavramlar içeriyor mu?
- Tahmin edilen cevabı gerçek cevapla karşılaştırdığınızda, tahmin edilen cevabın bilgilendirici/detaylı olduğunu düşünüyor musunuz?
- Tahmin edilen cevap, gerçek cevap kadar nesnel mi?
- ahmin edilen cevap, gerçek cevap gibi dil ve üslup özelliklerini içeriyor mu?
- Tahmin edilen cevabı gerçek cevapla karşılaştırdığınızda, tahmin edilen cevabın sizde yarattığı duygusal tepki ve etki aynı mı?

Tahmin edilen cevaplar ve gerçek cevaplar arasında anlam, ortak tema, ikna, inandırıcılık, bilgilendiricilik, nesnellik, ifade ve duygu kavramlarını inceleyerek mantıksal bir puan elde etmeye çalıştık. Anket sorularını cevaplamak için tasarladığımız arayüz Şekil 6.3 ve Şekil 6.4'deki gibidir.

Şekil 6.3 Anket için kullanıcı arayüzü 1



Şekil 6.4 Anket için kullanıcı arayüzü 2

Bu çalışmada, beş hakeme sekiz anket sorarak anketi kaç dakikada tamamladıklarını hesaplamak için toplam 600 SQuAD veri kümesi ve 100 MARCO veri kümesi kullandık. Bu sonuçlar Tablo 6.1'da gösterilmiştir.

Tablo 6.1 Hakem anketi tamamlama süreleri

	1. Hakem	2. Hakem	3. Hakem	4. Hakem	5. Hakem
Süre (dakika)	306	158	253	411	398

Hakemler, Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü'nde lisansüstü eğitimine devam eden öğrenciler arasından seçilmiş olup, seçim sürecinde daha önce doğal dil işleme (NLP) alanında lisansüstü düzeyde ders almış olmalarına özellikle dikkat edilmiştir. Bu seçim kriteri, değerlendiricilerin NLP alanında temel bilgi ve kavrayışa sahip olmalarını sağlayarak, daha tutarlı ve güvenilir değerlendirmeler gerçekleştirmelerine olanak tanımayı amaçlamıştır. Veri Bölme Modülü: Bu adımda, hakemler tarafından değerlendirilen SQuAD veri setini eğitim ve test veri setlerine böldük ve toplam verinin %83'ünü eğitim veri seti için, %17'sini ise test veri seti için ayrıldı. Değerlendirme Metrik Modeli Oluşturma Modülü: Bu adımda, önceki adımda elde edilen eğitim veri seti

üzerinde Mistral-7B büyük dil modelini eğitildi. Eğitimden sonra, gerçek cevap ile tahmin edilen cevap arasındaki doğruluğu tespit etmede yetkin bir değerlendirme modeli MQA-metrik[187] oluşturuldu. Bu modeli herkesin kullanabilmesi için hugface.co'ya yüklenildi. Değerlendirme Metrik Modellerine Göre Benzerlik Puanı Tahminleri Üretme Modülü: Oluşturulan modelden MACRO ve SQuAD test verilerinden 1 ile 5 arasında değerlendirme puanı sonuçları elde edildi. Bu puanları 0 ile 1 arasına düşecek şekilde normalleştirildi. Geleneksel Değerlendirme Metriklerinin Başarısını Değerlendirme Modülü: Bu adımda, F1, BLEU, ROUGE, METEOR ve kosinüs benzerliği geleneksel metriklerini kullanarak test veri setindeki tahmin edilen yanıtlar ile gerçek yanıtlar arasındaki benzerlik ölçüldü. Korelasyon Puanı Hesaplama Modülü'nü Hesaplama: Bu noktada, test verilerinden çıkartılan geleneksel metrik sonuçları ve model tabanlı metrik sonuçları vardı. Bu metriklerden hangisinin kategorilerine göre daha başarılı olduğunun belirlenmesi gerekiyordu. Bu amaçla Kendall ve Spearman korelasyon yöntemleri kullanıldı. 3 kategorinin tümü için korelasyon sonuçları MACRO ve SQuAD test verilerinden çıkartıldı.

6.7 Uygulama ve Değerlendirme

Bu bölüm, teorik planlamaların ötesinde, gerçek dünya uygulamalarının karşılaştığı pratik zorluklara odaklanmaktadır. İlk olarak, çalışmada kullanılan veri kümesi olan Stanford Soru-Cevap Veri Kümesi (SQuAD) 2.0 ele alınmaktadır. Geniş, çeşitli ve karmaşık soru-cevap çiftlerinden oluşan bu veri kümesi, önerilen yöntemin uygulanmasına yönelik stratejilerin temelini oluşturmuştur. Devamında, önerilen değerlendirme metriğinin geçerliliğini ve güvenilirliğini ortaya koymayı amaçlayan test tasarımı detaylandırılmaktadır. Bölümün son kısmında ise, önerilen metriğe ilişkin elde edilen sonuçlar açık ve ölçülebilir bir biçimde sunulmakta; bu bulgular, metriğin etkisi, uygulanabilirliği ve potansiyel iyileştirme alanları üzerine eleştirel bir değerlendirme yapılmasına zemin hazırlamaktadır.

6.7.1 Veri Kümesi

Bu çalışmada, iki farklı veri kümesi oluşturulmuştur. Bu kümeler, ya tahmin edilen yanıtın gerçek yanıtla oldukça düşük düzeyde benzerlik gösterdiği durumları ya da tahmin edilen yanıtın gerçek yanıttan daha ayrıntılı olduğu örnekleri içermektedir. SQuAD v2.0 eğitim veri kümesi kullanılarak, BERT modeli 86.821 örnek ile ince ayar aşamasından geçirilmiş ve bir soru-cevap (SC) modeli oluşturulmuştur. Elde edilen model, 10.000 adet tahmin edilen yanıttan oluşan bir test verisi üzerinde değerlendirilmiştir. Sonrasında, veri kalitesini artırmak ve değerlendirme

sürecini iyileştirmek amacıyla geleneksel metrikler (BLEU, ROUGE, METEOR, F1 skoru ve kosinüs benzerliği) kullanılarak analiz gerçekleştirilmiş ve 0,5'in altında puanlanan örnekler seçilmiştir. Yinelenen, boş veya anlam bütünlüğü taşımayan örnekler ayıklanmış ve nihai olarak toplamda 600 yüksek kaliteli örnek belirlenmiştir: 500 tanesi eğitim, 100 tanesi ise test amacıyla kullanılmıştır. 500 eğitim verisinin 400'ü geleneksel metriklere göre 0,5'in altında, 100'ü ise 0,5'in üzerinde puan almıştır. Benzer şekilde, 100 test örneğinin 80'i 0,5'in altında, 20'si ise üzerinde puanlanmıştır. Ayrıca, MARCO veri kümesinden 100 örnek üretilmiş ve bu örneklerde gerçek yanıtlar eşanlamlılarla değiştirilmiştir. Bu sayede, modelin yalnızca semantik benzerlik temelinde nasıl bir metrik çıktısı ürettiği analiz edilmiştir.

Çalışma süresince, etiketleme arayüzleri Python programlama dili ile geliştirilmiştir. İş akışının ikinci adımında, BERT tabanlı bir soru-cevap tahmin modülü tasarlanmıştır. Önerilen değerlendirme metriği olan MQA-metrik ise Mistral-7B büyük dil modeli temel alınarak yapılandırılmıştır. Model eğitimi ve deneysel süreçler, Google Collaboratory Pro+ ortamında NVIDIA A100 GPU donanımı kullanılarak gerçekleştirilmiştir.

6.7.2 Test Tasarımı

BERT modelinde kullanılan eğitim veri seti, toplam SQuAD veri setinin 81%'i ve test veri seti 19%'uydu. Mistrall-7B modeliyle eğitilen değerlendirme metriği modelinde, SQuAD veri setinin 0,47%'si ve test veri setinin 0,09%'u kullanıldı. Ek olarak, eş anlamlılar için değerlendirme metriği sonuçlarının nasıl olduğunu belirlemek için ayrı bir veri seti olan MARCO'dan yaklaşık 100 veri seçildi ve bunları test verisi olarak kullanıldı.

Önerilen metriği değerlendirirken, ROUGE, METEOR, F1 puanı ve Kosinüs benzerliği gibi yerleşik ölçütler kullanıldı. ROUGE, hatırlama ve metin benzerliğine odaklanarak oluşturulan ve referans metinler arasındaki örtüşmeyi ölçer. METEOR, kelime seçimi ve dizi doğruluğunu vurgular, benzerliği 0 ile 1 arasında puanlar ve daha yüksek puanlar daha iyi eşleşmeleri gösterir. F1 puanı, özellikle dengesiz veri setleri için yararlı olan sınıflandırma doğruluğunu değerlendirmek için kesinlik ve hatırlamayı birleştirir. Kosinüs benzerliği, -1 ile 1 arasında değişen metin vektörleri arasındaki yön benzerliğini ölçer. Bu metrikler, açıklık, kesinlik ve dil özgü zorluklar gibi metin üretiminin çeşitli yönlerini değerlendirir.

Önerilen yaklaşım, geleneksel metriklerin belirli bir sınırlamasını ele alır: n-gram

ve yüzey düzeyinde benzerliğe güvenmeleri, özellikle eş anlamlılar veya parafraz içeren durumlarda genellikle anlamsal anlamı yakalamada başarısız olurlar. Düşük sözdizimi benzerliğine sahip verileri seçmenin amacı, MQA metriğinin, sözdizimsel olarak farklı olsalar bile, tahmin edilen ve referans yanıtlar arasındaki anlamsal yakınlığı değerlendirebilmesini sağlamaktır.

6.7.3 Deneysel Çalışma Tasarımı ve Uygulama Detayları

Değerlendirme sürecinde iki farklı model yapılandırması kullanılmıştır. Bunlardan ilki, referans olarak kullanılan BERT tabanlı soru-cevap modelidir. İkincisi ise tez kapsamında önerilen model tabanlı metriği üretmek için kullanılan Mistral-7B Instruct modelidir. Bu iki modelin çıktıları, klasik metriklerle ve insan yargısıyla kıyaslanarak korelasyon analizi yapılmıştır.

Korelasyon analizinde, sıralı veriler arasındaki ilişkinin gücünü değerlendirmek için Spearman () ve sıralar arası bağıl tutarlılığı ölçmek için Kendall () korelasyon yöntemleri tercih edilmiştir. Bu analizler hem insan yargısı ile klasik metrikler (BLEU, ROUGE, METEOR, F1, CosineSim) hem de insan yargısı ile önerilen model tabanlı skorlar arasında gerçekleştirilmiştir. Böylece, önerilen metriğin insan yargısıyla olan uyumu, geleneksel metriklerle kıyaslamalı olarak ortaya konmuştur. Deneyler, Google Colab Pro+ platformunda, yüksek bellek kapasitesine sahip NVIDIA A100 40GB GPU üzerinde gerçekleştirilmiştir. Kodlama ve süreç yönetimi Python programlama dili ile yapılmış; özellikle HuggingFace Transformers, PEFT ve LoRA gibi açık kaynak kütüphanelerden faydalanılmıştır. Çalışmada kullanılan Mistral-7B modeli, yüksek parametre sayısı ve büyük model boyutu nedeniyle eğitim sürecinde önemli kaynak gereksinimleri doğurmuştur. İlk etapta tam model eğitimi uygulanmaya çalışılmış; ancak bu yaklaşım sırasında GPU bellek sınırlarının aşılması, batch size kısıtlamaları ve eğitim süresinin uzaması gibi çeşitli zorluklar ortaya çıkmıştır. Bu durum, hem modelin stabil bir şekilde çalışmasını engellemiş hem de prototipleme döngüsünü önemli ölçüde yavaşlatmıştır. Bu zorlukları aşmak ve eğitim sürecini daha verimli bir şekilde yürütebilmek amacıyla Parameter-Efficient Fine-Tuning (PEFT) tekniklerinden biri olan LoRA (Low-Rank Adaptation) yöntemi tercih edilmiştir. LoRA, yalnızca düşük-rank matris bileşenlerini eğittiği için hem bellek kullanımını azaltmakta hem de eğitim süresini ciddi ölçüde kısaltmaktadır. Modelin tüm ağırlıklarını güncellemek yerine, yalnızca belirli katmanlara düşük boyutlu adaptasyon parametreleri eklenerek öğrenme işlemi yapılmakta ve bu sayede hem hesaplama maliyeti hem de donanım ihtiyacı düşürülmektedir. Çalışmada HuggingFace PEFT kütüphanesi aracılığıyla Mistral modeli LoRA yöntemiyle yeniden yapılandırılmıştır. Bu sayede, eğitim işlemleri daha düşük

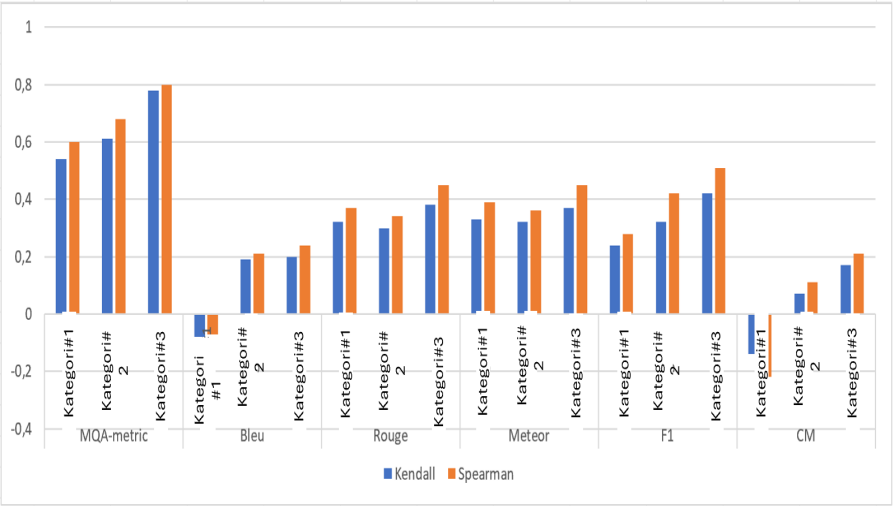
kaynak tüketimiyle başarılı bir şekilde tamamlanabilmiş ve model tabanlı metrik üretim sürecine entegre edilebilmiştir. LoRA'nın uygulanması sayesinde yüksek kapasiteli bir dil modeli olan Mistral, veri kümesine özel olarak uyarlanmış ve klasik metriklerin yetersiz kaldığı durumlarda daha anlamlı skorlar üretebilir hâle gelmiştir.

6.7.4 Sonuçlar

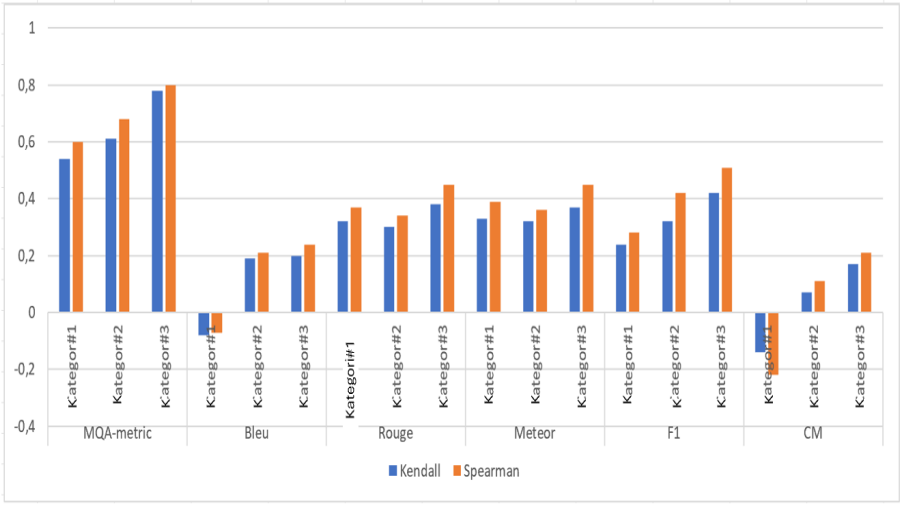
Metriğin birincil amacı, özellikle eş anlamlılar, detaylı veya farklı ifadelerin mevcut olduğu durumlarda semantik doğruluğu değerlendirmektir. Sözdizimi benzerliği yüksek olsa bile, semantik anlamda ince farklılıklar yine de ortaya çıkabilir. MQA metriği, bu ifadeleri yakalamak ve cevaplar arasındaki daha derin semantik ilişkiye odaklanarak SC sistemlerinin daha kapsamlı ve doğru bir şekilde değerlendirilmesini sağlamak için tasarlanmıştır. MQA metriği, BLEU, ROUGE ve METEOR gibi geleneksel metriklerin tamamen yerini almak için değil, onları tamamlamak için tasarlanmıştır. Geleneksel metrikler sözdizimsel benzerliği değerlendirmek için yararlı olsa da özellikle eş anlamlılar veya SC sistemlerine özgü daha uzun üretken cevaplarla uğraşırken, semantik ifadeleri yakalamada genellikle yetersiz kalırlar. MQA metriği, SC çıktılarının adil değerlendirilmesi için çok önemli olan semantik doğruluğun daha sağlam bir değerlendirmesini sağlayarak bu sınırlamaları ele almak üzere tasarlanmıştır.

MQA metriğinin önemi, insan yargısıyla daha yakın bir şekilde uyum sağlayarak daha bağlam farkında bir değerlendirme sunma becerisinde yatmaktadır. Sunulan deneysel sonuçlar, MQA metriğinin geleneksel metriklere kıyasla insan puanlarıyla daha yüksek korelasyona ulaştığını ve böylece SC sistemlerinin gerçek dünya uygulamalarındaki performansını değerlendirmede bir iyileştirme sağladığını göstermektedir. MQA metriği, formdan ziyade anlama odaklanarak değerlendirme sürecini geliştirirken, daha bütünsel bir değerlendirme sağlamak için geleneksel metriklerle birlikte kullanılabilir. Şekil 6.5 ve Şekil 6.6'daki grafiklere baktığımızda, Squad ve Marco veri kümelerinden üç kategorinin hepsinden sonuçlar elde edilmiştir. Model tabanlı değerlendirme metriğimiz MQA metriği, yanıtların geniş anlamlarını sözdizimsel benzerliklerinden değerlendirdiği için METEOR, Kosinüs Benzerliği ve ROUGE gibi geleneksel metriklerden daha iyi performans göstermektedir. N-gram örtüşmesine ve kelime eşleştirmesine dayalı bu metriklerin aksine, yaklaşımımız altta yatan anlama odaklanır ve parafraze edilmiş veya eşanlamlılarla zenginleştirilmiş cevapları doğru şekilde değerlendirir. Bu, özellikle çeşitli ve bağlam açısından zengin veri kümelerinde üstün performansla sonuçlanır, çünkü kelime farklılıklarına rağmen anlamsal olarak eşdeğer cevapları tanır. Sonuç olarak, daha adil, daha iyi performans ve kullanıcı merkezli

değerlendirme sağlayarak gerçek dünya ihtiyaçlarını daha iyi karşılayan gelişmiş SC sistemlerine katkıda bulunur.



Şekil 6.5 SQUAD Veriseti üzerindeki sonuçlar



Şekil 6.6 Marco Veriseti üzerindeki sonuçlar

Özetle, MQA-metrik modeli değerlendirme puanları açısından başarılı sonuçlar üretmiştir. Bu model eşanlamlı cümleler için başarılı sonuçlarını göstermek için üretilen veri setinden bazı örnekler seçildi ve bunları Tablo 6.2 ve Tablo 6.3'de sunuldu. Bu tablolarda, gerçek cevabı eşanlamlı bir kelimeyle değiştirerek güncellenmiş bir cevap elde edildi ve bu tahmin edilen cevap ile metrik sonuçları çıkarıldı. Buna karşılık, tablo geleneksel metriklerdeki eşzamanlı cevapların başarısız olduğunu açıkça göstermektedir.

Ayrıca, tahmin edilen yanıt gerçek yanıttan daha uzun ve ayrıntılı olduğunda MQA-metrik modelinin ve geleneksel metriklerin verilerimizden nasıl metrik

Tablo 6.2 Eş anlamlı örneklerde metrik sonuçları 1

Gerçek Cevap	**a major part**
Güncellenen Cevap	**a significant portion**
Tahmini Cevap	**it is the form that is a major part of the earth's atmosphere (see occurrence)...**
MQA-metrik model	1
Bleu	0,01
Rouge	0,15
Meteor	0,38
F1	0,12
CM	0,68

Tablo 6.3 Eş anlamlı örneklerde metrik sonuçları 2

Gerçek Cevap	**to resolve issue**
Güncellenen Cevap	**fix problem**
Tahmini Cevap	**to resolve issue. answered. in politics and government. if the bill is passed by the senate, both the house and senate bills are returned to the house with a note indicating any changes...**
MQA-metrik model	1
Bleu	0,01
Rouge	0,05
Meteor	0,13
F1	0,02
CM	0,54

ürettiğine dair bazı örnekler seçildi; bunlar Tablo 6.4 ve Tablo 6.5'te gösterilmiştir. Bu tablolarda görülebileceği gibi, tahmin edilen yanıtları doğru yanıt olmasına rağmen, geleneksel metrikler düşük metrikler üretirken, MQA-metrik modelinin ürettiği metrik yüksektir. Sonuç olarak, büyük dil modelleri tarafından üretilen yanıtlar kısa değil, aksine uzun ve ayrıntılı yanıtlardır. Bu durumda, tablolardaki sonuçların da gösterdiği gibi, geleneksel metrikler doğruluğu hesaplamada yetersiz kalmakta ve model tabanlı metriklere ihtiyaç duyulmaktadır.

6.7.5 Tartışma

Bu çalışma, temelde sözdizimi ve n-gram benzerliğine odaklanan ve çoğunlukla anlamsal doğruluğu ve bağlamsal nüansı yakalamada yetersiz kalan geleneksel

Tablo 6.4 Daha ayrıntılı tahmin cevaplarında metrik sonuçları 1

Gerçek Cevap	**January 1979**
Tahmini Cevap	**1979, saudi arms purchases from the us exceeded five times israel ' s. another motive for the large scale purchase of arms from the us by saudi arabia was the failure of the shah during january 1979**
MQA-metrik model	0,75
Bleu	0,01
Rouge	0,1
Meteor	0,34
F1	0,12
CM	0,51

Tablo 6.5 Daha ayrıntılı tahmin cevaplarında metrik sonuçları 2

Gerçek Cevap	**Oxygen therapy**
Tahmini Cevap	**oxygen supplementation is used in medicine. treatment not only increases oxygen levels in the patient's blood**
MQA-metrik model	1
Bleu	0
Rouge	0,1
Meteor	0,13
F1	0,11
CM	0,54

değerlendirme ölçütlerinin sınırlamalarını ele almaktadır. Bu sınırlamaları aşmak amacıyla, yeni bir model tabanlı değerlendirme ölçütü önerilmiş ve insan yargısına dayalı olarak, squad-qametrik ve marco-qametrik veri kümeleri geliştirilmiştir. BLEU, ROUGE ve METEOR gibi geleneksel metriklerle yapılan kapsamlı deneyler ve karşılaştırmalar, önerilen MQA-metrik modelinin insan yargısıyla daha yüksek korelasyon gösterdiğini ve SC sistemlerinin performansını değerlendirmek için daha güvenilir ve etkili bir araç sunduğunu ortaya koymuştur. Ayrıca, bu çalışma, soru-cevap sistemlerinin değerlendirilmesindeki eksiklikleri bütüncül bir şekilde ele almayı amaçlamaktadır. Bu bağlamda, uygun bir değerlendirme veri kümesi üretilmiş ve MQA-metrik ile geleneksel değerlendirme ölçütlerinin karşılaştırıldığı, çok aşamalı bir iş akışı geliştirilmiştir. Elde edilen sonuçlar, MQA-metrik modelinin, sözdizimi ve n-gram benzerliğine dayalı

geleneksel değerlendirme metriklerinden daha başarılı performans sergilediğini göstermektedir. Bu bağlamda, daha karmaşık sistemler için insan yargısına ve model tabanlı değerlendirme yöntemlerine dayalı veri kümelerinin gerekliliği açıkça ortaya çıkmaktadır. MQA-metrik, büyük dil modelleri aracılığıyla semantik anlayışı ve bağlamsal ifadeleri daha derinlemesine yakalayarak, geleneksel metriklerin gözden kaçırdığı önemli ifadeleri ve eşanlamlı varyasyonları başarıyla tespit etmektedir. Bu nedenle, önerilen yaklaşımın, yalnızca sözdizimsel örtüşmeye dayalı değil, semantik doğruluğu vurgulayarak SC değerlendirmeleri için yeni bir standart belirlediği savunulmaktadır. Yapılan grafik analizleri, yanıt kalitesi arttıkça metrik sonuçlarının da iyileştiğini göstermektedir. Yüksek kaliteli gerçek yanıtların, metriklerin başarı oranlarını artırdığı gözlemlenmiştir. Sonuçlar, tahmin edilen yanıtın uzunluğu arttıkça ve geleneksel metriklerin sonuçlarının zayıfladığı durumlarda, MQA-metrik modelinin daha başarılı olduğunu ortaya koymaktadır. Geleneksel yöntemler, sözdizimi benzerliğine odaklandığından, MQA-metrik modelinin eşzamanlı kelimeleri yakalama yeteneği, bu eksiklikleri önemli ölçüde telafi etmektedir. Ayrıca, tahmin edilen yanıtın uzunluğu gerçek yanıttan ne kadar farklılaşırsa, geleneksel metriklerin başarısızlık oranı artmakta, ancak MQA-metrik modelimiz bu durumu başarıyla ele almaktadır, çünkü üretken SC modellerinde üretilen yanıtlar genellikle uzun ve ayrıntılı olmaktadır.

Çalışmada kullanılan veri kümeleri, squad-qametrik ve marco-qametrik, İngilizce SC sistemlerine odaklanmıştır ve bu veri kümeleri, özellikle İngilizce dışı diller veya alan özgü uygulamalarda (örneğin, hukuki veya tıbbi alanlarda) gerçek dünya SC veri kümelerinin tüm spektrumunu yansıtmayabilir. Bu sınırlamayı aşmak amacıyla, gelecekteki çalışmaların, veri kümesini çok dilli ve alana özgü örneklerle zenginleştirerek, MQA metriğinin çeşitli diller ve alanlar genelinde uyarlanabilirliğini ve doğruluğunu artırması hedeflenmektedir. Bir diğer önemli zorluk, MQA metriğinin performansının, özellikle Mistral-7B gibi büyük dil modellerine olan bağımlılığıdır. Metrik, üretken model çıktısına dayalı olduğundan, altta yatan model mimarisindeki herhangi bir değişiklik veya iyileştirme, metriğin doğruluğunu ve tutarlılığını etkileyebilir. Bu bağlamda, gelecekteki çalışmalar, alternatif büyük dil modelleri (BDM'ler) ve ince ayar yaklaşımlarının incelenmesini içerecek, böylece metriğin sağlamlığı artırılacaktır. Ayrıca, model teknikleri entegrasyonu, MQA metriğinin farklı üretken modeller genelinde kararlılığını korumasına olanak sağlayacaktır. MQA metriğinin doğrulaması, SQuAD ve MARCO gibi standart SC veri kümeleri kullanılarak gerçekleştirilmiş, ancak bu veri kümelerinin dinamik ve konuşma tabanlı açık alan SC sistemlerinin karmaşıklıklarını tam olarak yansıtmadığına dikkat edilmiştir.

7
SONUÇ

SC sistemlerinin ürettiği cevapların kaliteli üretilmesi kullanıcıların güvenini kazanma ve SC sistemlerinin geliştirilebilmesi noktasında cevapların doğruluğunun değerlendirilme tekniği çok önemlidir. Bu çalışmada SC sistemleri için kullanılan geleneksel değerlendirme metriklerinin boşluklarına odaklanarak model tabanlı metrik ve veri kümesi önerilmiştir.

AS1: Tez kapsamında oluşturulan veri kümeleri nasıl oluşturulmuştur ve etiketlenme süreci nasıl tasarlanmıştır?

Tez kapsamında oluşturulan veri kümeleri, öncelikle çalışmanın amacı ve hedeflerine uygun olarak tasarlanmıştır. Veri toplama aşamasında, doğal dil işleme alanında yaygın olarak kullanılan metin tabanlı kaynaklar olan Squad ve Marco veri kümelerinden faydalanılmıştır. Bu kaynaklar arasında akademik makaleler, haber metinleri gibi çeşitli veri havuzları yer almıştır. Çeşitliliği sağlamak için, farklı metin türlerinden alınan örneklerin dengeli bir şekilde veri kümelerine dahil edilmesine özen gösterilmiştir. Toplanan ham veriler, ön işleme aşamasına tabi tutulmuştur. Bu aşamada gereksiz karakterler, boşluklar ve tekrarlanan veriler hariç tutulmuştur. Veri kümelerin etiketleme süreci, çalışma kapsamında kullanılacak değerlendirme kriterlerine uygun olarak tasarlanmıştır. İlk aşamada, metinler belirli kategorilere veya hedeflere göre gruplandırılmıştır. Etiketleme kriterlerinin objektif olmasını sağlamak için, bir web arayüzü tasarlanarak etiketleme sürecinde görev alacak hakemlerin erişimine açılmıştır.

Etiketleme sürecinde görev alacak hakem seçimi, dikkatlice planlanlanmıştır. Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü lisansüstü mezun olan ve doğal dil işleme eğitimini almış gönüllü katılımcılar tercih edilmiştir. Katılımcıların, lisansüstü seviyesinde temel NLP derslerini tamamlamış olması bir ön koşul olarak belirlenmiştir. Etiketleme sürecinde katılımcılara detaylı eğitimler verilip ve onların etik kurallara uymaları, tarafsız bir şekilde etiketleme yapmaları sağlanmıştır. Etiketleme tamamlandıktan sonra, sonuçların doğruluğunu

ve tutarlılığını değerlendirmek için kalite kontrol aşaması başlatılmıştır. Bu aşamada, her bir etiketin doğruluğu, farklı etiketleyiciler arasında tutarlılığı ve belirlenen yönergelere uygunluğu incelenmiştir. Belirli bir metin üzerinde birden fazla etiketleyicinin çalışması sağlanarak, etiketler arası uyum ölçüldü. Uyuşmazlık olan durumlar tekrar gözden geçirilip gerekirse uzman bir ekip tarafından yeniden etiketleme yapılmıştır.

AS2: Model tabanlı değerlendirme metriği üretirken iş akışı tasarımı nasıl olmalıdır?

Model tabanlı değerlendirme metriği üretiminde iş akışı tasarımı, sistemin amaçlarını net bir şekilde tanımlama ve bu amaçlara uygun bir süreç oluşturma ile başlamıştır. İlk adım, değerlendirme yapılacak alan veya problem türünün belirlenmesidir. Örneğin, SC sistemi metriğin geliştirilmesi, yanıtların doğruluğunu, bağlamsal uygunluğunu ve kapsamlılığını ölçmeye odaklanmalıdır. İş akışı tasarımı sırasında, metriklerin hangi kriterlere dayanacağı ve bu kriterlerin nasıl ölçüleceği netleştirilmiştir. Bu aşamada literatür taraması yapılarak mevcut metriklerin güçlü ve zayıf yönleri analiz edilir, böylece geliştirilmekte olan metriğin yenilikçi ve etkili olmasını sağlayacak bir temel oluşturulmuştur. İş akışının bir sonraki adımı, metriğin eğitimi ve değerlendirilmesi için uygun veri kümelerinin hazırlanmasıdır. Bu süreçte, veri kaynaklarının seçimi, çeşitliliğin sağlanması ve veri kalitesinin güvence altına alınması büyük önem taşır. Örneğin, SC sistemleri için soruların zorluk seviyelerini, bağlamlarını ve yanıtlarının uzunluklarını içeren dengeli bir veri seti hazırlanmıştır. Ayrıca, metriğin insan yargılarına dayalı değerlendirme yapabilmesi için, insan etiketleyiciler tarafından işaretlenmiş bir alt veri kümesi oluşturulmuştur. Bu, metriğin doğruluk ve güvenilirlik açısından değer kazanmasına olanak tanımıştır. Metriğin tasarımı sırasında, büyük dil modellerinden Mistral-7B modeli tercih edilmiştir. Modelin eğitimi sırasında, belirlenen kriterlere göre oluşturulan veri kümeleriyle eğitim yapılır ve metriğin her bir çıktısı için uygun skorlar belirlenir. Bu aşamada, iş akışının modüler bir yapıya sahip olması önemlidir, böylece farklı veri türlerine veya problemlere uyarlanabilir bir metrik elde edilir.

İş akışının son aşaması, metriğin performansını değerlendirme ve optimize etme sürecidir. Geliştirilen metrik, insan yargılarına dayalı veri kümeleri üzerinde test edilerek doğruluğu ve tutarlılığı analiz edilir. Metriğin, diğer popüler metriklerle (örneğin, ROUGE, BLEU, METEOR) karşılaştırılması, performansını değerlendirmek için kritik bir adımdır.

AS3: Soru cevaplama sistemlerinin sonuçlarının kalitesini değerlendirmek

için oluşturulan MQA-metrik modelinin etkinliğini ve başarısını nasıl değerlendirebiliriz?

MQA-metrik modelinin etkinliğini değerlendirmek için modelin verdiği puanları insan yargılarıyla karşılaştırılmıştır. Gönüllü olarak seçilen ve NLP alanında eğitim almış kişiler tarafından etiketlenen veri kümeleri, MQA-metriğin insan değerlendirmelerine ne kadar uyumlu olduğunu analiz etmek için kullanılabilir. Modelin çıktılarının insan yargılarıyla yüksek korelasyon göstermesi, metriğin doğruluğunu ve güvenilirliğini göstermiştir. Özellikle, semantik benzerlik, bağlama uygunluk ve detaylı cevaplar gibi geleneksel metriklerdeki boşluğu bu metrik çalışmasında başarısını görmekteyiz. MQA-metrik modelinin etkinliğini değerlendirmek için ROUGE, METEOR, BLEU ve F1 gibi yaygın olarak kullanılan metriklerle kıyaslama yapılmıştır. Bu tür bir karşılaştırma, MQA-metriğin hangi açılardan mevcut metriklerden üstün olduğunu veya hangi durumlarda eksik kaldığını ortaya koymaktadır. Örneğin, ROUGE genellikle kelime düzeyindeki benzerliklere odaklanırken, MQA-metriğin semantik anlamı daha derinlemesine analiz edebilmesi beklenir. Böyle bir kıyaslama, metriğin yalnızca doğruluğunu değil, aynı zamanda kapsamlılığını ve bağlamsal anlayışını da değerlendirmek için fırsat sunmaktadır.

Metriğin performansını çeşitli soru türleri ve veri kümeleri üzerinde test etmektir. MQA-metriğin, farklı zorluk seviyelerinde ve uzunluklarda sorularla uyumlu bir şekilde çalışıp çalışmadığı incelenmek için Marco veri kümesinden oluşturulan test veri kümesinden sonuçlar üretilip incelenmiş ve başarılı sonuçlar üretilmiştir. Ve böylece metriğin genelleştirilebilirliğini ve farklı bağlamlarda ne kadar tutarlı çalıştığını göstermiştir. MQA-metriğin kısa yanıtlarla uzun yanıtları adil bir şekilde değerlendirebilmesi önemlidir. Yanıt uzunluğuna bağlı olarak metriğin puanlamasında bir sapma olup olmadığı test edilmiştir. Benzer şekilde, bağlamı doğru anlayıp anlamadığı da analiz edilmiştir. Metriğin yalnızca yüzeysel değerlendirme yapmadığını, aynı zamanda semantik anlamı ve bağlam ilişkisini doğru bir şekilde yorumlayabildiğini göstermektedir

7.0.1 Tez Kapsamında Yapılan Araştırma Sonuçlarının Geçerliliğini Tehdit Eden Unsurlar

Bu tez kapsamında geliştirilen model tabanlı değerlendirme metriği, SQuAD temel alınarak yapılandırılmış ve Mistral-7B Instruct adlı büyük dil modeli kullanılarak eğitilmiştir. Elde edilen bulgular, önerilen metodolojinin SQuAD veri kümesi üzerindeki performansına dayanmaktadır. Bu durum, genel geçerliliğin veri kümesine özgü özellikler nedeniyle sınırlı olabileceğini göstermektedir. Ancak

önerilen iş akışının, içerik bakımından benzer yapıdaki farklı QA (Soru-Cevap) veri setlerine de aynı şekilde uygulanabilir olduğu düşünülmektedir. Bu bağlamda, metodolojinin genellenebilirliğini sağlamak amacıyla, ileride farklı soru-cevap veri kümeleri kullanılarak tekrar test edilmesi gerekmektedir.

Model tabanlı değerlendirme metriği, yalnızca Mistral BDM üzerinde geliştirilmiştir. Ancak, bu metrik oluşturma sürecinin yapısı model bağımsız olarak tasarlandığı için; aynı iş akışı, farklı büyük dil modelleri kullanılarak da tekrar uygulanabilir ve böylece model tabanlı metriklerin alternatif versiyonları üretilebilir. Bu durum, önerilen yaklaşımın esnekliğini ve sürdürülebilirliğini göstermektedir. Ayrıca, önerilen değerlendirme metriği, belirli bir SC sisteminden (örneğin BERT tabanlı) alınan cevaplar üzerine uygulanmış ve değerlendirme sonuçları buna göre analiz edilmiştir. Bu yaklaşım, kullanılan SC sisteminin kalitesine ve modelin ince ayar düzeyine duyarlıdır. Ancak önerilen metrik oluşturma süreci, farklı SC modelleri için de aynı şekilde uygulanabilir. Yani, başka büyük dil modellerine dayalı SC sistemleri (örneğin T5, RoBERTa, DeBERTa) de benzer şekilde incelenerek, ilgili sistemler için ayrı model tabanlı metrikler geliştirilebilir.

Bu bağlamda, çalışmada sunulan model tabanlı metrik üretim süreci; yalnızca belirli bir veri seti ve BDM ile sınırlı kalmamakta, aynı zamanda alternatif veri kümeleri ve dil modelleri ile de ölçeklenebilir bir yapıya sahiptir. Bu yönüyle, metodolojinin tekrarlanabilirliği yüksektir; ancak her yeni uygulamada, modelin ve veri kümesinin kendine özgü özellikleri dikkate alınarak yeniden eğitilmesi ve değerlendirilmesi gerekmektedir. Bu da çalışmanın geçerliliği açısından hem bir esneklik hem de potansiyel bir tehdit unsuru olarak değerlendirilebilir. Sonuç olarak, önerilen yaklaşımın geçerliliği, kullanılan veri kümesi, model mimarisi ve SC sistemine bağlı olmakla birlikte; önerilen metodoloji farklı veri kümeleri ve modeller üzerinde de uygulanabilir nitelikte olup, yeniden üretilebilirlik ve genellenebilirlik açısından önemli bir potansiyele sahiptir.

7.0.2 Gelecekteki Araştırma Fırsatları

SC sistemlerinin kullanım alanlarının genişlemesiyle birlikte, bu sistemlerin performansının güvenilir ve anlamlı bir biçimde değerlendirilmesine olan ihtiyaç da giderek artmaktadır. Özellikle kullanıcı memnuniyetinin ön planda olduğu web arama motorları ve sohbet robotları gibi alanlarda, değerlendirme süreçlerinin doğruluğu kritik öneme sahiptir. Geleneksel değerlendirme metrikleri genellikle yüzeysel benzerliklere odaklandığı için yazım hatalarına, eşanlamlı kelimelere ve ayrıntılı cevaplara karşı yetersiz kalmakta, bu durumda gelişmiş değerlendirme

yöntemlerine olan ihtiyacı ortaya koymaktadır. Bu bağlamda, bu boşlukları gidermek üzere model tabanlı bir metrik önerilmiş ve bu metriği desteklemek amacıyla özel olarak tasarlanmış bir veri kümesi geliştirilerek sistematik bir biçimde uygulanmıştır. Önerilen yaklaşımın geçerliliğini ve işlevselliğini göstermek adına kapsamlı bir kullanılabilirlik analizi gerçekleştirilmiştir. Bunun yanı sıra, SC sistemleri üzerine gerçekleştirilen önceki çalışmaları incelemek ve gelecekte yapılacak araştırmalara yön göstermek amacıyla kapsamlı bir sistematik literatür taraması yapılmıştır. Bu analiz sonucunda, büyük dil modellerinin değerlendirme metriklerine entegrasyonunun henüz yeterince araştırılmadığı ve bu konudaki çalışmaların artırılması gerektiği sonucuna ulaşılmıştır.

Büyük dil modelleri sürekli evrim geçiren yapısı nedeniyle, bu sistemlerin değerlendirme süreçlerine entegrasyonu düzenli olarak güncellenmeli ve sürdürülebilir biçimde ele alınmalıdır. Ayrıca, bu tez kapsamında kullanılan ve geliştirilen veri kümesi, çalışmanın en temel bileşenlerinden biri olarak öne çıkmaktadır. Literatürde değerlendirici metriklere yönelik kamuya açık veri kümesi sayısının oldukça sınırlı olduğu göz önünde bulundurulduğunda, bu tez kapsamında oluşturulan veri kümesinin literatüre önemli bir katkı sunduğu ve bu alandaki veri kümelerinin sayısının artırılmasına yönelik çalışmalara ihtiyaç duyulmaktadır. Bu tezde geliştirilen model tabanlı değerlendirme metriği, insan yargılarına dayalı olarak oluşturulan özel bir veri setiyle desteklenerek, değerlendirme süreçlerindeki yüzeyselliği azaltmayı ve semantik doğruluğu artırmayı hedeflemektedir. Çalışmada, mevcut geleneksel metriklerin (ROUGE, METEOR, F1 skoru ve Cosine benzerliği) sınırlılıkları ele alınmış, önerilen metriğin bu metriklerle karşılaştırmalı analizleri gerçekleştirilmiştir. Önerilen metrik, özellikle yanıtların bağlama uygunluğunu ve semantik anlamını daha doğru bir şekilde ölçme kapasitesiyle öne çıkmaktadır. Ayrıca, hakemlerin anket süreci, değerlendirme aşamaları ve önyargıların minimize edilmesine yönelik kullanılan yöntemler ayrıntılı bir şekilde sunulmuştur. Son olarak, çalışma yalnızca metriğin güçlü yönlerini değil, aynı zamanda sınırlılıklarını ve bu sınırlıkların giderilmesine yönelik önerileri de kapsamlı biçimde ele alarak, SC sistemlerinin daha sağlıklı bir biçimde değerlendirilmesine katkı sunmaktadır.

KAYNAKÇA

[1] H. Lee ve diğ., *KPQA: A Metric for Generative Question Answering Using Keyphrase Weights*, 2021. arXiv: 2005.00192 [cs.CL]. erişim adresi: https://arxiv.org/abs/2005.00192.

[2] E. Durmus, H. He M. Diab, "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.454. erişim adresi: http://dx.doi.org/10.18653/v1/2020.acl-main.454.

[3] A. Mohammadshahi ve diğ., *RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question*, 2023. arXiv: 2211.01482 [cs.CL]. erişim adresi: https://arxiv.org/abs/2211.01482.

[4] J. P. Chirag Shah, "Evaluating and predicting answer quality in community QA," *SIGIR '10: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010, ss. 411–418.

[5] D. R. Daniel Deutsch Tania Bedrax-Weiss, "Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary," *Transactions of the Association for Computational Linguistics*, c. 9, ss. 774–789, 2021.

[6] V. G. Jaspreet Kaur, "Effective Question Answering Techniques and their Evaluation Metrics," *International Journal of Computer Applications*, c. 65, ss. 30–37, 2013.

[7] J. Risch, T. Möller, J. Gutsch M. Pietsch, *Semantic Answer Similarity for Evaluating Question Answering Models*, 2021. arXiv: 2108.06130 [cs.CL]. erişim adresi: https://arxiv.org/abs/2108.06130.

[8] D. Z. W. Yang Bai, "More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering," 2021.

[9] D. Bakir M. S. Aktas, "A Systematic Literature Review of Question Answering: Research Trends, Datasets, Methods," *Computational Science and Its Applications – ICCSA 2022 Workshops*, Springer International Publishing, 2022, ss. 47–62.

[10] D. Bakir M. S. Aktas, "A Business Workflow For Providing Open-Domain Question Answering Reader Systems on The Wikipedia Dataset," *2022 IEEE International Conference on Big Data (Big Data)*, 2022, ss. 3308–3314. doi: 10.1109/BigData55660.2022.10020945.

[11] D. Bakir M. S. Aktas, “An Approach to Decentralized Hybrid Question Answering Systems,” *Computational Science and Its Applications – ICCSA 2023 Workshops*, Springer Nature Switzerland, 2023, ss. 29–38, isbn: 978-3-031-37129-5.

[12] D. Bakir, M. S. Aktas B. Yildiz, “A Model-Based Evaluation Metric for Question Answering Systems,” *International Journal of Software Engineering and Knowledge Engineering*, 2024. doi: 10.1142/S0218194025500032.

[13] K. Papineni, S. Roukos, T. Ward W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, seri ACL ’02, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, ss. 311–318. doi: 10.3115/1073083.1073135. erişim adresi: https://doi.org/10.3115/1073083.1073135.

[14] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Tem. 2004, ss. 74–81. erişim adresi: https://aclanthology.org/W04-1013/.

[15] S. Banerjee A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan: Association for Computational Linguistics, Haz. 2005, ss. 65–72. erişim adresi: https://aclanthology.org/W05-0909/.

[16] D. Bakir, B. Yildiz M. S. Aktas, “ Developing and Evaluating a Model-Based Metric for Legal Question Answering Systems,” *2023 IEEE International Conference on Big Data (BigData)*, Los Alamitos, CA, USA: IEEE Computer Society, Ara. 2023, ss. 2745–2754. doi: 10.1109/BigData59044.2023.10386689. erişim adresi: https://doi.ieeecomputersociety.org/10.1109/BigData59044.2023.10386689.

[17] F. Al-Khateeb, N. Dey, D. Soboleva J. Hestness, *Position Interpolation Improves ALiBi Extrapolation*, 2023. arXiv: 2310.13017 [cs.CL]. erişim adresi: https://arxiv.org/abs/2310.13017.

[18] C. Wohlin, E. Mendes, K. R. Felizardo M. Kalinowski, “Guidelines for the search strategy to update systematic literature reviews in software engineering,” *Information and Software Technology*, c. 127, s. 106 366, 2020, issn: 0950-5849. doi: https://doi.org/10.1016/j.infsof.2020.106366. erişim adresi: https://www.sciencedirect.com/science/article/pii/S095058491930223X.

[19] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” *Information and Software Technology*, c. 51, no. 1, ss. 7–15, 2009, Special Section - Most Cited Articles in 2002 and Regular Research Papers, issn: 0950-5849. doi: https://doi.org/

10.1016/j.infsof.2008.09.009. erişim adresi: https://www.sciencedirect.com/science/article/pii/S0950584908001390.

[20] M. Unterkalmsteiner, T. Gorschek, A. M. Islam, C. K. Cheng, R. B. Permadi R. Feldt, "Evaluation and Measurement of Software Process Improvement—A Systematic Literature Review," *IEEE Transactions on Software Engineering*, c. 38, no. 2, ss. 398–424, 2012. doi: 10.1109/TSE.2011.26.

[21] X. Yao, "Feature-driven Question Answering With Natural Language Alignment," 2014. erişim adresi: https://api.semanticscholar.org/CorpusID:63858843.

[22] C. Sammut G. I. Webb, *Encyclopedia of Machine Learning*, 1st. Springer Publishing Company, Incorporated, 2011, isbn: 0387307680.

[23] E. M. Voorhees, "Question answering in TREC," *Proceedings of the Tenth International Conference on Information and Knowledge Management*, seri CIKM '01, Atlanta, Georgia, USA: Association for Computing Machinery, 2001, ss. 535–537, isbn: 1581134363. doi: 10.1145/502585.502679. erişim adresi: https://doi.org/10.1145/502585.502679.

[24] J. Herrera, A. Peñas F. Verdejo, "Question Answering Pilot Task at CLEF 2004," *Multilingual Information Access for Text, Speech and Images*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, ss. 581–590, isbn: 978-3-540-32051-7.

[25] A. Nentidis ve diğ., "Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering," içinde *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer Nature Switzerland, 2023, ss. 227–250, isbn: 9783031424489. doi: 10.1007/978-3-031-42448-9_19. erişim adresi: http://dx.doi.org/10.1007/978-3-031-42448-9_19.

[26] G.-I. Brokos, P. Malakasiotis I. Androutsopoulos, "Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering," *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, Berlin, Germany: Association for Computational Linguistics, Ağus. 2016, ss. 114–118. doi: 10.18653/v1/W16-2915. erişim adresi: https://aclanthology.org/W16-2915/.

[27] Y. Cao ve diğ., "AskHERMES: An online question answering system for complex clinical questions," *Journal of biomedical informatics*, c. 44 2, ss. 277–88, 2011. erişim adresi: https://api.semanticscholar.org/CorpusID:15241604.

[28] S. Tellex, B. Katz, J. Lin, A. Fernandes G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," seri SIGIR '03, Toronto, Canada: Association for Computing Machinery, 2003, ss. 41–47, isbn: 1581136463. doi: 10.1145/860435.860445. erişim adresi: https://doi.org/10.1145/860435.860445.

[29] M. W. Bilotti, J. Elsas, J. Carbonell E. Nyberg, "Rank learning for factoid question answering with linguistic and semantic constraints," *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, seri CIKM '10, Toronto, ON, Canada: Association for Computing Machinery, 2010, ss. 459–468, isbn: 9781450300995. doi: 10.1145/1871437.1871498. erişim adresi: https://doi.org/10.1145/1871437.1871498.

[30] M. Pardiño ve diğ., "Adapting IBQAS to work with Text Transciptions in QAst Task: IBQAst," *Conference and Labs of the Evaluation Forum*, 2008. erişim adresi: https://api.semanticscholar.org/CorpusID:7573984.

[31] B. Roth, C. Conforti, N. Poerner, S. K. Karn H. Schütze, "Neural architectures for open-type relation argument extraction," *Natural Language Engineering*, c. 25, no. 2, ss. 219–238, Ara. 2018, issn: 1469-8110. doi: 10.1017/s1351324918000451. erişim adresi: http://dx.doi.org/10.1017/S1351324918000451.

[32] G. H. Yun Niu, *Identifying cores of semantic classes in unstructured text with a semi-supervised learning approach*, 2019. erişim adresi: https://www.cs.toronto.edu/pub/gh/Niu+Hirst-RANLP-2007.pdf.

[33] Y. Chen, X. Zhang, A. Chen, X. Zhao Y. Dong, "QA system for food safety events based on information extraction. Nongye Jixie Xuebao," *Trans. Chin. Soc. Agric. Mach*, c. 51, ss. 442–448, 2020.

[34] D. Pappas I. Androutsopoulos, "A Neural Model for Joint Document and Snippet Ranking in Question Answering for Large Document Collections," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Ağus. 2021, ss. 3896–3907. doi: 10.18653/v1/2021.acl-long.301. erişim adresi: https://aclanthology.org/2021.acl-long.301/.

[35] H.-Y. Lin, T.-H. Lo B. Chen, "Enhanced Bert-Based Ranking Models for Spoken Document Retrieval," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, ss. 601–606, 2019. erişim adresi: https://api.semanticscholar.org/CorpusID:211243670.

[36] P. Nie, Y. Zhang, A. Ramamurthy L. Song, *Answering Any-hop Open-domain Questions with Iterative Document Reranking*, 2021. arXiv: 2009.07465 [cs.CL]. erişim adresi: https://arxiv.org/abs/2009.07465.

[37] B. Kratzwald S. Feuerriegel, "Adaptive Document Retrieval for Deep Question Answering," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Ekim 2018, ss. 576–581. doi: 10.18653/v1/D18-1055. erişim adresi: https://aclanthology.org/D18-1055/.

[38] Y. Cong, Y. Wu, X. Liang, J. Pei Z. Qin, "PH-model: enhancing multi-passage machine reading comprehension with passage reranking and hierarchical information," c. 51, no. 8, ss. 5440–5452, Ağus. 2021, issn: 0924-669X. doi: 10.1007/s10489-020-02168-3. erişim adresi: https://doi.org/10.1007/s10489-020-02168-3.

[39] T. M. Nguyen, V.-L. Tran, D.-C. Can, Q.-T. Ha, L. T. Vu E.-S. Chng, "QASA: Advanced Document Retriever for Open-Domain Question Answering by Learning to Rank Question-Aware Self-Attentive Document Representations," seri ICMLSC '19, Da Lat, Viet Nam: Association for Computing Machinery, 2019, ss. 221–225, isbn: 9781450366120. doi: 10.1145/3310986.3310999. erişim adresi: https://doi.org/10.1145/3310986.3310999.

[40] Q.-l. Guo M. Zhang, "Semantic information integration and question answering based on pervasive agent ontology," *Expert Systems with Applications*, c. 36, no. 6, ss. 10 068–10 077, 2009, issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2009.01.056. erişim adresi: https://www.sciencedirect.com/science/article/pii/S0957417409000542.

[41] B. Grau, "Finding an answer to a question," seri IWRIDL '06, Kolkata, India: Association for Computing Machinery, 2006, isbn: 1595936084. doi: 10.1145/1364742.1364751. erişim adresi: https://doi.org/10.1145/1364742.1364751.

[42] D. Radev, W. Fan, H. Qi, H. Wu A. Grewal, "Probabilistic question answering on the web," *Proceedings of the 11th International Conference on World Wide Web*, seri WWW '02, Honolulu, Hawaii, USA: Association for Computing Machinery, 2002, ss. 408–419, isbn: 1581134495. doi: 10.1145/511446.511500. erişim adresi: https://doi.org/10.1145/511446.511500.

[43] J. Lin ve diğ., "The role of context in question answering systems," *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, seri CHI EA '03, Ft. Lauderdale, Florida, USA: Association for Computing Machinery, 2003, ss. 1006–1007, isbn: 1581136374. doi: 10.1145/765891.766119. erişim adresi: https://doi.org/10.1145/765891.766119.

[44] M. Pérez-Coutiño, T. Solorio, M. Montes-y-Gómez, A. López-López L. Villaseñor-Pineda, "Question Answering for Spanish Based on Lexical and Context Annotation," *Advances in Artificial Intelligence – IBERAMIA 2004*, C. Lemaître, C. A. Reyes J. A. González, ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ss. 325–333, isbn: 978-3-540-30498-2.

[45] D. Pappas I. Androutsopoulos, "A Neural Model for Joint Document and Snippet Ranking in Question Answering for Large Document Collections," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online: Association for Computational Linguistics, Ağus. 2021, ss. 3896–3907. doi: 10.

18653/v1/2021.acl-long.301. erişim adresi: https://aclanthology.org/2021.acl-long.301/.

[46] Y. Fan, J. Guo, X. Ma, R. Zhang, Y. Lan X. Cheng, "A Linguistic Study on Relevance Modeling in Information Retrieval," *Proceedings of the Web Conference 2021*, seri WWW '21, ACM, Nis. 2021, ss. 1053–1064. doi: 10.1145/3442381.3450009. erişim adresi: http://dx.doi.org/10.1145/3442381.3450009.

[47] M. Kaiser, "Incorporating User Feedback in Conversational Question Answering over Heterogeneous Web Sources," seri SIGIR '20, Virtual Event, China: Association for Computing Machinery, 2020, s. 2482, isbn: 9781450380164. doi: 10.1145/3397271.3401454. erişim adresi: https://doi.org/10.1145/3397271.3401454.

[48] A. Lamurias, D. Sousa F. M. Couto, "Generating Biomedical Question Answering Corpora From QA Forums," *IEEE Access*, c. 8, ss. 161 042–161 051, 2020. doi: 10.1109/ACCESS.2020.3020868.

[49] M. Sarrouti S. Ouatik El Alaoui, "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions," *Artificial Intelligence in Medicine*, c. 102, s. 101 767, 2020, issn: 0933-3657. doi: https://doi.org/10.1016/j.artmed.2019.101767. erişim adresi: https://www.sciencedirect.com/science/article/pii/S0933365718302756.

[50] A. A. Shah, S. D. Ravana, S. Hamid M. A. Ismail, "Accuracy evaluation of methods and techniques in Web-based question answering systems: a survey," c. 58, no. 3, ss. 611–650, Mar. 2019, issn: 0219-1377. doi: 10.1007/s10115-018-1203-0. erişim adresi: https://doi.org/10.1007/s10115-018-1203-0.

[51] B. Roth, C. Conforti, N. Poerner, S. K. Karn H. Schütze, "Neural architectures for open-type relation argument extraction," *Natural Language Engineering*, c. 25, no. 2, ss. 219–238, Ara. 2018, issn: 1469-8110. doi: 10.1017/s1351324918000451. erişim adresi: http://dx.doi.org/10.1017/S1351324918000451.

[52] *IRQA '08: Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, Manchester, UK: Association for Computational Linguistics, 2008, isbn: 9781905593552.

[53] V. Novotný P. Sojka, "Weighting of Passages in Question Answering," *RASLAN*, 2018. erişim adresi: https://api.semanticscholar.org/CorpusID:67769713.

[54] M. Sarrouti S. Ouatik El Alaoui, "A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering," *Journal of Biomedical Informatics*, c. 68, ss. 96–103, 2017, issn: 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.03.001. erişim adresi: https://www.sciencedirect.com/science/article/pii/S1532046417300503.

[55] H. Bolat B. Şen, "Document Retrieval System for Biomedical Question Answering," *Applied Sciences*, c. 14, no. 6, s. 2613, 2024.

[56] S. A. Aroussi, E. H. Nfaoui O. E. Beqqali, "Improving question answering systems by using the explicit semantic analysis method," *2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, ss. 1–6, 2016. erişim adresi: https://api.semanticscholar.org/CorpusID:18144232.

[57] A. Omari, D. Carmel, O. Rokhlenko I. Szpektor, "Novelty based Ranking of Human Answers for Community Questions," *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, seri SIGIR '16, Pisa, Italy: Association for Computing Machinery, 2016, ss. 215–224, isbn: 9781450340694. doi: 10.1145/2911451.2911506. erişim adresi: https://doi.org/10.1145/2911451.2911506.

[58] M. M. Hoque P. Quaresma, "An effective approach for relevant paragraph retrieval in Question Answering systems," *2015 18th International Conference on Computer and Information Technology (ICCIT)*, 2015, ss. 44–49.

[59] G.-I. Brokos, P. Malakasiotis I. Androutsopoulos, "Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering," *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, K. B. Cohen, D. Demner-Fushman, S. Ananiadou J.-i. Tsujii, ed., Berlin, Germany: Association for Computational Linguistics, Ağus. 2016, ss. 114–118. doi: 10.18653/v1/W16-2915. erişim adresi: https://aclanthology.org/W16-2915/.

[60] A. Nentidis ve diğ., "Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering," içinde *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer Nature Switzerland, 2023, ss. 227–250, isbn: 9783031424489. doi: 10.1007/978-3-031-42448-9_19. erişim adresi: http://dx.doi.org/10.1007/978-3-031-42448-9_19.

[61] M. Neves, "HPI Question Answering System in the BioASQ 2015 Challenge," *Conference and Labs of the Evaluation Forum*, 2015. erişim adresi: https://api.semanticscholar.org/CorpusID:17887641.

[62] Z. Liu, X. Wang, Q. Chen, Y. Zhang Y. Xiang, "A Chinese question answering system based on web search," *2014 International Conference on Machine Learning and Cybernetics*, c. 2, ss. 816–820, 2014. erişim adresi: https://api.semanticscholar.org/CorpusID:534196.

[63] A. M. Ferguson, D. McLean E. F. Risko, "Answers at your fingertips: Access to the Internet influences willingness to answer questions," *Consciousness and Cognition*, c. 37, ss. 91–102, 2015, issn: 1053-8100. doi: https://doi.org/10.1016/j.concog.2015.08.008. erişim adresi: https://www.sciencedirect.com/science/article/pii/S1053810015300234.

[64] W. Sun, C. Fu Q. Xiao, "A text inference based answer extraction for Chinese question answering," *9th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2012, 29-31 May 2012, Chongqing, China*, IEEE, 2012, ss. 2870–2874. doi: 10.1109/FSKD.2012.6234145. erişim adresi: https://doi.org/10.1109/FSKD.2012.6234145.

[65] Z. Rasool ve diğ., "Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset," *Natural Language Processing Journal*, c. 8, s. 100083, 2024, issn: 2949-7191. doi: https://doi.org/10.1016/j.nlp.2024.100083. erişim adresi: https://www.sciencedirect.com/science/article/pii/S2949719124000311.

[66] C. Monz, "Machine learning for query formulation in question answering," *Nat. Lang. Eng.*, c. 17, no. 4, ss. 425–454, Ekim 2011, issn: 1351-3249. doi: 10.1017/S1351324910000276. erişim adresi: https://doi.org/10.1017/S1351324910000276.

[67] Z. Wei, D. Liguo C. Junjie, "Reasoning and realization based on ontology model and Jena," *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, ss. 1057–1060, 2010. erişim adresi: https://api.semanticscholar.org/CorpusID:207872837.

[68] W.-H. Lu, C.-M. Tung C.-W. Lin, "Question Intention Analysis and Entropy-Based Paragraph Extraction for Medical Question Answering," *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore*, C. T. Lim J. C. H. Goh, ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ss. 1582–1586, isbn: 978-3-642-14515-5.

[69] X. Li E. Chen, "Graph-Based Answer Passage Ranking for Question Answering," *2010 International Conference on Computational Intelligence and Security*, 2010, ss. 634–638. doi: 10.1109/CIS.2010.144.

[70] W.-H. Lu, C.-M. Tung C.-W. Lin, "Question intention analysis and entropy-based paragraph extraction for medical question answering," *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore: In Conjunction with 14th International Conference on Biomedical Engineering (ICBME) and 5th Asia Pacific Conference on Biomechanics (APBiomech)*, Springer, 2010, ss. 1582–1586.

[71] D. T. Nguyen, T. N. Pham Q. T. Phan, "A Semantic Model for Building the Vietnamese Language Query Processing Framework in e-Library Searching Application, accepted paper," *The 2nd International Conference on Machine Learning and Computing (ICMLC 2010)*, ss. 9–11.

[72] W.-H. Lu, C.-M. Tung C.-W. Lin, "Question Intention Analysis and Entropy-Based Paragraph Extraction for Medical Question Answering," *6th World Congress of Biomechanics (WCB 2010). August 1-6, 2010 Singapore*, C. T. Lim J. C. H. Goh, ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ss. 1582–1586, isbn: 978-3-642-14515-5.

[73] D. Buscaldi, P. Rosso, J. M. Gómez-Soriano E. Sanchis, "Answering questions with an n-gram based passage retrieval engine," *J. Intell. Inf. Syst.*, c. 34, no. 2, ss. 113–134, Nis. 2010, issn: 0925-9902. doi: 10.1007/s10844-009-0082-y. erişim adresi: https://doi.org/10.1007/s10844-009-0082-y.

[74] S. Momtazi D. Klakow, "A word clustering approach for language model-based sentence retrieval in question answering systems," *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, seri CIKM '09, Hong Kong, China: Association for Computing Machinery, 2009, ss. 1911–1914, isbn: 9781605585123. doi: 10.1145/1645953.1646263. erişim adresi: https://doi.org/10.1145/1645953.1646263.

[75] N. T. Dang D. T. T. Tuyen, "Document Retrieval Based on Question Answering System," *2009 Second International Conference on Information and Computing Science*, c. 1, 2009, ss. 183–186. doi: 10.1109/ICIC.2009.53.

[76] Q.-l. Guo M. Zhang, "Semantic information integration and question answering based on pervasive agent ontology," *Expert Syst. Appl.*, c. 36, no. 6, ss. 10 068–10 077, Ağus. 2009, issn: 0957-4174. doi: 10.1016/j.eswa.2009.01.056. erişim adresi: https://doi.org/10.1016/j.eswa.2009.01.056.

[77] N. T. Dang D. T. T. Tuyen, "Document Retrieval Based on Question Answering System," *2009 Second International Conference on Information and Computing Science*, c. 1, 2009, ss. 183–186. doi: 10.1109/ICIC.2009.53.

[78] N. T. Dang, "e-Document Retrieval by Question Answering System," 2009. erişim adresi: https://api.semanticscholar.org/CorpusID:59808496.

[79] L. Abouenour, K. Bouzoubaa P. Rosso, "Structure-Based Evaluation of an Arabic Semantic Query Expansion Using the JIRS Passage Retrieval System," *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, M. Rosner S. Wintner, ed., Athens, Greece: Association for Computational Linguistics, Mar. 2009, ss. 62–68. erişim adresi: https://aclanthology.org/W09-0808/.

[80] D. Ortiz-Arroyo, "Flexible Question Answering System for mobile devices," *2008 Third International Conference on Digital Information Management*, 2008, ss. 266–271. doi: 10.1109/ICDIM.2008.4746794.

[81] L. V. Lita J. G. Carbonell, "Cluster-Based Query Expansion for Statistical Question Answering," *International Joint Conference on Natural Language Processing*, 2008. erişim adresi: https://api.semanticscholar.org/CorpusID:17672649.

[82] J. Kürsten M. Eibl, "Putting It All Together: The Xtrieval Framework at Grid@CLEF 2009," *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ss. 570–577, isbn: 978-3-642-15754-7.

[83] M. Pardiño, J. Gómez, H. Llorens, P. Moreda M. Palomar, "Adapting IBQAS to work with text transcriptions in QAst task," *IBQAst: CEUR Workshop Proceedings*, 2008.

[84] P. R. Comas J. Turmo, "Robust Question Answering for Speech Transcripts: UPC Experience in QAst 2009," *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, C. Peters ve diğ., ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ss. 297–304, isbn: 978-3-642-15754-7.

[85] H. Bao, "An Answer Extraction Algorithm Based on Syntax Structure Feature Parsing and Classification," *Chinese Journal of Computers*, 2008. erişim adresi: https : / / api . semanticscholar . org / CorpusID:57918625.

[86] E. M. Voorhees L. P. Buckland, ed., *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, c. 500-274, NIST Special Publication, National Institute of Standards ve Technology (NIST), 2007. erişim adresi: http://trec.nist.gov/pubs/trec16/t16%5C_proceedings.html.

[87] N. Schlaefer, J. Ko, J. Betteridge, M. A. Pathak, E. Nyberg G. Sautter, "Semantic Extensions of the Ephyra QA System for TREC 2007," *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, E. M. Voorhees L. P. Buckland, ed., seri NIST Special Publication, c. 500-274, National Institute of Standards ve Technology (NIST), 2007. erişim adresi: http://trec.nist.gov/pubs/trec16/papers/cmu-ukarlsruhe.qa.final.pdf.

[88] A. Hickl ve diğ., "Question Answering with LCC's CHAUCER-2 at TREC 2007," *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*, seri NIST Special Publication, c. 500-274, National Institute of Standards ve Technology (NIST), 2007. erişim adresi: http://trec.nist.gov/pubs/trec16/papers/lcc-hickl.qa.final.pdf.

[89] M. Paşca, "Lightweight web-based fact repositories for textual question answering," seri CIKM '07, Lisbon, Portugal: Association for Computing Machinery, 2007, ss. 87–96, isbn: 9781595938039. doi: 10.1145/1321440.1321455. erişim adresi: https://doi.org/10.1145/1321440.1321455.

[90] C. Peters, "Multilingual information access: the contribution of evaluation," seri IWRIDL '06, Kolkata, India: Association for Computing Machinery, 2006, isbn: 1595936084. doi: 10.1145/1364742.1364761. erişim adresi: https://doi.org/10.1145/1364742.1364761.

[91] Y. Yang, S. Liu, S. Kuroiwa F. Ren, "Question Answering System of Confucian Analects based on Pragmatics Information and Categories," *2007 International Conference on Natural Language Processing and Knowledge Engineering*, 2007, ss. 361–366. doi: 10.1109/NLPKE.2007.4368056.

[92] C. Wartena, "Comparing segmentation strategies for efficient video passage retrieval," *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012, ss. 1–6. doi: 10.1109/CBMI.2012.6269850.

[93] M. A. Yarmohammadi, M. Shamsfard, M. A. Yarmohammadi M. Rouhizadeh, "Using WordNet in extracting the final answer from retrieved documents in a question answering system," GWC, 2008.

[94] J. Tiedemann, "Comparing document segmentation strategies for passage retrieval in question answering," *RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING*, 2007, s. 576.

[95] M. Hussain, A. Merkel D. Klakow, *Dedicated backing-off distributions for language model based passage retrieval*. Universitätsbibliothek Hildesheim, 2011.

[96] D. Jinguji ve diğ., "The University of Washington's UWclmaQA System.," *TREC*, Citeseer, 2006.

[97] S. Balantrapu, M. Khan A. Nagubandi, "TREC 2006 Q&A Factoid: TI Experience.," *TREC*, 2006.

[98] S. Balantrapu, M. Khan A. Nagubandi, "TREC 2006 Q&A Factoid: TI Experience.," *TREC*, 2006.

[99] D. Ferrés H. Rodríguez, "TALP at GeoCLEF 2006: experiments using JIRS and Lucene with the ADL feature type thesaurus," *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer, 2006, ss. 962–969.

[100] M. Garcıa-Cumbreres, L. Urena-López, F. Martınez-Santiago J. M. Perea-Ortega, "BRUJA System. The University of Jaén at the Spanish task of CLEFQA 2006," *Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.

[101] C. Blake, "A comparison of document, sentence, and term event spaces," *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, ss. 601–608.

[102] Z.-T. Yu, Z.-Y. Zheng, S.-P. Tang J.-Y. Guo, "Query expansion for answer document retrieval in Chinese question answering system," *2005 International Conference on Machine Learning and Cybernetics*, IEEE, c. 1, 2005, ss. 72–77.

[103] F. Jousse, I. Tellier, M. Tommasi P. Marty, "Learning to Extract Answers in Question Answering: Experimental Studies.," *CORIA*, 2005, s. 85.

[104] D. Ferrés, S. Kanaan, D. Dominguez-Sal, M. Surdeanu J. Turmo, "Experiments using a voting scheme among three heterogeneous QA systems," *NIST Special Publication*, 2005.

[105] G.-C. Yang H. U. Oh, "ANEX: An Answer Extraction System based on Conceptual Graphs.," *IKE*, 2005, ss. 17–24.

[106] J. Tiedemann, "Integrating linguistic knowledge in passage retrieval for question answering," *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, ss. 939–946.

[107] H. Isozaki, "An analysis of a high-performance Japanese question answering system," *ACM Transactions on Asian Language Information Processing (TALIP)*, c. 4, no. 3, ss. 263–279, 2005.

[108] J. Tiedemann, "Integrating linguistic knowledge in passage retrieval for question answering," *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, ss. 939–946.

[109] C. Amaral, H. Figueira, A. Martins, A. Mendes, P. Mendes C. Pinto, "Priberam's question answering system for Portuguese," *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer, 2005, ss. 410–419.

[110] J. Weston, S. Chopra A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[111] S. Antol ve diğ., "Vqa: Visual question answering," *Proceedings of the IEEE international conference on computer vision*, 2015, ss. 2425–2433.

[112] C. Xiong, S. Merity R. Socher, "Dynamic memory networks for visual and textual question answering," *International conference on machine learning*, PMLR, 2016, ss. 2397–2406.

[113] C. Raffel ve diğ., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, c. 21, no. 140, ss. 1–67, 2020.

[114] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[115] G. Izacard E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint arXiv:2007.01282*, 2020.

[116] M. Seo, A. Kembhavi, A. Farhadi H. Hajishirzi, "Bidirectional attention flow for machine comprehension. arXiv 2016," *arXiv preprint arXiv:1611.01603*, 2016.

[117] V. Karpukhin ve diğ., "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[118] X. Zhang, L. Lu M. Lapata, "Top-down tree long short-term memory networks," *arXiv preprint arXiv:1511.00060*, 2015.

[119] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[120] Q. You, H. Jin, Z. Wang, C. Fang J. Luo, "Image captioning with semantic attention," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, ss. 4651–4659.

[121] A. Rush, S. Harvard, S. Chopra J. Weston, "A neural attention model for sentence summarization. ACLWeb," *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2017.

[122] D. Yu ve diğ., "Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention.," *Interspeech*, 2016, ss. 17–21.

[123] M. Zanfir, E. Marinoiu C. Sminchisescu, "Spatio-temporal attention models for grounded video captioning," *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, Springer, 2017, ss. 104–119.

[124] Y. Chen, L. Wu M. J. Zaki, "Reinforcement learning based graph-to-sequence model for natural question generation," *arXiv preprint arXiv:1908.04942*, 2019.

[125] D. Gao, R. Wang, S. Shan X. Chen, "Cric: A vqa dataset for compositional reasoning on vision and commonsense," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, c. 45, no. 5, ss. 5561–5578, 2022.

[126] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiskỳ P. Blunsom, "Reasoning about entailment with neural attention," *arXiv preprint arXiv:1509.06664*, 2015.

[127] W. Yin, H. Schütze, B. Xiang B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for computational linguistics*, c. 4, ss. 259–272, 2016.

[128] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[129] D. Chen, J. Bolton C. D. Manning, "A thorough examination of the cnn/daily mail reading comprehension task," *arXiv preprint arXiv:1606.02858*, 2016.

[130] S. Wang, "Machine comprehension using match-LSTM and answer pointer," *arXiv preprint arXiv:1608.07905*, 2016.

[131] S. Liu, X. Zhang, S. Zhang, H. Wang W. Zhang, "Neural machine reading comprehension: Methods and trends," *Applied Sciences*, c. 9, no. 18, s. 3698, 2019.

[132] D. Chen, "Reading Wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[133] S. Wang ve diğ., "R3: Reinforced ranker-reader for open-domain question answering," *Proceedings of the AAAI Conference on Artificial Intelligence*, c. 32, 2018.

[134] R. Das, S. Dhuliawala, M. Zaheer A. McCallum, "Multi-step retriever-reader interaction for scalable open-domain question answering," *arXiv preprint arXiv:1905.05733*, 2019.

[135] K. Guu, K. Lee, Z. Tung, P. Pasupat M. Chang, "Retrieval augmented language model pre-training," *International conference on machine learning*, PMLR, 2020, ss. 3929–3938.

[136] Y. Lin, H. Ji, Z. Liu M. Sun, "Denoising distantly supervised open-domain question answering," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, ss. 1736–1745.

[137] Y. Feldman R. El-Yaniv, "Multi-hop paragraph retrieval for open-domain question answering," *arXiv preprint arXiv:1906.06606*, 2019.

[138] W. Yang ve diğ., "End-to-end open-domain question answering with bertserini," *arXiv preprint arXiv:1902.01718*, 2019.

[139] S. Min, D. Chen, L. Zettlemoyer H. Hajishirzi, "Knowledge guided text retrieval and reading for open domain question answering," *arXiv preprint arXiv:1911.03868*, 2019.

[140] P. Lewis ve diğ., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 2021. arXiv: 2005.11401 [cs.CL]. erişim adresi: https://arxiv.org/abs/2005.11401.

[141] W. Xiong ve diğ., *Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval*, 2021. arXiv: 2009.12756 [cs.CL]. erişim adresi: https://arxiv.org/abs/2009.12756.

[142] D. Bakır M. S. Aktas, "A systematic literature review of question answering: Research trends, datasets, methods," *International Conference on Computational Science and Its Applications*, Springer, 2022, ss. 47–62.

[143] G. Aydin ve diğ., "Building and applying geographical information system Grids," *Concurr. Comput.: Pract. Exper.*, c. 20, no. 14, ss. 1653–1695, Eyl. 2008, issn: 1532-0626.

[144] M. Aktas ve diğ., "iSERVO: Implementing the International Solid Earth Research Virtual Observatory by Integrating Computational Grid and Geographical Information Web Services," *Computational Earthquake Physics: Simulations, Analysis and Infrastructure, Part II*, Basel: Birkhäuser Basel, 2007, ss. 2281–2296, isbn: 978-3-7643-8131-8.

[145] G. Aydin, M. S. Aktas, G. C. Fox, H. Gadgil, M. Pierce A. Saya, "SERVOGrid complexity computational environments (CCE) integrated performance analysis," *The 6th IEEE/ACM International Workshop on Grid Computing, 2005.*, IEEE, 2005, 6–pp.

[146] M. E. Pierce ve diğ., "The QuakeSim project: Web services for managing geophysical data and applications," *Earthquakes: Simulations, Sources and Tsunamis*, ss. 635–651, 2008.

[147] G. C. Fox ve diğ., "Algorithms and the Grid," *Computing and visualization in science*, c. 12, ss. 115–124, 2009.

[148] A. Tufek, A. Gurbuz, O. F. Ekuklu M. S. Aktas, "Provenance collection platform for the weather research and forecasting model," *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*, IEEE, 2018, ss. 17–24.

[149] M. S. Aktas M. Astekin, "Provenance aware run-time verification of things for self-healing Internet of Things applications," *Concurrency and Computation: Practice and Experience*, c. 31, no. 3, e4263, 2019.

[150] B. Dundar, M. Astekin M. S. Aktas, "A Big Data Processing Framework for Self-Healing Internet of Things Applications," *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, 2016, ss. 62–68. doi: 10.1109/SKG.2016.017.

[151] C. Callison-Burch, M. Osborne P. Koehn, "Re-evaluating the Role of Bleu in Machine Translation Research," *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy: Association for Computational Linguistics, Nis. 2006, ss. 249–256. erişim adresi: https://aclanthology.org/E06-1032/.

[152] A. R, P. Bhattacharyya, M. Sasikumar R. M. Shah, "Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU," 2006. erişim adresi: https://api.semanticscholar.org/CorpusID:5690091.

[153] C. Callison-Burch, "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk," *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, P. Koehn R. Mihalcea, ed., Singapore: Association for Computational Linguistics, Ağus. 2009, ss. 286–295. erişim adresi: https://aclanthology.org/D09-1030/.

[154] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger Y. Artzi, *BERTScore: Evaluating Text Generation with BERT*, 2020. arXiv: 1904.09675 [cs.CL]. erişim adresi: https://arxiv.org/abs/1904.09675.

[155] M. Eyal, T. Baumel M. Elhadad, "Question Answering as an Automatic Evaluation Metric for News Article Summarization," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Haz. 2019, ss. 3938–3948. doi: 10.18653/v1/N19-1395. erişim adresi: https://aclanthology.org/N19-1395/.

[156] K. S. Jones J. Galliers, "Book Reviews: Evaluating Natural Language Processing Systems: An Analysis and Review," *Evaluating Natural Language Processing Systems*, 1996. erişim adresi: https://api.semanticscholar.org/CorpusID:34479739.

[157] S. Chakraborty P. Pakray, "Abstractive Summarization Evaluation for Prompt Engineering," *Advances in Visual Informatics*, Singapore: Springer Nature Singapore, 2024, ss. 629–640, isbn: 978-981-99-7339-2.

[158] Y. Gao, C. Sun R. J. Passonneau, "Automated Pyramid Summarization Evaluation," *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China: Association for Computational Linguistics, Kas. 2019, ss. 404–418. doi: 10.18653/v1/K19-1038. erişim adresi: https://aclanthology.org/K19-1038/.

[159] S. Narayan, S. B. Cohen M. Lapata, "Ranking Sentences for Extractive Summarization with Reinforcement Learning," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Haz. 2018, ss. 1747–1759. doi: 10.18653/v1/N18-1158. erişim adresi: https://aclanthology.org/N18-1158/.

[160] E. Durmus, H. He M. Diab, "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.454. erişim adresi: http://dx.doi.org/10.18653/v1/2020.acl-main.454.

[161] S. Gu, X. Luo H. Wang, "Enhancing Answer Selection via Ad-Hoc Knowledge Extraction from Unstructured Web Texts," *International Journal of Software Engineering and Knowledge Engineering*, c. 33, no. 06, ss. 933–951, 2023. doi: 10.1142/S0218194023500201. eprint: https://doi.org/10.1142/S0218194023500201. erişim adresi: https://doi.org/10.1142/S0218194023500201.

[162] K. Al-Sabahi Z. Zuping, "Document Summarization Using Sentence-Level Semantic Based on Word Embeddings," *International Journal of Software Engineering and Knowledge Engineering*, c. 29, no. 02, ss. 177–196, 2019. doi: 10.1142/S0218194019500086. eprint: https://doi.org/10.1142/S0218194019500086. erişim adresi: https://doi.org/10.1142/S0218194019500086.

[163] S. Yuan, H. Qin, X. Gu B. Shen, "Clean and Learn: Improving Robustness to Spurious Solutions in API Question Answering," *International Journal of Software Engineering and Knowledge Engineering*, c. 32, no. 07, ss. 1101–1123, 2022. doi: 10.1142/S0218194022500449. eprint: https://doi.org/10.1142/S0218194022500449. erişim adresi: https://doi.org/10.1142/S0218194022500449.

[164] K. Chen, G. Shen, Z. Huang H. Wang, "Improved Entity Linking for Simple Question Answering Over Knowledge Graph," *International Journal of Software Engineering and Knowledge Engineering*, c. 31, no. 01, ss. 55–80, 2021. doi: 10.1142/S0218194021400039. eprint: https://doi.org/10.1142/S0218194021400039. erişim adresi: https://doi.org/10.1142/S0218194021400039.

[165] E. Sezerer, S. Tenekeci, A. Acar, B. Baloğlu S. Tekir, "Author Reputation Measurement on Question and Answer Sites by the Classification of Author-Generated Content," *International Journal of Software Engineering and Knowledge Engineering*, c. 31, no. 10, ss. 1421–1445, 2021. doi: 10.1142/S0218194021500479. eprint: https://doi.org/10.1142/S0218194021500479. erişim adresi: https://doi.org/10.1142/S0218194021500479.

[166] C. Li, Y. Wang, D. Li, D. Chu M. Ma, "An Effective Method of Evaluating Pension Service Quality Using Multi-Dimension Attention Convolutional Neural Networks," *International Journal of Software Engineering and Knowledge Engineering*, c. 31, no. 04, ss. 533–543, 2021. doi: 10.1142/S0218194021400064. eprint: https://doi.org/10.1142/S0218194021400064. erişim adresi: https://doi.org/10.1142/S0218194021400064.

[167] S. LI Z. LI, "QUESTION-ORIENTED ANSWER SUMMARIZATION VIA TERM HIERARCHICAL STRUCTURE," *International Journal of Software Engineering and Knowledge Engineering*, c. 21, no. 06, ss. 877–889, 2011. doi: 10.1142/S0218194011005475. eprint: https://doi.org/10.1142/S0218194011005475. erişim adresi: https://doi.org/10.1142/S0218194011005475.

[168] T. SAKAI, "Advanced Technologies for Information Access," *International Journal of Computer Processing of Languages*, c. 18, no. 02, ss. 95–113, 2005. doi: 10.1142/S0219427905001274. eprint: https://doi.org/10.1142/S0219427905001274. erişim adresi: https://doi.org/10.1142/S0219427905001274.

[169] P. FRAGKOU, "INFORMATION EXTRACTION VERSUS TEXT SEGMENTATION FOR WEB CONTENT MINING," *International Journal of Software Engineering and Knowledge Engineering*, c. 23, no. 08, ss. 1109–1137, 2013. doi: 10.1142/S0218194013500332. eprint: https://doi.org/10.1142/S0218194013500332. erişim adresi: https://doi.org/10.1142/S0218194013500332.

[170] C. Song E. Cho, "XL-BPMN Model-Based Service Similarity Measurement Technique," *International Journal of Software Engineering and Knowledge Engineering*, c. 33, no. 05, ss. 697–732, 2023. doi: 10.1142/S0218194023500122. eprint: https://doi.org/10.1142/S0218194023500122. erişim adresi: https://doi.org/10.1142/S0218194023500122.

[171] F. Cirett-Galán ve diğ., "Assessing the Use of GitHub Copilot on Students of Engineering of Information Systems," *International Journal of Software Engineering and Knowledge Engineering*, c. 34, no. 11, ss. 1717–1734, 2024. doi: 10.1142/S0218194024500335. eprint: https://doi.org/10.1142/S0218194024500335. erişim adresi: https://doi.org/10.1142/S0218194024500335.

[172] H. Lee ve diğ., *KPQA: A Metric for Generative Question Answering Using Keyphrase Weights*, 2021. arXiv: 2005.00192 [cs.CL]. erişim adresi: https://arxiv.org/abs/2005.00192.

[173] G. Aydin ve diğ., "Building and applying geographical information system Grids," *Concurr. Comput.: Pract. Exper.*, c. 20, no. 14, ss. 1653–1695, Eyl. 2008, issn: 1532-0626.

[174] M. A. N. E. Al., "VLab: Collaborative Grid services and portals to support computational material science," *Concurrency and Computation: Practice and Experience*, ss. 1717–1728, 2007.

[175] G. C. Fox ve diğ., "Algorithms and the Grid," *Comput. Vis. Sci.*, c. 12, no. 3, ss. 115–124, Mar. 2009, issn: 1432-9360.

[176] M. Aktas ve diğ., "iSERVO: Implementing the International Solid Earth Research Virtual Observatory by Integrating Computational Grid and Geographical Information Web Services," *Computational Earthquake Physics: Simulations, Analysis and Infrastructure, Part II*, Basel: Birkhäuser Basel, 2007, ss. 2281–2296, isbn: 978-3-7643-8131-8.

[177] M. E. Pierce ve diğ., "The QuakeSim Project: Web Services for Managing Geophysical Data and Applications," içinde *Earthquakes: Simulations, Sources and Tsunamis*. Basel: Birkhäuser Basel, 2008, ss. 635–651, isbn: 978-3-7643-8757-0. doi: 10.1007/978-3-7643-8757-0_11. erişim adresi: https://doi.org/10.1007/978-3-7643-8757-0_11.

[178] M. Kapdan, M. Aktas M. Yigit, "On the Structural Code Clone Detection Problem: A Survey and Software Metric Based Approach," *Computational Science and Its Applications – ICCSA 2014*, Cham: Springer International Publishing, 2014, ss. 492–507, isbn: 978-3-319-09156-3.

[179] M. Sahinoglu, K. Incki M. S. Aktas, "Mobile Application Verification: A Systematic Mapping Study," *Computational Science and Its Applications – ICCSA 2015*, O. Gervasi ve diğ., ed., Cham: Springer International Publishing, 2015, ss. 147–163, isbn: 978-3-319-21413-9.

[180] M. K. H. Briman B. Yildiz, "Beyond ROUGE: A Comprehensive Evaluation Metric for Abstractive Summarization Leveraging Similarity, Entailment, and Acceptability," *International Journal on Artificial Intelligence Tools*, c. 33, no. 05, s. 2 450 017, 2024. doi: 10.1142/S0218213024500179. eprint: https://doi.org/10.1142/S0218213024500179. erişim adresi: https://doi.org/10.1142/S0218213024500179.

[181] Y. Uygun, R. F. Oguz, E. Olmezogullari M. S. Aktas, "On the Large-scale Graph Data Processing for User Interface Testing in Big Data Science Projects," *2020 IEEE International Conference on Big Data (Big Data)*, Ara. 2020, ss. 2049–2056. doi: 10.1109/BigData50022.2020.9378153.

[182] E. Olmezogullari M. S. Aktaş, "Representation of Click-Stream DataSequences for Learning User Navigational Behavior by Using Embeddings," *2020 IEEE International Conference on Big Data (Big Data)*, ss. 3173–3179, 2020. erişim adresi: https://api.semanticscholar.org/CorpusID:232374538.

[183] E. Olmezogullari M. S. Aktaş, "Representation of Click-Stream DataSequences for Learning User Navigational Behavior by Using Embeddings," *2020 IEEE International Conference on Big Data (Big Data)*, ss. 3173–3179, 2020. erişim adresi: https://api.semanticscholar.org/CorpusID:232374538.

[184] M. S. A. Dilan Bakır B. Yıldız, *"Survey interface design codes" and "User interface website"*, 2024. erişim adresi: https://github.com/dilanbakr-/qametrichub,%20https://qa-metric-hub.azurewebsites.net.

[185] P. Rajpurkar, J. Zhang, K. Lopyrev P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Kas. 2016, ss. 2383–2392. doi: 10.18653/v1/D16-1264. erişim adresi: https://aclanthology.org/D16-1264/.

[186] M. S. A. Dilan Bakır B. Yıldız, *Evaluation metric train and test dataset created using SQUAD dataset*, 2024. erişim adresi: https://huggingface.co/datasets/dilanbakr/squad-qametric-traindata, %20https://huggingface.co/datasets/dilanbakr/squad-qametric-testdata.

[187] M. S. A. Dilan Bakır B. Yıldız, *"MQA-metric model" and "Evaluation metric test dataset created using MARCO dataset"*, 2024. erişim adresi: https://huggingface.co/dilanbakr/mistral-updated-model-for-qa%5C_scoresystem-evaluation, %20https://huggingface.co/datasets/dilanbakr/marco-qametric-testdata.

TEZDEN ÜRETİLMİŞ YAYINLAR

Makale

1. D. Bakir, M. S. Aktas, and B. Yildiz, "A model-based evaluation metric for question answering systems," *Int. J. Softw. Eng. Knowl. Eng.*, 2024, doi: 10.1142/S0218194025500032.

Konferans Bildirisi

1. D. Bakir and M. S. Aktas, "A systematic literature review of question answering: Research trends, datasets, methods," in *Computational Science and Its Applications: ICCSA 2022 Workshops*, Springer International Publishing, 2022, pp. 47–62.

2. D. Bakir and M. S. Aktas, "A business workflow for providing open-domain question answering reader systems on the Wikipedia dataset," in *Proc. 2022 IEEE Int. Conf. Big Data (Big Data)*, Osaka, Japan, 2022, pp. 3308–3314, doi: 10.1109/BigData55660.2022.10020945.

3. D. Bakir and M. S. Aktas, "An approach to decentralized hybrid question answering systems," in *Computational Science and Its Applications: ICCSA 2023 Workshops*, Springer Nature Switzerland, 2023, pp. 29–38.

4. D. Bakir, M. S. Aktas, and B. Yildiz, "Developing and evaluating a model-based metric for legal question answering systems," in *Proc. 2023 IEEE Int. Conf. Big Data (Big Data)*, Sorrento, Italy, 2023, pp. 2745–2754, doi: 10.1109/BigData59044.2023.10386689.

Proje

1. Website link: https://sites.google.com/view/qa-metric-improve/home

2. Github Kod link: https://github.com/dilanbakr/qametrichub

3. Youtube Link: Youtube link: https://youtu.be/ubL2NyZT9tQ?si=HbWIb7nL3JbL2MHB