Soru Cevaplama Sistemleri Üzerine Detaylı Bir Çalışma: Veri Kümeleri, Yöntemler ve Açık Araştırma Alanları

Literatür Makalesi/Review Article

Gülsüm YİĞİT[1]*, Mehmet Fatih AMASYALI[2]

[1] Bilgisayar Mühendisliği, Kadir Has Üniversitesi, İstanbul, Türkiye
[2] Bilgisayar Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul, Türkiye
gulsum.yigit@khas.edu.tr, amasyali@yildiz.edu.tr
(Geliş/Received:14.10.2020; Kabul/Accepted:25.05.2021)
DOI: 10.17671/gazibtd.810362

Özet— Soru Cevaplama (QA) sistemleri, kullanıcıların doğal dilde sordukları sorulara belge veya bağlantıları listelemek yerine doğrudan cevap almalarını sağlayan sistemlerdir. Bu çalışmada, QA sistemlerinde yaygın kullanılan veri kümeleri tanıtılmış ve çeşitli özelliklere göre karşılaştırılmıştır. Ayrıca, QA alanındaki diğer çalışmalardan farklı olarak bu çalışmada son yıllarda literatürde yer alan QA sistemlerinin arkasında kullanılan yöntemlere odaklanılmıştır. Bu yöntemler dört farklı grupta ele alınmış olup literatürdeki güncel çalışmaları ve teknolojileri içermektedir. Bu modeller kullanılan teknikler, harici bilgi kaynaklarının veya dil modelinin kullanılıp kullanılmadığı gibi faktörlere göre karşılaştırılmıştır. Dikkat mekanizmasının, dil modellerinin, çizge işleyen ağların, harici bilgi kaynaklarının, kolektif öğrenmenin ve derin öğrenme mimarilerinin QA sistemlerinin başarısı üzerinde genel olarak olumlu etkisi olduğu görülmüştür. Ayrıca, bu çalışmada QA sistemlerinin günümüzdeki açık araştırma alanları ve olası çözüm yolları belirlenerek gelecekteki QA sistemleri için önerilerde bulunulmuştur. Gelecekteki araştırma alanları olarak yeterli veriye sahip olmayan diller üzerindeki sistemler, birden fazla dil üzerinde çalışabilen sistemler, çok sayıda bilgi kaynağının kullanılmasının gerekli olduğu sistemler ve karşılıklı konuşmaya dayalı sistemler öne çıkmaktadır.

Anahtar Kelimeler— soru cevaplama, derin öğrenme, bilgi tabanları, bellek ağları, seq2seq

A Comprehensive Study on Question Answering Systems: Datasets, Methods and Open Research Areas

Abstract— Question Answering (QA) systems allow users to get direct answers to questions they ask in natural language instead of listing documents or links. In this study, current QA datasets are introduced and compared according to various properties. Unlike other studies in QA, this study focuses on the methods used in current QA systems. These methods are discussed in four different categories and include recent studies and technologies. The models are compared with various factors such as techniques used, external knowledge, or language model. In general, attention mechanisms, language models, graph neural networks, external knowledge, collective learning, and deep learning architectures positively affect the success of QA systems. In addition, current open research areas of QA systems and possible solutions are determined, and suggestions for future QA systems are given. Systems on languages that do not have enough data, systems that can work on more than one language, systems that require the use of many information sources, and speech-based systems stand out as future research areas.

Keywords— question answering, deep learning, knowledge bases, memory networks, seq2seq

1. GİRİŞ (INTRODUCTION)

Teknoloji ve internetin her geçen gün yaygınlaşması ile birlikte etkili hesaplamaya olan ilgi artmakta ve her gün büyük miktarda veri üretilerek kullanıma sunulmaktadır. Bilgi toplamadaki bu hızlı artış ile beraber internetteki büyük miktardaki verinin birleştirilip sorgulanması bilgiye erişmeyi karmaşık ve ciddi zaman alan bir görev haline getirmektedir. Bilgiyi etkili kullanmada karşılaşılan bu

zorluk günümüzde Soru Cevaplama (QA) sistemleri adı verilen araçların geliştirilmesine yol açmıştır.

QA sistemleri, kullanıcılar tarafından belirli bir sorgu dili, sorgu tipi, sorgu oluşturma kuralları ve belirli bir alana sınırlı kalmadan kendi doğal dilleri ile yapılandırılmış bir veri tabanı veya doğal dil belgeleri kullanarak sorguları otomatik olarak cevaplayan sistemlerdir. Günümüzde kullanılan arama motorları artık doğal dil ile yöneltilen sorulara bağlantı listesi sıralamak yerine doğrudan sorunun cevabını sunmaya başlamışlardır. Ancak, arama motorlarındaki bu gelişme kullanılan doğal dillere bağlı olarak farklılık göstermekle beraber belirli soru tipleri ile sınırlı kalmış, henüz karmaşık muhakeme gerektiren sorularda yeterli başarıya ulaşılamamıştır.

QA sistemlerindeki son dönemlerdeki ilerlemenin kilit faktörlerinden biri 2013 yılında Mikolov ve arkadaşlarının [1]'deki çalışmalarında anlamsal olarak birbirine yakın olan kelimelerin vektör uzayında birbirine yakın koordinatlara kümelenmesini sağlayan kelime vektörlerinin kullanımını önermesi gösterilebilir [2], [3]. Ayrıca, sistemin bir metin içerisinde hedeflenen bir alana odaklanmasına olanak sağlayan dikkat mekanizmasının kullanılması QA sistemlerindeki gelişmelerde büyük bir öneme sahiptir [4–6]. Literatürdeki QA sistemleri incelendiğinde bir diğer önemli faktör bilgi tabanlı (KB) sistemlerin kullanılmasıdır. Freebase [7] ve YAGO [8] gibi büyük yapılandırılmış KB'lerin ortaya çıkmasından sonra yapılandırılmış KB'leri kullanarak doğal dil sorularını yanıtlayan KB temelli QA (KB-QA) doğal dil işlemenin (NLP) araştırma alanlarından biri haline gelmiştir [9–12]. KB-QA sistemlerinde sorunun anlamsal bir gösterimi (zaman, tarih, yer, varlık, sayısal büyüklükler) oluşturulur. Örneğin "Ankara'da kaç kişi yaşıyor?" sorusundan "nüfus" etiketinin bulunması ile sorunun cevabına ulaşılabilmektedir.

Literatürde QA sistemleri üzerine birçok derleme çalışması bulunmaktadır. Wasim ve arkadaşları (2017) Biyomedikal Soru Cevaplama (BioQA) sistemlerine odaklanmış ve sistem performanslarını değerlendirmek için kullanılan veri kümelerinin incelemesi ve özelliklerini analiz eden bir çalışma yapmıştır [13]. Kolomiyets ve arkadaşları (2011) QA sistemlerinde bilginin çıkarılması, NLP ve bilgi erişimi ile etkileşimlere odaklanmıştır [14]. Diefenbach ve arkadaşları (2017) KB'lerden cevabı çıkarmak için kullanılan tekniklere odaklanmış soru analizi, kelime öbeği eşleme, belirsizliği giderme, sorgu oluşturma ve dağıtılmış bilgiyi sorgulama aşamalarında kullanılan yöntemleri incelemiştir [15]. Genel olarak bu yöntemler NLP'nin klasik dilbilimsel tekniklerinden olan sözcük türü (POS) etiketleme, varlık tanıma (NER) ve ayrıştırıcılarıdır.

Bu çalışmada, QA sistemlerinin eğitimi için kullanılan veri kümelerine ve kullanılan yöntemlere odaklanılmıştır. Günümüzde birçok alanda olduğu gibi QA sistemlerinde de uygulanan yöntemlerin performanslarını karşılaştırmak için çeşitli veri kümeleri yayınlanmaktadır. Böylece, geliştirilen yöntemler aynı veri kümeleri üzerinde eğitilerek adil biçimde karşılaştırılabilmektedir. Bu veri kümeleri için karşılaştırmalı sonuçların yer aldığı web siteleri de bulunmaktadır. QA veri kümelerinin yer aldığı bu tarz web sitelerinin incelenmesi ile çalışma kapsamındaki veri kümeleri ve yöntemler belirlenmiştir. Veri kümelerinin seçiminde QA sistemlerinin çeşitli uygulama alanları üzerinde en çok çalışma yapılanlar, yöntemlerin seçiminde ise veri kümelerinde en çok başarı sağlayan çalışmalar ele alınmıştır. Ayrıca, bu çalışmada QA sistemlerinin eksiklikleri incelenmiş gelecekteki QA sistemleri için araştırma alanları belirlenmeye çalışılmıştır.

Bölüm 2'de, popüler QA veri kümeleri incelenmiş olup belirli niteliklere göre karşılaştırılması yapılmıştır. Bölüm 3'te QA sistemlerinin değerlendirilmesi için kullanılan başarı kriterleri verilmiştir. Bölüm 4'te, ana katkılarının niteliğine göre Seq2Seq, Bellek Ağları, Çizge İşleyen Ağlar ve Klasik Yöntemler olmak üzere dört ana kategoride QA yöntemlerinin derinlemesine bir araştırması verilmiştir. Bölüm 5'te, Bölüm 4'te belirlenen yöntemlerin QA veri kümeleri üzerindeki karşılaştırmalı değerlendirmesi sunulmuştur. Bölüm 6'te literatürdeki QA sistemlerinin eksiklikleri belirlenerek olası araştırma yönleri ortaya konmuştur. Bölüm 7 çalışmanın sonuç bölümüdür.

2. SORU CEVAPLAMADA KULLANILAN VERİ KÜMELERİ (DATASETS USED IN QUESTION ANSWERING)

Bu bölümde, QA sistemlerinin eğitimi ve değerlendirilmesi için yaygın kullanılan veri kümeleri tanıtılmıştır. Çeşitli uygulama alanları bulunan farklı yapılardaki bu veri kümelerinden İngilizce olanlar Türkçe çevirileri ile birlikte verilmiştir.

2.1 bAbi Görevleri (bAbi Tasks)

Facebook'un bAbi görevleri veri kümesi 20 farklı görevden oluşan her biri farklı türde ve zorluk derecesine sahip sorular içeren sentetik bir veri kümesidir [16]. Her bir görev için eğitim kümesinde 1.000 veya 10.000, test kümesinde ise 1.000 soru bulunmaktadır. bAbi görevleri veri kümesinden alınan örnekler Şekil 1'te verilmiştir.

en	Example	
Task 17	1 The pink rectangle is to the left of the triangle. 2 The triangle is to the left of the red square. 3 Is the pink rectangle to the right of the red square? 4 Is the pink rectangle to the left of the red square?	 no yes

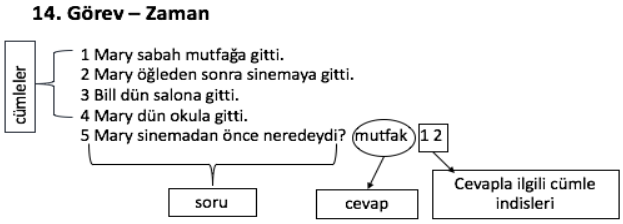
tr	Örnek	
17. Görev	1 Pembe dikdörtgen, üçgenin solundadır. 2 Üçgen, kırmızı karenin solundadır. 3 Pembe dikdörtgen kırmızı karenin sağında mı? 4 Kırmızı karenin solundaki pembe dikdörtgen mi?	 hayır evet

Şekil 1. bAbitasks veri kümesinden örnekler (Examples from bAbitasks dataset)

2.2 Türkçe bAbi Görevleri (Turkish bAbitasks)

Türkçe için sınırlı sayıda bulunan QA sistemleri veri kümelerine katkı sağlamak amacıyla [17]'da Türkçe QA veri kümesi oluşturulmuştur. [16]'de önerilen bAbitasks veri kümesi Türkçe diline makine çevirisi ile çevrildikten sonra anlam bütünlüğünü sağlamak amacıyla her bir

görevin metinleri üzerinde düzeltmeler yapılmıştır. 20 farklı görevden oluşan veri kümesi [16]'deki İngilizce sürümünden farklı olarak her bir görevdeki sözlük boyutunun artırılması, metindeki cümle sayılarının artırılması gibi değişiklikler uygulanarak yeniden oluşturulmuştur. Kelime sayısındaki artış cümlelerdeki yer, aktör ve nesne sayılarının artırılmasıyla elde edilmiştir. Türkçe bAbitasks veri kümesinden örnekler Şekil 2'de verilmiştir.



Şekil 2. Türkçe bAbi veri kümesinden örnekler (Examples from turkish bAbitasks dataset)

2.3 HotpotQA (HotpotQA)

HotpotQA, kalabalık bir insan grubu tarafından İngilizce Wikipedia makalelerini kullanarak oluşturulan büyük ölçekli bir QA sistemleri veri kümesidir [18]. 113.000 soru-cevap çiftinden oluşmaktadır. HotpotQA birden fazla belgedeki bilgilerin kullanılmasıyla cevaplandırılabilecek sorular içermektedir. Şekil 3'de HotpotQA veri kümesinden örnek metin ve soru-cevap verilmiştir.

Paragraph A, Return to Olympus: **[1]** Return to Olympus is the only album by the alternative rock band Malfunkshun. **[2]** It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. **[3]** Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records. **Paragraph B, Mother Love Bone:** **[4]** Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987. **[5]** The band was active from 1987 to 1990. **[6]** Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene. **[7]** Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success. **[8]** The album was finally released a few months later. **Q:** What was the former band of the member of Mother Love Bone who died just before the release of "Apple"? **A:** Malfunkshun **Supporting facts:** 1, 2, 4, 6, 7	**Paragraf A, Olympus'a Dönüş:** **[1]** Olympus'a Dönüş, alternatif rock grubu Malfunkshun'un tek albümüdür. **[2]** Grubun dağılmasından ve baş vokalisti Andrew Wood'un (daha sonra Mother Love Bone'un) 1990 yılında aşırı dozda uyuşturucudan ölmesinden sonra yayınlandı. **[3]** Pearl Jam'den Stone Gossard, şarkıları derlemiş ve albümü Loosegroove Records etiketiyle yayınlamıştı. **Paragraf B, Mother Love Bone :** **[4]** Mother Love Bone, 1987'de Seattle, Washington'da kurulan bir Amerikan rock grubuydu. **[5]** Grup 1987'den 1990'a kadar etkindi. **[6]** Öncü Andrew Wood'un kişiliği ve bestleri, grubu 1980'lerin sonu / 1990'ların başı Seattle müzik sahnesinin gelişmekte olan zirvesine ulaştırmasına yardımcı oldu. **[7]** Wood, grubun ilk albümü olan "Apple" ın planlanan çıkışından yalnızca günler önce öldü ve böylece grubun başarı umutları sona erdi. **[8]** Albüm nihayet birkaç ay sonra yayınlandı. **S:** "Apple" ın yayınlanmasından hemen önce ölen Mother Love Bone üyesinin eski grubu neydi? **C:** Malfunkshun **Destekleyici cümleler:** 1, 2, 4, 6, 7

Şekil 3. HotpotQA veri kümesindeki çoklu geçiş sorularına örnek [18] (HotpotQA: An example of multi-hop questions in HotpotQA dataset [18])

2.4 SQuAD (SQuAD)

SQuAD 1.1 536 Wikipedia makalesinden toplanan 100.000'den fazla soru bulunan bir veri kümesidir. Kullanılan makaleler şekil ve tablolardan arındırılarak metin paragraflara ayrılmıştır [19]. Ayrıca, bu veri kümesinde her bir sorunun cevabı ilgili paragrafta bir metin parçasıdır.

SQuAD 1.1 sürümündeki cevaplanabilir sorular aynı paragraflar ile ilgili 53.775 yeni, cevabı bulunmayan sorular ile birleştirilip veri kümesinin bir başka versiyonu olan SQuAD 2.0 oluşturulmuştur [20]. Şekil 4'te SQuAD 2.0 veri kümesinden örnek paragraf, soru ve cevap verilmiştir.

Text	Metin
Beyonce Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. ..	Beyonce Giselle Knowles-Carter (d. 4 Eylül 1981) Amerikalı şarkıcı, söz yazarı, plak yapımcısı ve oyuncudur. Houston, Teksas'ta doğup büyümüştür. Çocukken çeşitli şarkı ve dans yarışmalarında performans sergiledi ve 1990'ların sonunda R&B kız grubu Destiny's Child'ın solisti olarak ün kazandı. ..
Question: In what city and state did Beyonce, grow up? **Answer:** Houston, Texas	**Soru:** Beyonce hangi şehir ve eyalette büyüdü? **Cevap:** Houston, Texas
Question: When did Beyonce start becoming popular? **Answer:** in the late 1990s	**Soru:** Beyonce ne zaman popüler olmaya başladı? **Cevap:** 1990'ların sonunda

Şekil 4. SQuAD veri kümesinden örnek metin ve soru-cevaplar (Example story and question-answer pair in SQuAD dataset)

2.5 CoQA (CoQA)

CoQA, karşılıklı konuşmaya dayalı QA sistemleri için kullanılan bir veri kümesidir [21]. Çeşitli alanlardan metinler hakkındaki 8.000 konuşmadan elde edilen 127.000 soru içermektedir. CoQA'daki bazı soruların birden fazla geçerli cevabı olabilir. Ayrıca, veri kümesi cevabı olmayan soruları da içermektedir ve bu tür soruların oranı %1.3 civarındadır. Şekil 5'te CoQA veri kümesinden örnek paragraf, soru-cevap verilmiştir.

Text	Metin
Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. …	Jessica sallanan koltuğuna oturmaya gitti. Bugün onun doğum günüydü ve 80 yaşına basıyordu. Torunu Annie öğleden sonra geliyordu ve Jessica onu gördüğü için çok heyecanlıydı. …
Round #1	**1. TUR**
Question: Who had a birthday? **Answer:** Jessica **Supporting Text:** Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.	**Soru:** Kimin doğum günü vardı? **Cevap:** Jessica **Destekleyici Metin:** Jessica sallanan sandalyesine oturmaya gitti. Bugün onun doğum günüydü ve 80 yaşına basıyordu.
Round #2	**2. TUR**
Question: How old would she be? **Answer:** 80 **Supporting Text :** "she was turning 80"	**Soru:** Kaç yaşında olurdu? **Cevap:** 80 **Destekleyici Metin :** "80 yaşına giriyordu"
Round #3	**3. TUR**
Question: Did she plan to have any visitors? **Answer:** Yes **Supporting Text :** "Her granddaughter Annie was coming over"	**Soru:** Ziyaretçileri olmasını planlıyor muydu? **Cevap:** Evet **Destekleyici Metin :** "Torunu Annie geliyordu"

Şekil 5. CoQA veri kümesinden örnek (An example from CoQA Dataset)

2.6 MovieQA (MovieQA)

Movie Name: Moon	**Film Adı:** Ay
Text	**Metin**
After a heated argument and physical altercation, they together coerce GERTY into revealing that they are both clones of the original Sam Bell.	Hararetli bir tartışma ve fiziksel çekişmeden sonra, onlar GERTY'yi her ikisinin de orijinal Sam Bell'in klonları olduğunu ortaya çıkarmaya zorlarlar.
Question	**Soru**
What do the two Sams learn from Gerty about themselves?	İki Sams, Gerty'den kendileri hakkında ne öğreniyor?
Options	**Seçenekler**
Answer 1: That they are both clones of Sam Bell	**Cevap 1:** Her ikisinin de Sam Bell'in klonları olduğu
Answer 2: That one of them is a clone of the other **Answer 3:** That both of them were activated after the rover crash **Answer 4:** That they are both going back to Earth **Answer 5:** That they are both clones of Gerty	**Cevap 2:** Birinin diğerinin klonu olduğunu **Cevap 3:** Her ikisinin de gezici kazasından sonra etkinleştirildiğini **Cevap 4:** Her ikisinin de Dünya'ya geri döneceği **Cevap 5:** Her ikisinin de Gerty'nin klonları olduğu

Şekil 6. MovieQA veri kümesinden örnek (Example from MovieQA dataset)

MovieQA, otomatik hikâye anlama özelliğini hem video hem de metinden değerlendirmeyi amaçlayan bir QA sistemleri veri kümesidir [22]. MovieQA, yüksek anlamsal çeşitliliğe sahip 408 film hakkındaki 14.944 sorudan oluşmaktadır. Veri kümesinin büyük bir bölümü "Neden", "Nasıl" gibi muhakeme temelli soruları içermek ile birlikte "Kim", "Ne", "Kime" gibi sorulara da cevap aramaktadır. Kalabalık bir grup insan tarafından oluşturulmuş olan veri kümesinde her soru için beş olası cevap seçeneği bulunmaktadır. Şekil 6'da MovieQA veri kümesinden örnek metin, soru ve cevap seçenekleri verilmiştir.

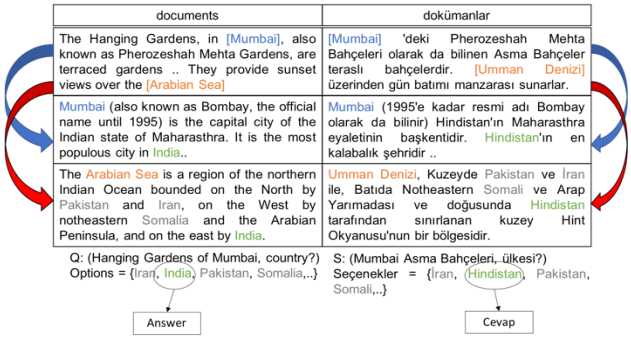
2.7 NewsQA (NewsQA)

CNN veri kümesinin 12.744 haber makalesini kullanarak oluşturulan 100.000 soru ve cevap çiftinden oluşmaktadır [23]. Bu veri kümesindeki soruların önemli bir kısmı basit kelime ve bağlam eşleşmesinin ötesinde bir muhakeme gerektirir. Bunun nedeni NewsQA veri kümesindeki soruların makalelerin ana metnine erişmeden sadece özetlere dayanarak formüle edilmesidir. Şekil 7'de NewsQA veri kümesinden örnek hikâye, soru ve cevabın metindeki karakter aralığı verilmiştir.

Story ID	**Öykü numarası**
./cnn/stories/b18f5c532055a819392300026 27746464d158602.story	./cnn/stories/b18f5c532055a819392300026 27746464d158602.story
Question	**Soru**
Was Justice of Peace Keith Bardwell breaking the law when he refused to issue a marriage license to a biracial couple?	Barış Adaleti Keith Bardwell çift ırklı bir çifte evlilik izni vermeyi reddettiğinde yasayı çiğnemiş miydi?
Answer char ranges	**Cevabın Metindeki Karakter Aralığı**
2692:2729	2692:2729

Şekil 7. NewsQA veri kümesinden örnek (Example from NewsQA dataset)

2.8 Wikihops (Wikihops)



Şekil 8. Wikihops veri kümesi örneği[24] (Example from Wikihop dataset [24])

Welbl ve arkadaşları, çoklu geçiş çıkarımına odaklanan iki yeni veri kümesi sunmuşlardır [24]. Bu veri kümelerinden biri olan Wikihop yaklaşık 51.000 örnek içermektedir. Şekil 8'de, hedefin Mumbai Asma Bahçe'lerinin hangi ülkede olduğunu belirlemek olan bir örnek verilmiştir. Metinde cevap açıkça belirtilmediği için cevabı bulmak için birinci dokümanda Asma Bahçelerinin Mumbai'de bulunduğunun daha sonra ikinci dokümandaki bilgilerle Mumbai'nin Hindistan'da bir şehir olduğunun çıkarımlarının yapılması gerekmektedir.

2.9 TriviaQA (TriviaQA)

TriviaQA veri kümesi 40.478 farklı cevap ve 95.956 soru-cevap ikilisinden oluşmaktadır [25]. TriviaQA veri kümesi birden fazla cümle üzerinde gerekçelendirilmesi ile cevaplanabilecek karmaşık soruları içermektedir. Veri kümesine ait örnek Şekil 9'da verilmiştir.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?	**Soru:** American Callan Pinckney'in adını taşıyan sistemi hangi türde en çok satan (1980'ler-2000'ler) kitap / video oldu?
Answer: Fitness	**Cevap:** Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became inter- national best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.	**Alıntı:** Callan Pinckney, Amerikalı bir fitness uzmanıydı. Callanetics egzersizleriyle benzeri görülmemiş bir başarı elde etti. 9 kitabının tümü uluslararası en çok satanlar listesine girdi ve ardından gelen video serisi 6 milyondan fazla kopya sattı. Pinckney'in ilk video yayını "Callanetics: 10 Saatte 10 Yıl Genç", ABD'deki diğer tüm fitness videolarının satışını geçti.

Şekil 9. TriviaQA veri kümesi örneği (Example from TriviaQA dataset)

2.10 WebQuestions (WebQuestions)

WebQuestions soruları Google Suggest API kullanılarak toplanan ve cevapları Amazon MTurk kullanılarak Freebase KB'den alınarak elde edilen 5.810 soru-cevap çiftinden oluşur [26]. Freebase KB delillerin ve cevapların kaynağı olarak kullanılmıştır. Şekil 10'da WebQuestions veri kümesine ait örnek metin, soru-cevap verilmiştir.

Article: Atom **Paragraph:** .. The atomic mass of these isotopos varied by integer amounts, called the whole number rule. The explanation for these isotopes awaited the discovery of the **neutron**, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932,.. **Q:** What part of the atom did Chadwick discover? **A:** neutron	**Makale:** Atom **Paragraf:** .. Bu izotopların atom kütlesi, tam sayı kuralı adı verilen tamsayı miktarlarına göre değişiyordu. Bu izotopların açıklaması, 1932'de fizikçi James Chadwick tarafından protona benzer bir kütleye sahip yüksüz bir parçacık olan **nötronun** keşfini bekliyordu. **S:** Chadwick atomun hangi bölümünü keşfetti? **C:** nötron

Şekil 10. WebQuestions veri kümesi örneği (Example from WebQuestions)

2.11 MSMarco (MSMarco)

MSMarco, 100.000 soru ve bunlara karşılık gelen cevapları içeren bir veri kümesidir [27]. Cevaplar Bing arama motoru kullanılarak gerçek web belgelerinden çıkarılmıştır. Sorguların cevapları kalabalık bir insan grubu tarafından üretilmiştir. Ayrıca, birden fazla cevap içeren sorular içermektedir.

Tablo 1'de ise QA veri kümelerindeki cevaplar ile ilgili bir karşılaştırılma verilmiştir. Cevabın metin parçacıklarından oluştuğu veri kümelerinin yanı sıra birden fazla cevaba sahip veri kümeleri belirlenmiştir. Ayrıca, cevabı olmayan soruları da içeren veri kümeleri belirtilmiştir.

3. DEĞERLENDİRME KRİTERLERİ (EVALUATION METRICS)

Bu bölümde, QA veri kümelerinin değerlendirilmesi için kullanılan başarı değerlendirme kriterleri olan hassasiyet, kesinlik, EM, F-Skoru, Bleu, Rouge kriterleri verilmiş olup

hangi veri türlerinde kullanıldığı belirtilmiştir (Tablo 2). Değerlendirme kriterlerinin seçimi büyük ölçüde görevin türüne bağlıdır.

Tablo 1. QA veri kümelerinin karşılaştırılması (Comparison of QA datasets)

Veri Kümesi	Cevap metinde geçer	Çoklu cevap içeren soru içerir	Cevabı olmayan soru içerir
SQuAD (Rajpurkar ve arkadaşları, 2016) [19]	+	+	+
NewsQA (Trischler ve arkadaşları, 2016) [23]	+	+	+
TriviaQA (Joshi ve arkadaşları, 2017) [25]	+	-	-
WebQuestions (Berant ve arkadaşları, 2013) [26]	-	+	+
MovieQA (Tapaswi ve arkadaşları, 2016) [22]	-	-	-
SimpleQuestions (Bordes ve arkadaşları, 2015) [28]	-	-	-
CoQA (Reddy ve arkadaşları, 2019) [21]	+	-	+
MSMarco (Nguyen ve arkadaşları, 2016) [27]	-	+	+

Çoktan seçmeli görevler doğru cevaplar veya sınıf etiketleri ile eşit olarak dağıtılıyorsa, tam eşleşme (EM) doğruluğu kullanır. Bu tür görevlerde kesinlik ve hassasiyet de kullanılabilir, fakat kesinlik ve hassasiyet değerleri ile hesaplanan F skoru son dönemlerde daha fazla kullanılmaktadır. Çoktan seçmeli, sınıflandırma görevleri bu kriterler ile değerlendirilebilir. Sentetik oluşturulmuş bAbi görevleri gibi hem tek bir doğru cevabın olduğu hem de cevapların tek kelime veya kelime listeleri olduğu veri kümeleri üzerinde EM kullanılarak başarı karşılaştırılmaları yapılmaktadır.

BLEU sistem tarafından oluşturulan özet kelimelerin (ve / veya n-gramların) insan tarafından çevrilmiş referans özetlerinde ne kadar göründüğünü ölçmeye yarayan bir değerlendirme kriteridir [29]. Genellikle metin özetleme ve makine çevirisinde kullanılan BLEU, cevabın birden fazla kelimeden oluştuğu QA veri kümeleri değerlendirilmesinde kullanılmaktadır.

Tablo 3'de verilen BLEU ölçütü hesaplamasında ilk olarak n-gram sayıları her bir cümle için hesaplanır. Daha sonra, tüm aday cümlelere kırpılmış n-gram sayıları eklenir. Kırpılmış n- gram sayıları, kelimenin herhangi bir referans özetinde gözlemlenen en büyük sayıyı geçmeyecek şekilde her kelimenin sayımı kesilerek belirlenir. Elde edilen toplam test derlemindeki aday n-gram sayısına bölünerek BLEU skoru hesaplanır.

Tablo 2. Değerlendirme kriterleri (Evaluation metrics)

Metrik	Açıklama	Amaç	Formül
Kesinlik [30]	Doğru sonuçların, bilginin tamamına oranı	Çoktan seçmeli, sınıflandırma, boşluk doldurma, cevabın metinde geçtiği sorular	$= \frac{\text{"İlgili Getirim"} \cup \text{"Bütün Veri Çıkarımı"}}{\text{"Bütün Veri Çıkarımı"}}$
Hassasiyet [30]	Getirilen doğru sonuçların, olması gereken doğru sonuçlara oranı	Çoktan seçmeli, sınıflandırma, boşluk doldurma, cevabın metinde geçtiği sorular	$= \frac{\text{"İlgili Getirim"} \cup \text{"Bütün Veri Çıkarımı"}}{\text{"İlgili Veri Çıkarımı"}}$
F Skoru	Kesinlik ve hassasiyetin harmonik ortalaması	Çoktan seçmeli, sınıflandırma, boşluk doldurma, cevabın metinde geçtiği sorular	$= 2x \frac{Kesinlik\ x\ Hassasiyet}{Kesinlik + Hassasiyet}$
EM	Doğru cevapların sayısı, cevabın metnin belli bir parçasında bulunması	Çoktan seçmeli, kısa cevaplı sorular	$= \frac{Doğru\ Cevaplandırılmış\ Soru\ Sayısı}{Toplam\ Soru\ Sayısı}$
Bleu [29]	Sistem tarafından oluşturulan özet kelimelerin (ve / veya n-gramların) insan tarafından çevrilmiş referans özetlerinde ne kadar göründüğünü ölçer.	Metin Özetleme, Makine Çevirisi, Cevabın metinde geçtiği sorular, cevabın birçok kelimeden oluşması	$p_n = \frac{\sum_{C \in \{Adaylar\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Adaylar\}} \sum_{n-gram' \in C} Count(n-gram')}$
Rouge [31]	İnsan tarafından çevrilmiş referans özetlerdeki kelimelerin (ve / veya n-gramların) sistem tarafından oluşturulan özetlerde ne kadar göründüğünü ölçer.	Metin Özetleme, Makine Çevirisi, Cevabın metinde geçtiği sorular, cevabın birçok kelimeden oluşması	$\text{ROUGE} - \text{N} = \frac{\sum_{S \in \{ReferansÖzetler\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferansÖzetler\}} \sum_{gram_n \in S} Count(gram_n)}$

ROUGE başarı değerlendirme ölçütü ise hem makine çevirisi hem de metin özetlemenin otomatik olarak değerlendirildiği n-gram kelime dizileri ile hesaplanmaktadır [31]. İnsan tarafından oluşturulan özet bilginin ile makine ile oluşturulmuş özetler arasında oluşan kesişimler kullanılmaktadır. Tablo 3'teki ROUGE formülünde yer alan n, $gram_n$'ın uzunluğunu ifade etmektedir. $Count\_match$ $(gram_n)$, aday özetteki ve referans özetteki kesişen maksimum n-gram sayısını ifade etmektedir. Eşitliğin paydası referans özette oluşan n-

gramların sayılarının toplamı olarak hesaplandığından ROUGE performans ölçütü, duyarlılık ile ilişkili bir ölçüttür [31].

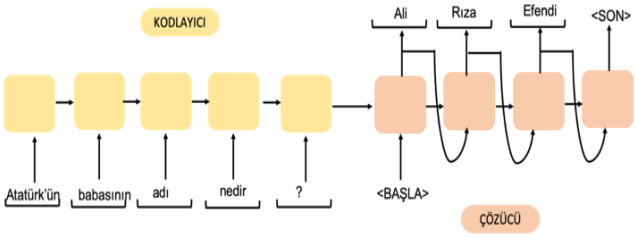
QA veri kümelerindeki başarı değerlendirme ölçütlerine yakından bakıldığında yaygın olarak SQuAD, NewsQA, TriviaQA, WebQuestions, Wikihop, bAbi Tasks veri kümelerinde EM ve F1 performans ölçütleri kullanılmaktadır. MSMarco veri kümesinde ise BLEU ve ROUGE performans ölçütleri kullanılmaktadır.

4. SORU CEVAPLAMADA KULLANILAN YÖNTEMLER (METHODS USED IN QUESTION ANSWERING)

Bu bölümde, QA sistem literatüründe yaygın kullanılan yöntemler Seq2Seq, Bellek Ağları, Çizge İşleyen Ağlar, Klasik Yöntemler olmak üzere 4 kategoride ele alınmıştır. Yöntemlerin QA sistemlerine nasıl uygulandığı özetlenmiş olup güncel QA sistem modelleri detaylı bir şekilde analiz edilmiştir.

4.1 Seq2Seq (Seq2Seq)

Seq2Seq temelli model bir girdi dizisini bir çıkış dizisine dönüştüren öğrenme modelidir. Bu bağlamda, dizi cümledeki kelimelere karşılık gelen semboller listesidir. Yaygın olarak makine çevirisi, diyalog sistemleri, soru cevaplama ve metin özetleme gibi alanlarda büyük başarı elde etmiştir [32]–[34]. Genellikle tekrarlayan bir derin öğrenme ağı ve bir dikkat bileşeni içermektedir. Uzun kısa süreli bellek (LSTM) ağları [35] ve Kapılı Tekrarlayan Ünite (GRU) [36] ağları sıralı girdiler problemleri için yaygın kullanılan yöntemlerdir.



Şekil 11. Seq2Seq modelin temel yapısı (Basic structure of Seq2Seq model)

Şekil 11'de bir kodlayıcı ve bir çözücüden oluşan Seq2Seq temelli QA sistem modeli verilmiştir. Kodlayıcı, girdi dizisini alıp LSTM veya GRU tekrarlayan ağlarını kullanarak eğitir. Buradaki girdi dizisi sorudaki kelimelerin vektörel karşılıkları ile elde edilir. Kodlayıcı, tekrar eden katmanının son durumunu bir başlangıç durumu olarak kod çözücünün ilk tekrarlayan katmanına gönderir. Kodlayıcı tarafından elde edilen vektör kod çözücünün doğru tahminler yapabilmesi için girdideki bütün bilgileri kapsar. Çözücüden elde edilen çıktı dizisi sorunun cevabının kelimelerini temsil etmektedir. Seq2Seq temelli modellerin özellikle uzun metinler için paralelleştirilemeyen doğası nedeniyle hem eğitim hem de çıkarım için genellikle yavaş olmaları karşılaşılan zayıflıkları olarak gösterilebilir.

Tan ve arkadaşları (2018), cevaptaki kelimelerin verilen metinde yer almasının gerekmediği bir çıkarım-sentezleme ağı olan S-Net ağını önermiştir [37]. Buradaki yaklaşım kanıt çıkarma ve cevap sentezleme olmak üzere iki aşamadan oluşmaktadır. Kanıt çıkarımı aşamasında soru metin ile eşleştirilip metinde soru ile ilgili en önemli alt alanlar kanıt olarak çıkarılır. Cevap sentezleme aşamasında ise kanıt çıkarımı aşamasında elde edilen kanıt parçacıklarına dayanarak cevabın oluşturulması amaçlanmaktadır. Burada, soru ve metin çıkarılan kanıt metninin başlangıç ve bitiş konumlarının özellik olarak etiketlendiği çift yönlü bir tekrarlayan ağ tarafından kodlanır. Cevabı oluşturmak için Rocktaschel ve arkadaşlarının (2016) önerdikleri dikkat mekanizmalı kod çözücü kullanılmıştır [38].

Hao ve arkadaşları (2017) çalışmalarında KB-QA sistemleri için uçtan uca bir sinir ağı modeli önermiştir [39]. İlk olarak soruları ve onlara karşılık gelen skorları farklı aday cevaplara göre dinamik olarak temsil etmek için çapraz dikkat mekanizmalı bir model kullanılmıştır. Böylece, her soru kelimesinde cevap üzerinde odaklanılır ve buna göre dikkat ağırlıkları belirlenir. Verilen metindeki kelimelerden önceki ve sonraki bilgiyi de kullanmak için kelime vektörleri üzerinde biLSTM (çift yönlü LSTM) kullanılmıştır.

Golub ve arkadaşları (2016) çalışmalarında LSTM ve CNN mimarilerini kullanan dikkat mekanizması tabanlı KB-QA için yaklaşım önermişlerdir [6]. Bu çalışmada ilk olarak soru dizileri üzerinde LSTM ve dikkat mekanizması uygulanarak bağlam vektörü elde edilmiştir. Bütün olası KB girişleri için soru, metin ve tahminlerde yalnızca karakter düzeyinde temsiller kullanılmıştır. Karakter düzeyinde temsil kullanılmasının sebebi daha önce eğitim sırasında görülmeyen yeni kelimelere daha iyi bir şekilde genelleme yapılabilmesidir. Daha sonra, elde edilen bu bağlam vektörü KB'ye girdi olarak verilerek cevap aranmaktadır. Bu çalışma ile standart kodlayıcı-kod çözücü çerçeveleri için zorlu bir görev olan kelime dağarcığımızda bulunmayan KB girişleri için olasılık puanlarını başarı ile elde edilmesi amaçlanmıştır.

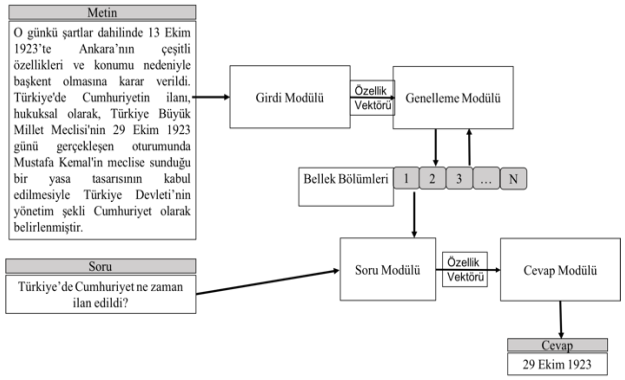
Seo ve arkadaşları (2016) çalışmalarında, paragrafın farklı ayrıntı düzeylerinde temsillerini modellemek için hiyerarşik bir mimari olan Çift Yönlü Dikkat Akışı (BIDAF) ağını tanıtmıştır [40]. Bu çalışmada, kelimeler arasındaki etkileşimi modellemek için biLSTM ağı kullanılmıştır. BIDAF karakter seviyesi, kelime seviyesi ve bağlama duyarlı kelime vektörlerini, sorguya duyarlı bir içerik temsili elde etmek için çift yönlü dikkat mekanizması kullanmıştır. Ayrıca, metindeki ve sorgu sözcüklerindeki bilgilerin birbirine bağlanması ve kaynaştırılmasından sorumlu olan dikkat akış katmanı bulunmaktadır. Dikkat akış katmanı, metindeki her bir paragrafı sabit uzunluklu bir vektör olarak oluşturmaz. Bunun yerine önceki katmanlardan elde edilen vektörlerle birlikte her bir dikkat vektörünün sonraki katmana

akmasına izin verir. Böylece erken özetlemenin neden olduğu bilgi kaybını azaltılması amaçlanmaktadır.

Xiong ve arkadaşları (2017) bir kodlayıcı ve dinamik bir kod çözücüden oluşan DCN (Ortak Dinamik Dikkat Ağları) ağını önermiştir [5]. DCN ilk önce sorunun ve belgenin ilgili kısımlarına odaklanmak için soru ve belgenin vektör temsillerini birleştirir. Ardından LSTM tabanlı sıralı model olan dinamik bir kod çözücü ile potansiyel cevap aralıkları üzerinde yinelenir. DCN cevabın başlangıç ve bitiş noktalarını her seferinde önceki tahminlerine göre şartlandırılmış olarak defalarca tahmin etme kabiliyetine sahip bir model olarak tasarlanmıştır.

Wang ve çalışma arkadaşları (2017) çalışmalarında R-NET ismini verdikleri QA sistemini tasarlamışlardır [41]. Metinden gelen bilgileri etkili bir şekilde kodlamaya yarayan her bir soruyu cevaplandırabilmek için metindeki kelimelerin farklı öneme sahip olduğunu dikkate alan tekrarlayan ağ tabanlı dikkat mekanizması önerilmiştir. Önerilen dikkat mekanizmasının çalışma prensibi tekrarlayan ağdaki hücrelere soruya uygunluklarına bağlı olarak önemsiz geçit parçalarını maskeleme ve önemli olanları vurgulama gibi farklı önem düzeyleri atamaktır. Cevabı bulmak için metindeki bütün bilgileri etkin bir şekilde toplayabilen bir eşleştirme mekanizması kullanılmıştır. Bu mekanizma ile olası bir cevap adayı metindeki bütün bilgileri kullanan dinamik bir yapı haline getirilmiştir.

4.2 Bellek Ağları (Memory Networks)



Şekil 12. Bellek ağlarının genel yapısı (Structure of memory networks)

Bellek Ağları (MN), bir bellek bileşeni ve çeşitli çıkarım bileşenlerine sahip modüler mimari temelli yapay sinir ağlarıdır [42]. MN veri etiketleme, metin sınıflandırma, metin özetleme gibi NLP'de sıkça karşılaşılan problemlerin yanı sıra birçok görüntü problemlerinde de kullanılmaktadır [4], [28], [42]–[44]. MN QA sistemleri üzerinde de başarılı performans gösterdiği için bu alanda yaygın olarak kullanılan bir model sınıfıdır.

MN'ler QA sistemleri için ele alındığında Şekil 12'de görüldüğü gibi giriş modülü, genelleme modülü, soru modülü ve cevap modülü olmak üzere 4 farklı bileşenden oluşmaktadır [1], [42]. Giriş modülünde kelime vektörleri kullanılarak her bir cümle için ayrı birer özellik vektörü oluşturulur ve metindeki her bir cümle vektör olarak ifade edilir. Genelleme modülünde giriş modülünde metinden elde edilen özellik vektörü ve mevcut bellek kullanılarak yeni girişi uygun bellek birimlerine kaydeder. Yeni bilgilere dayanarak daha önce kaydedilmiş herhangi bir bellek birimi değiştirilebilir. Bu modülde yapılan işlem depolanan bilginin yeni bilgi parçaları eklendiğinde genelleştirilmesi olarak düşünülebilir. Soru modülünde ise soru özellik vektörü elde edilir. Soru özellik vektörü ve genelleme vektöründen elde edilen mevcut bellek vektörü kullanılarak cevap için bir özellik vektörü oluşturur. En basit şekilde, tüm bellek birimleri üzerinde bir sıralama fonksiyonu kullanılarak soru ile bellek birimlerinin içeriği arasındaki eşleşmenin puanlanması yapılarak en yüksek puana sahip bellek birimi seçilir. Cevap modülünde ise elde edilen özellik vektörü kullanılarak sistem tarafından bir doğal dil ifadesi cevap olarak oluşturur.

Kumar ve arkadaşları (2016) çalışmalarında Dinamik Bellek Ağları (DMN) adını verdikleri bir mimari önermişlerdir [4]. DMN giriş modülü, soru modülü, bellek modülü ve cevap modülü olmak üzere 4 bileşenden oluşmaktadır. Giriş modülü, metindeki cümleleri "gerçekler" olarak adlandırılan bir dizi vektöre dönüştürür. Soru modülü, GRU tekrarlayan ağı kullanılarak sorunun bir vektör gösterimini hesaplamaktadır. Bellek modülü, soruyu cevaplamak için gerekli gerçekleri giriş modülünden alır. Ayrıca, bu modül soru ile metindeki ilgili cümleleri (gerçekleri) seçen bir dikkat mekanizması ve mevcut durumu ile elde edilen gerçekler arasındaki etkileşimlerden oluşan bir güncelleme mekanizması içermektedir. Cevap modülü ise bellek modülünden elde edilen vektörün son halini ve soru vektörünü kullanarak sorunun cevabını tahmin eder. Xiong ve arkadaşları (2016) çalışmalarında bu mimarinin giriş modülünde kullanılan GRU yerine biGRU kullanarak ve bellek modülünde kullanılan dikkat skorunun hesaplanmasında yaptıkları basit değişiklikler ile yeni bir mimari tasarlamış ve daha iyi performans elde etmişlerdir [43].

Chen ve arkadaşları (2019) çalışmalarında KB tabanlı QA sistemlerine BAMnet ismini verdikleri yeni bir Çift Yönlü Dikkat Mekanizması Tabanlı MN modeli önermişler [9]. Etkili bilgiyi çıkarabilmek için soru ve KB'ler arasında karşılıklı etkileşim yaklaşımı kullanılmıştır. Bu çalışmada önerilen çift yönlü dikkat mekanizmasının birincil dikkat ağı KB ışığında bir sorunun önemli kısımlarına ve sorunun ışığında önemli KB özelliklerine odaklanmayı amaçlamaktadır. İkincil dikkat ağı ise iki yönlü dikkat mekanizmasının kullanılmasıyla soruyu ve KB temsillerini geliştirmeyi amaçlamaktadır. Bu hiyerarşik çift yönlü dikkat mekanizması yaklaşımıyla soruyu cevaplayabilmek için metindeki en uygun bilgilerin çıkarılması amaçlanmaktadır.

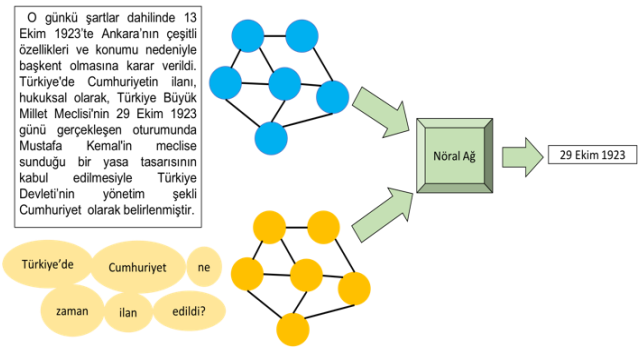
Ma ve arkadaşları (2017) çalışmalarında MN temelli bir QA sistemi tasarlamışlar [45]. QA sistemlerinde verilen sorular için çok kelimeli cevaplar üreten MN ve LSTM'nin etkileşimi ile etkili bir sinir ağı mimarisi önerilmiştir. Önerilen mimaride, Kumar ve arkadaşlarının (2016) [4]'deki önerdikleri DMN mimarisinden esinlenerek elde

ettikleri mimaride LSTM kullanılmıştır. Ayrıca, Facebook'un tekli cevap içeren bAbi görevleri veri kümesi tekil cevapların yerine çoklu cevaplar içeren kelimeler ile yer değiştirerek yeniden oluşturulmuş, deneyler çoklu cevap içeren veri kümesi üzerinde yapılmıştır.

Shen ve arkadaşları (2017) çalışmalarında ReasoNet adlı modüler yapıya sahip bir sinir ağı mimarisi önermiştir [46]. ReasoNets sorgular, belgeler ve cevaplar arasındaki ilişkiden etkili bir şekilde yararlanabilmek için tatmin edici bir cevap oluşturana kadar çoklu dönüşlerden yararlanarak dokümanın farklı kısımlarına odaklanır. Takviyeli öğrenme kullanan ReasoNets mevcut bilgilerin bir cevap üretmek için yeterli olduğu sonucuna vardığında okumayı sonlandırmayı dinamik olarak belirlemektedir.

Dong ve arkadaşları (2015) çalışmalarında soruları birçok açıdan analiz etmek için Çok Sütunlu Evrişimli Sinir Ağları (MCCNN) ismini verdikleri MN temelli yeni bir mimari tasarlamışlardır [47]. Bu çalışmada, giriş sorularından yanıt türlerini, ilişkileri ve bağlam bilgilerini çıkarmak için farklı sütun ağları kullanılır. MCCNN'ler üç farklı yönden cevap yolu, cevap içeriği ve cevap tipi olmak üzere soruları anlamak ve dağıtılmış temsillerini öğrenmek amacıyla oluşturulmuştur. Mimaride, bir sorunun dağıtılmış temsili ile cevabı arasındaki benzerliğin en üst düzeye çıkarılması amaçlanmaktadır. Freebase KB bilgi kaynağı olarak kullanılmış olup WebQuestions veri kümesi üzerinde deneyler yapılmıştır. Xu ve arkadaşları (2016) tarafından MCCNN'yi soru cevaplamak için ilk olarak kullanan [47]'daki Dong ve arkadaşlarının çalışmalarından esinlenerek ilişki çıkarımında sözdizimi ve cümlecik bilgilerinden yararlanmak için MCCNN'yi kullanmıştır [48]. Önerilen model iki ana adımdan oluşmaktadır. İlk adımda, bir nöral ağ kullanılarak bir KB (Freebase) aracılığıyla verilen bir sorunun olası cevapları çıkarılmaktadır. Daha sonra, yanlış cevapları bulmak ve doğru cevapları seçmek için yapılandırılmamış bir kaynak olan Wikipedia kullanılmaktadır. Bir arama motoru tarafından alınan belgeleri kullanarak KB-QA sistemlerini geliştiren bir uygulama olarak görülebilir.

4.3 Çizge İşleyen Ağlar (Graph Neural Networks)



Şekil 13. Çizge işleyen ağların yapısı (Structure of graph neural networks)

Son dönemlerde Çizge İşleyen Ağlar'ın birçok NLP problemi üzerinde uygulanmasının performansı arttırdığı görülmektedir [49], [50]. Şekil 13'de Çizge İşleyen Ağların QA sistemleri üzerinde uygulaması verilmiştir. İlk olarak metin bir çizgeye dönüştürülür. Metindeki bütün kelimeler çizgenin düğümleri olarak belirlenir. Her kenar metindeki bir kelimeden başlar ve bitişik kelimelerle devam eder. Sorunun kelimeleri de aralarındaki sözdizimsel bağımlılıklarını temsil eden bir çizge olarak oluşturulur. Oluşturulan çizgelerin tüm düğümleri birbirine bağlıdır ve her kenar metindeki kelimelerin arasındaki ilişkiyi ifade etmektedir. Bir sinir ağı bu temsiller üzerinde bir tahmin olarak uygun bir cevap üretmek için eğitilir. Her bir düğümde bir kelime vektörü ve kullanılan çizgenin türüne göre düğümler arasındaki her bir kenarın ağırlık veya kenar vektörü bulunabilmektedir.

Cao ve arkadaşlarının (2019) Çizge tabanlı Evrişimsel Ağ (GCN) modeli önerilmiştir [51]. GCN, komşu düğümlerine göre düğüm temsillerini güncelleyen algoritması ile farklı belgelerden kanıt toplayarak soruları yanıtlamayı öğrenir. Cao ve arkadaşları (2019) çalışmalarında ise çizge işleyen Çift Yönlü Dikkat Mekanizması Temelli Çizge Evrişimli Ağı (BAG) ismini verdikleri modeli önermişlerdir [52]. Önerdikleri model, [51]'deki GCN modelinin genişletilmişidir. Bu çalışmada öncelikle içerik olarak kullanılan dokümanlar bir çizgeye dönüştürülür. Bu çizgede düğümler varlıkları ifade ederken kenarlar düğümler arasındaki ilişkilerdir. Oluşturulan çizge daha sonra cevabın bulunması için çoklu muhakemeyi sağlayan GCN'e aktarılır. Ayrıca, bu çalışmada nihai tahmin için gerekli karşılıklı bilgiyi türetmek için çizge ile sorgu arasında çok seviyeli özelliklere sahip yeni bir çift yönlü dikkat mekanizması tasarlanmıştır. Bu dikkat mekanizması çizge ve sorgu arasındaki ortak bilginin oluşturulmasından sorumludur.

Sorokin ve arkadaşları (2018) çalışmalarında Çizge Sinir Ağları kullanılarak KB tabanlı bir QA sistem modeli tasarlamıştır [10]. Bu model komşu düğümlere ve ilişkilere dayalı çizge düğümlerinin gösterimlerini yinelemeli olarak güncelleyerek çizgeyi işler. Sonuçlar en iyi çizge kullanmayan modele karşı F1 skorunda %27.4 iyileşme olduğunu göstermektedir. Bu da çizge tabanlı modellerin QA sistemleri üzerindeki başarısını göstermektedir.

4.4 Klasik Yöntemler (Traditional Methods)

Klasik QA sistemlerinde kullanılan yaklaşımlardan biri dilbilimsel tekniklerdir. Bunlar sözcük türü (POS) etiketleme, varlık tanıma (NER) ve ayrıştırma gibi NLP tekniklerinden yararlanılmaktadır. Dilbilimsel teknikler kullanıcının sorusuna uygulanır, böylece soru belirli bir kaynaktan ilgili cevabı çıkarmak için bir belge veya bir veri tabanından uygun bir sorguya yeniden formüle edilebilir. Bir diğer yaklaşım olarak sistemin büyük miktarda veri kullanarak istatistiksel bir yolla "öğrendiği" ve tahminleri ve benzerlikleri hesaplamak için algoritmalar kullandığı istatistiksel yaklaşımdır. Bu yaklaşımda çeşitli problemler üzerinde uygulanabilen destek vektör makinesi (SVM) sınıflandırıcıları, Bayes ağı sınıflandırıcıları, Maksimum Entropi gibi makine öğrenmesi teknikleri kullanılmaktadır [53,54]. İstatistiksel yaklaşımların temel dezavantajlardan biri herhangi bir dilbilimsel özelliği tanımlamadan ve sorgu kelimeleri arasındaki herhangi bir dilbilimsel ilişkiyi

tespit etmeden, bir sorgunun her belirtecini bağımsız olarak ele almasıdır. Bu bölümde, bu yöntemler kullanılarak oluşturulan modeller analiz edilmiştir.

Klasik yöntemlere dayanan öncü çalışmalardan biri Ittycheriah ve arkadaşlarının (2001) geliştirdikleri IBM'in istatistiksel QA sistemidir [55]. Bu sistem çeşitli N-gram veya sözcük torbası özelliklerine dayanarak soru/cevap sınıflandırması için Maksimum Entropi modelini kullanmıştır. Green ve arkadaşları (1961) BASEBALL ismini verdikleri doğal dilde yöneltilen soruları yapısal bir veri tabanında sorgulayan ilk QA sistemlerinden biridir [56]. Önerilen sistem soruların belli bir sorgu formuna çevrilerek bilgi tabanlarından sorguların cevaplarının bulunması yaklaşımını kullanmıştır.

Banko ve arkadaşları (2002) çalışmalarında AskMSR ismini verdikleri manuel olarak tanımlanmış kurallara dayanarak kullanıcının sorusunu değiştirerek Google arama motoruna gönderen bir sistem tasarlamışlardır [57]. Sonuç kümesindeki tüm sayfa içeriğini aramak yerine Google tarafından oluşturulan sonuç özetlerinde 1 gram, 2 gram ve 3 gram kullanılmıştır. Bu n-gramlar, sıklıkları ve sorgunun ağırlığı ile ağırlıklandırılır. Bu yaklaşım AskMSR'nin sorguları hızlı işlemesini sağlamaktadır.

QA sistemlerinden arama motorlarını (Google, Yahoo, Altavista vb.) kullanan en eski çalışmalardan biri Zheng ve arkadaşlarının (2002) AnswerBus ismini verdikleri sistemdir [58]. AnswerBus, arama motorlarının her kelime için puanlandığı bir kelime öbeği stratejisini kullanmaktadır. Toplam puanlarına göre en iyi arama motoru soruyu cevaplamak için belirlenir. AnswerBus'taki bir eksiklik doğrudan bir cevap vermek yerine cevabı içerme olasılığı en yüksek web sayfalarını döndürmesidir. Dolayısıyla bu sistemde sorunun cevabını bulmak kullanıcıya bırakılmıştır. Bu bakımdan bir QA sisteminden ziyade belli bir sorgu için en iyi bir arama motorunu bulan bir sistem olarak ifade edilebilir. Stoyanchev ve arkadaşları (2008) çalışmalarında sorguları aramak için bileşenler olarak sorulardan otomatik olarak tanımlanan tam ifadeleri kullandılar. Cümlelerin ayıklanması işlemi adlandırılmış varlık (NER) tanıma, engellenecek kelime listeleri ve kelime öbeği (POS) etiketleyicileri kullanılmıştır [59].

Zhang ve arkadaşları (2010) çalışmalarında Çince için bir QA sistem modeli önermişlerdir [60]. Bu sistem genel bir soru analiz süreci, anahtar kelime çıkarma ve soru sınıflandırma süreçlerini içermektedir. Soru sınıflandırmasını uygulamak için kelimelerin özelliklerine POS (Sözcük Türü Etiketleme) ve SVM (Karar Destek Makinaları) sınıflandırıcısı kullanılmıştır. Cevap çıkarımı için kümeleme yapılmıştır. Moy'awiah ve arkadaşları (2019) ise çalışmalarında Arapça QA sistemleri için bilgilerini çeşitli alanlardaki dokümanlardan alan paragraf düzeyinde cevaplar sunan iki aşamalı bir sistem önermiştir [61]. Birinci aşamada SVM kullanarak sorgunun sınıfı bulunurken, ikinci aşamada ilgili belgeler içinden cevabın olduğu paragrafları bulmak için Gizli Anlamsal İndeksleme (LSI) kullanılmıştır. Deneyler toplam 10.000 doküman ve 10 sınıftan oluşan veri kümesi üzerinde uygulanmıştır.

İlhan ve arkadaşları (2014) tarafından Türkçe için oluşturdukları QA sisteminde metin madenciliğini kullanarak en uygun cevabı bulmayı amaçlamıştır [62]. Yine Türkçe QA sistemleri için Amasyalı ve arkadaşları (2005) tarafından yapılan çalışmada ise doğal dil ile yöneltilen sorudaki fiillerin eş anlamlılarının da kullanılmasıyla oluşturulan sorgular birleştirilmiş sorgular olarak arama motoruna verilir [63]. Arama motorundan elde edilen sonuç kümesi sorgulamada kullanılan anahtar kelimelerin terim sıklıklarının kullanıldığı puanlama yöntemi ile işlenmesiyle en yüksek puanı alan ilk beş cümle kullanıcıya cevap olarak verilmektedir.

5. YÖNTEMLERİN KARŞILAŞTIRMALI DEĞERLENDİRMESİ (COMPARATIVE EVALUATION OF METHODS)

QA sistemlerinde yeni teknolojilerin geliştirilmesi ile birlikte kullanılan yöntemler de değişmektedir. Literatürü incelediğimizde ilk olarak QA problemi için klasik yöntemler kullanılmıştır. Kelime torbası, POS, NER gibi klasik NLP tekniklerinin yanı sıra SVM, lojistik regresyon gibi makine öğrenmesi teknikleri de kullanılmıştır. Bununla birlikte kelime vektörlerinin ve sinir ağlarının popülerlik kazanmasıyla klasik yöntemlerden uzaklaşılmış derin öğrenme tabanlı mimarilere geçilmiştir. Derin öğrenme tabanlı mimariler klasik yöntemlere göre daha yüksek performans göstermektedir. Son dönemlerde ise çizge işleyen ağların ve klasik ve derin öğrenme tabanlı mimarilerin birlikte kullanılmaya başlandığı görülmektedir. Ayrıca, yine son yıllarda önceden eğitilmiş dil modellerinin QA sistemlerinde kullanıldığı ve performansı artırdığı görülmektedir.

Bu bölümde, Bölüm 4'te incelenmiş çalışmaların bir karşılaştırması sunulmuştur. Çalışmalar kullanılan yöntemlere, içerdikleri teknolojilere ve veri kümeleri üzerinde gösterdikleri performanslarına göre gruplanmıştır.

Tablo 3'de QA sistem modellerinin Bölüm 4'de yapılan sınıflandırması baz alınarak detaylı bir karşılaştırması yapılmıştır. Modeller oluşturulurken dikkat mekanizması, pekiştirmeli öğrenme, KB, harici bilgi ve çizge işleyen ağları kullanıp kullanmayan modeller gösterilmiştir. Ayrıca, bazı modeller soru ile ilgili olarak verilen doküman üzerinden birden fazla geçerek soruyu cevaplandırabildiği modeller tasarlanmıştır.

Tablo 3. QA sistem modellerinin karşılaştırılması (Comparison of QA system models)

Model	Metot	Teknoloji	Dikkat Mekanizması	Pekiştirmeli Öğrenme	Bilgi Tabanı	Harici Bilgi	Çizge İşleyen Ağlar	Çoklu Geçiş	Dil Modeli
R-NET (Wang ve arkadaşları, 2017) [41]	Seq2Seq	GRU	+	-	-	-	-	-	-
BIDAF (Seo ve arkadaşları, 2016) [40]	Seq2Seq	LSTM, CNN	+	-	-	-	-	-	-
BERT-BiDAF (Yang ve arkadaşları, 2019) [64]	Seq2Seq	BERT LSTM	+	-	-	-	-	-	+
Xlnet (Yang ve arkadaşları, 2019) [65]	Dil Modeli	Xlnet	+	-	-	-	-	-	+
Reasonet (Shen ve arkadaşları, 2017) [46]	Bellek Ağları	RNN, biGRU	+	+	-	-	-	+	-
SLQA (Wang ve arkadaşları, 2018) [66]	Seq2Seq	biLSTM,	+	-	-	-	-	+	-
DCN (Xiong ve arkadaşları, 2016) [5]	Seq2Seq	LSTM	+	-	-	-	-	-	-
KBQA (Cui ve arkadaşları, 2017) [67]	Klasik	KB	-	-	+	+	-	-	-
Xser (Xu ve arkadaşları, 2014) [68]	Klasik	POS, NER	-	-	+	-	-	-	-
BAMnet (Chen ve arkadaşları, 2019) [9]	Bellek Ağları	biLSTM	+	-	+	-	-	+	-
SAN (Liu ve arkadaşları, 2017) [69]	Bellek Ağları	biLSTM	+	+	-	-	-	-	-
S-net (Tan ve arkadaşları, 2017) [37]	Seq2Seq	GRU	+	-	-	-	-	-	-
Fastqa (Weissenborn ve arkadaşları, 2017) [70]	Seq2Seq	biLSTM	-	-	-	-	-	-	-
S-net (Tan ve arkadaşları, 2018) [71]	Seq2Seq	GRU	+	-	-	-	-	-	-
BAG (Cao ve arkadaşları, 2019) [52]	Seq2Seq; Klasik	biLSTM, NER, POS	+	-	-	-	+	+	-
EntityGCN (De ve arkadaşları, 2018) [51]	Seq2Seq	biLSTM, MLP	-	-	+	-	+	+	-
MHRC (Song ve arkadaşları, 2018) [72]	Seq2Seq	biLSTM	+	-	-	-	+	+	-
DMN (Kumar ve arkadaşları, 2016) [4]	Bellek Ağları	biGRU	+	-	-	-	-	-	-
(Golub ve arkadaşları, 2016) [6]	Seq2Seq	CNN, LSTM	+	-	+	-	-	-	-
AQAS (Moy ve arkadaşları, 2019) [61]	Klasik	SVM, LSI	-	-	-	-	-	-	-
GGNN (Sorokin ve arkadaşları, 2018) [10]	Çizge Ağları	Çizge Ağları	-	-	+	-	+	-	-

Tablo 4. Wikihops veri kümesinin karşılaştırmalı sonuçları (Comparative results on Wikihops dataset)

Model	Tekil	Kolektif öğrenme
BAG (Cao ve arkadaşları, 2019) [52]	%69	-
Entity-GCN (De ve arkadaşları, 2018) [51]	%67.6	%71.2
MHRC (Song ve arkadaşları, 2018) [72]	%65.4	-

Tablo 4'da Wikihop veri kümesi kullanılarak oluşturulan QA sistem modellerinin performans karşılaştırılması F1 başarı değerlendirme kriteri ile verilmiştir. Verilen bazı çalışmalarda tekil başarı oranının yanı sıra kolektif öğrenme uygulanan çalışmalar da bulunmaktadır. Tabloda kolektif öğrenme uygulanmayan çalışmalardaki başarı değerleri - ile gösterilmiştir. Cao ve arkadaşlarının (2019) [52]'deki ve De ve arkadaşlarının [51]'deki derin öğrenme ve klasik yöntemlerini bir arada kullanan çalışmalarının [72]'daki sadece derin öğrenme yöntemleri kullanılarak geliştirilen Seq2Seq temelli çalışmadan tekil başarı olarak daha yüksek performans gösterdiği görülmektedir. Ayrıca, [51]'deki çalışmada olduğu gibi kolektif öğrenme ile tekil öğrenmeden daha yüksek başarı değerlerine ulaşılabildiği görülmektedir.

Tablo 5'da MSMarco veri kümesi üzerinde çalıştırılan yöntemlerin karşılaştırılması Rouge ve Bleu başarı değerlendirme ölçütleri ile verilmiştir. Seq2Seq tabanlı modeller kullanılarak oluşturulan Reasonet [46], [70]'deki Fastqa çalışmalarında elde edilen başarı sonuçları MN temelli diğer çalışmalara göre daha düşük olduğu

görülmektedir. San ve arkadaşlarının (2017) [37]'da tasarladıkları mimarinin diğer tüm modellerden daha iyi performans göstererek %46.65 Rouge ve %44.78 Bleu başarı skorlarına ulaştığı görülmektedir.

Tablo 5. MSMarco Veri Kümesinin Karşılaştırmalı Sonuçları (Comparative Results on MSMarco Dataset)

Model	**Rouge**	**Bleu**
S-net (Tan ve arkadaşları, 2018) [37]	%46.65	%44.78
S-net (Tan ve arkadaşları, 2017) [59]	%46.65	%44.78
SAN (Liu ve arkadaşları, 2017) [69]	%46.14	%43.85
R-NET (Wang ve arkadaşları, 2017) [41]	%42.89	%42.22
Reasonet (Shen ve arkadaşları, 2017) [46]	%38.01	%38.62
Fastqa (Weissenborn ve arkadaşları, 2017) [70]	%33.67	%33.93

Tablo 6. SQuAD veri kümesinin karşılaştırmalı sonuçları (Comparative results on SQuAD dataset)

Model	Tekil	Kolektif Öğrenme
	EM/F1	*EM/F1*
Xlnet (Yang ve arkadaşları, 2019) [65]	%87.926 / %90.689	-
SLQA (Wang ve arkadaşları, 2018) [66]	%79.2 / %86.6	%82.4 / %88.6
SAN (Liu ve arkadaşları, 2017) [69]	%76.89 / %84.4	%79.61 / %86.5
BERT-BiDAF (Yang ve arkadaşları, 2019) [73]	%72.24 / %75.84	-
R-NET (Wang ve arkadaşları, 2017) [41]	%71.3 / %79.7	%75.9/ %82.9
FastQA (Weissenborn ve arkadaşları, 2017) [70]	%70.8 / %78.9	-
(Chen ve arkadaşları, 2017) [74]	%70.0 / %79.0	-
Reasonet (Shen ve arkadaşları, 2017) [46]	%69.1 / %78.9	%73.4/ %81.8
BIDAF (Seo ve arkadaşları, 2016) [40]	%68.0 / %77.3	%73.3 / %81.1
DCN (Xiong ve arkadaşları, 2016) [5]	%66.2 / %75.9	-/ %80.4
Multi-Perspective Matching (Wang ve arkadaşları, 2016) [75]	%65.5 / %75.1	-

Tablo 6'de SQuAD veri kümesi üzerinde uygulanmış çeşitli yöntemler ile tasarlanan QA sistem modellerinin başarı karşılaştırılması verilmiştir. [65], [66], [69]'deki çalışmaların Tablo 11'de verilen diğer çalışmalara göre daha iyi performans gösterdiği görülmektedir. [65]'deki Xlnet dil modeli %87.926 EM ve %90.629 F1 skoruna ulaşmışken, diğer mimarilerin kolektif öğrenme kullanılarak elde edilen başarı skorlarından daha yüksek başarı gösterdiği görülmektedir. Böylece, dil modeli ile eğitilen QA sistem modellerinde daha yüksek performans elde edildiği görülebilmektedir. [66]'deki çalışmada her soru için 15 deneme arasından en yüksek puana sahip cevabı seçilip elde edilen F1 skoru % 88.6 iken, [69]'deki SAN ağında farklı başlangıç değerleri verilerek oluşturulan 5 modelin olasılık dağılımının ortalaması kullanılarak %86.5 F1 skoru elde edilmiştir. Böylece, modellerin kolektif öğrenme ile eğitildiğinde tekil öğrenmeye göre daha yüksek başarı sonuçları elde ettikleri görülmektedir.

Tablo 7'de WebQuestions veri kümesinin çeşitli yöntemler uygulanarak tasarlanan QA sistem mimarilerinin başarı karşılaştırılması verilmiştir. Xu ve arkadaşları (2019) [76] tarafından önerilen hem KB hem de MN mimarisi kullanılarak oluşturulan modelin %54.60 F1 başarı skoru göstererek diğer tüm modellerden daha yüksek performans gösterdiği görülmektedir.

Tablo 7. WebQuestions veri kümesinin karşılaştırmalı sonuçları (Comparative results of WebQuestions dataset)

Model	F1
(Xu ve arkadaşları, 2019) [76]	%54.60
(Xu ve arkadaşları, 2016) [48]	%53.30
STAGG (Yih ve arkadaşları, 2015) [11]	%52.50
BAMnet (Chen ve arkadaşları, 2019)[9]	%51.80
QUINT (Abujabal ve arkadaşları, 2017) [77]	%51.00
(Bast ve arkadaşları, 2015) [69]	%49.40
(Yao ve arkadaşları, 2015) [67]	%44.30
MCCNNs (Dong ve arkadaşları, 2015) [47]	%40.80

Bu bölümde, QA sistemleri için önerilen mimarilerin karşılaştırılması yapılmıştır. Tablo 8'de Bölüm 4'teki yöntemlerin sınıflandırılmasına göre literatürdeki modeller gruplandırılmış olup uygulanan çeşitli tekniklere göre karşılaştırmaları yapılmıştır. Ayrıca, bu bölümde sırasıyla Wikihop, MSMarco, SQuAD ve WebQuestions veri kümeleri üzerinde uygulanan modellerin göstermiş oldukları EM, F1, ROUGE veya BLEU başarı ölçütleri listelenmiştir. Buna göre kolektif öğrenmenin genel olarak başarıyı artırdığı görülmektedir. Seq2seq temelli mimarilerin klasik yöntemlerle kullanılmasının QA sistemleri performansı üzerinde olumlu etkisi bulunmaktadır. Ayrıca, önceden eğitilmiş dil modellerinin kullanılmasının performans üzerinde olumlu etkisinin olduğu görülmektedir.

6. AÇIK ARAŞTIRMA ALANLARI (OPEN RESEARCH AREAS)

Bu bölümde QA sistemleri üzerindeki son gelişmeler ve eğilimler hakkında yardımcı olacak fikirler ve gelecekteki QA sistemleri için açık araştırma alanları belirlenmiştir.

6.1 Cevabın Liste Olması (List Type Answers)

Basit sorgular çoğu zaman basit modeller ile yanıtlanabilmektedir. Fakat, karmaşık sorgularda bilginin ortaya çıkarılması, bağlanması ve birleştirilmesi gerektiğinden basit sorgulara kıyasla cevabın bulunmasında zorluklarla karşılaşılabilmektedir. Örneğin, "Türkiye'nin başkenti neresidir?" sorusu alt sorgulara ayrılmaya ihtiyaç duyulmadan KB-QA sistemlerinde "başkent" etiket bilgisi ile cevaplandırılabilecek basit bir sorgudur. "Ankara'daki en büyük ilçeler hangileridir?" sorgusunda ise sorguyu alt sorgulara ayırmak gerekmekte

ve sorgunun cevabı birden fazla ilçeyi içerdiğinden bir *liste* olarak ifade edilmektedir. NLP'de cevabın bir liste olması problemi *liste-QA* olarak bilinmektedir. Bu tür sorgularda bütün olası farklı cevapları bulunması gerekmektedir. Bu sistemlerde cevap listesinin kaç elemandan oluşacağı her soru için farklılık göstermekle birlikte bu farklılık bu sorgu türlerinin zorluk nedenlerinden biri olarak görülmektedir. Bunun yanı sıra literatürdeki sistemlerde metinde geçen cevap adaylarından (kelimeler) frekansı düşük olanlar göz ardı edilmekte ve cevap listesinde bulunmamaktadır. Bu durumda sorgunun eksik veya yanlış cevap listesi oluşturulması söz konusu olmaktadır. Günümüzdeki *liste-QA* sistemlerinde büyük ölçekli veri üzerinden liste sorularına verilebilecek farklı cevapların tamamının bulunması hem performans hem de bellek açısından problem oluşturmakta ve çözülmeyi bekleyen problemlerden biri olarak görülmektedir.

6.2 Alt Sorgu İçeren Sorgular (Queries with subqueries)

Bazı sorguların cevaplandırılması karmaşık muhakeme gerektirmektedir. Günümüzde kullanılan birçok KB olmasına rağmen çoğu karmaşık muhakeme gerektiren problemlerin çözümü için doğrudan kullanılamaz. Karmaşık sorgular, ayrı ayrı ele alınan alt sorgular halinde ayrıştırılır. Bu ayrıştırmalar sorguyu hedef KB'ler üzerinde yürütmek için kullanılabilir. Örneğin, "Brad Pitt'in 2019 yılında oynadığı filmler hangileridir?" sorgusu "2019 yılındaki filmler" ve "Brad Pitt'in oynadığı filmler" alt sorgularına dönüştürülerek cevaplanabilir. Bu tür sorguların alt sorgular haline getirilip KB'ler üzerinde çalıştırılması ciddi bir hesaplama zamanı gerektirmektedir. Cevapların elde edilme hızı ve kalitesi üzerinde çalışılması QA sistemlerinin geleceği için önemli bir rol oynayacaktır. Ayrıca, KB'lerin bu tür sorgular için sorgulanmaya hazır hale getirilmesi, KB'de kullanılan bilgilerin vektörleştirilerek yeniden oluşturulması bir diğer araştırma yönü olarak gösterilebilir.

6.3 Eş Anlamlı İfadeler (Synonyms)

Doğal bir dilde aynı anlama gelen farklı kelimelerle bir metin çeşitli şekillerde ifade edilebilir. Günümüzdeki QA sistemlerinde popüler olarak kullanılan KB'ler belirli bir varlığa atıfta bulunabilecek tüm farklı terimleri içermemektedir. QA sistemlerinde herhangi bir sorguda kullanılan kelimeler KB'lerin etiketinde kullanılandan farklılık teşkil edebilir. QA sistemlerinde soruyu cevaplandırabilmek için bilgi kaynağı olarak kullanılan KB'lerde kullanılacak etiket bilgisinin tespitinde zorluklarla karşılaşılabilmektedir. QA sistemlerinde karşılaşılan bu problem literatürde sözcüksel boşluk olarak ifade edilir [78]. Bu sözcüksel boşluğu doldurmak QA sistemi tarafından doğru cevaplanabilecek soruların oranını önemli ölçüde artırmaktadır. Sözcüksel boşluğun tersi olan aynı kelimenin farklı anlamlara sahip olması durumu literatürde belirsizlik olarak adlandırılmaktadır [78]. Hem sözcüksel boşluk hem de belirsizlik problemleri QA sistemlerinin başarı oranlarını olumsuz etkilemektedir. Bu problemler göz önünde bulundurularak oluşturulacak KB'lerin QA sistemlerinin performansını olumlu yönde etkilemesi beklenmektedir. Sözcüksel boşluk problemi için bir kelimenin eş anlamlarının veya kelime öbeklerinin de oluşturulacak yeni KB'lerin etiketinde yer alması bir çözüm olarak gösterilebilir.

6.4 Harici Bilginin Kullanılması (Using External Knowledge)

Daha önce belirtildiği gibi karmaşık sorgular alt sorgulara ayrıldıktan sonra her bir alt sorgunun cevabının ayrı ayrı bulunup birleştirilmesi ile karmaşık sorular cevaplandırılabilir. Fakat, ayrıştırılan her bir alt sorgu aynı KB'lerin kullanılması ile cevaplanması mümkün olmayabilir. Muhtemel sebebi, bu KB'lerin bu alt görevleri çözmek için gerekli olan bilgiyi içermemesidir. Bu tür karmaşık sorgular üzerinde çoklu KB'lerin kullanımı çözüm olarak gösterilebilir. Fakat, birden fazla KB kullanmak yerine bu tür sorguları cevaplayabilecek büyük ölçekli KB'lerin oluşturulması QA sistemleri üzerinde hem performans hem de elde edilen cevabın kalitesi açısından başarılı sonuçlar verecektir. Çözüm olarak literatürde yeni yeni yer almaya başlayan özellikle çoktan seçmeli sorgular için kullanılan bilgi kaynaklarına ek olarak harici bilgilerinin (Wikipedia makaleleri, haberler vb.) kullanılması gösterilebilir. Harici bilgiler ile soruyu cevaplandırmak için eğitim aşamasında kullanılan bilgi zenginleştirilmiş olup daha başarılı sonuçlar vermesi beklenmektedir.

6.5 Çok Sayıda Bilgi Kaynağının Kullanılması (Using Multiple Knowledge Resources)

SQuAD veri kümesi gibi birçok QA veri kümesinde verilen tek bir paragraf ile yöneltilen sorulara cevap aranmaktadır. Verilen paragrafların uzunluğu veri kümelerine göre farklılık göstermekle birlikte metinler gerçek dünyada kişilerin herhangi bir soruyu cevaplarken kullandıkları bilgiye göre oldukça küçük ölçekli veriyi temsil etmektedir. Gerçek dünyada insanlar bir soruyu cevaplayabilmek için birden fazla doküman veya web sayfasından yararlanmaktadır. Bu da tek bir paragrafın verilmesiyle geliştirilen QA sistemlerinin gerçek dünyadaki kullanımını sınırlandırmaktadır.

Literatürde yeni yeni yer almaya başlayan birden fazla paragraf veya doküman içeren veri kümeleri tek bir paragrafın kullanıldığı veri kümelerine göre çok fazla gürültülü veri içermektedir. Bu tür veri kümeleri aynı bilgi ile ilgili çok fazla bilgi içerebildiğinden tek bir paragraf içeren veri kümelerine göre daha zor bir problemi çözmeyi amaçlamaktadır. Literatürdeki bu tür QA sistemlerinde soruyu cevaplandırabilmek için gereken bilgi doğru bir şekilde filtrelenemediğinden dolayı henüz yüksek başarı oranlarına ulaşılamamıştır. Metinden doğru cevabı çıkarabilmek için soru ile ilgili olmayan metin parçalarının çıkartılarak daha dar bir arama uzayı oluşturulmasıyla hem geliştirilecek QA sistemlerinin başarı performansı artırılacak hem de soruya cevap verme süresi hızlandırılacaktır.

6.6 Karşılıklı Konuşmaya Dayalı QA Sistemleri (Conversational QA Systems)

İnsanlar bir dizi birbirine bağlı soru ve cevaplar içeren konuşmalar yoluyla bilgi toplar. Aynı yaklaşım kullanılarak makinelerin bilgi toplaması sağlanabilir. Bu da makinelerin konuşma sorularını yanıtlayabilmesini sağlamak ile mümkündür. Son dönemlerde literatürde yer almaya başlayan karşılıklı konuşmaya dayalı sistemler popüler QA sistemleri araştırma alanlarından biridir. Karşılıklı konuşmaya dayalı QA sistemlerinde birbirini takip eden sorgulara cevap verebilmek için verilen bir metin ve oluşan diyalog bilgisine ihtiyaç duyulmaktadır. Literatürde QA sistemlerinde oldukça fazla kullanılan tekrarlayan sinir ağları geçmişteki bilgiyi kullanan derin öğrenme yöntemlerindendir. Fakat, insanlar karşılıklı konuşma yaparken geçmişteki bilgiyi farklı perspektiflerle değerlendirip karmaşık ve kapsamlı sorulara cevap verebilmektedir.

Geçmişteki diyalog bilgisini farklı yönlerden daha derin bir şekilde sentezleyebilen modellerin geliştirilmesi karşılıklı konuşma QA sistemlerinin başarısı üzerinde oldukça etkili olacaktır. Bunun yanı sıra, bir soruyu cevaplandırabilmek için çok uzun zaman önceki konuşmaların diyalog bilgisine ihtiyaç duyulabilir. Diyalog bilgisi henüz yöneltilmemiş soruların cevabını içerebildiğinden dolayı QA sistem modellerinin hangi verinin unutulup hangi verinin tutulması gerektiğine doğru bir şekilde karar vermesi ve bilgiyi doğru bir şekilde yönetmesi gerekmektedir. Bu tür problemlerde günümüzdeki QA sistemleri bellek limiti problemi ile karşı karşıya kalmaktadır. Gelecekteki araştırma yönü olarak soruyu cevaplandırmak için verilen bilginin belleğin tutamayacağı kadar büyük olması durumunda geliştirilebilecek çözüm yolları olabilir.

6.7 Yeterli Verinin Bulunmaması (The Lack of Data)

QA sistemleri üzerinde Türkçe için literatürdeki çalışmalar oldukça kısıtlıdır. Bunun nedenlerinden biri Türkçe'nin sondan eklemeli bir dil olması ve bunun sonucunda kelimelerin işlenmesinde zorluklar içermesidir. Bilgimiz dahilinde Türkçe QA sistemlerinde bilgi kaynağı olarak kullanılabilecek büyük ölçekli KB, çeşitli alanlarda olmak üzere QA veri kümeleri bulunmamaktadır. Eş anlamlı ve sözcük öbeklerinin etiket olarak kullanılabildiği KB'ler ve Türkçe için veri toplanmasında kullanılabilecek karşılıklı konuşma içeren veri kümeleri, karmaşık sorgular (alt sorgulara ayrılabilecek sorgular) vb. içeren veri kümelerinin oluşturulması durumunda Türkçe için QA alanında büyük ilerlemenin olacağı açıktır.

İnternet'teki bilgiler çeşitli doğal diller ile ifade edilmekte ve bazı doğal dillerde çeşitli alanlarla ile ilgili kaynak bulma problemi bulunmaktadır. Kullanıcıların farklı doğal dillerinin olması sebebiyle daha esnek bir yaklaşıma, bilgiyi kodlamak için kullanılan dilden farklı dilde olabilen, birden fazla giriş diline izin verebilen QA sistemlerine ihtiyaç duyulmaktadır. Literatürdeki çalışmalarda, sorgunun bölümlerinin veya tüm sorgunun otomatik çevirisi kullanılarak bu probleme çözüm aranmıştır. Ancak hatalı çevirilerden kaynaklı problemlerden dolayı bu şekilde geliştirilen sistemler düşük başarı göstermektedir. Bir doğal dilde ifade edilen metinle ilgili başka bir doğal dildeki sorguya cevap aranması problemi *çoklu dil destekleyen QA* olarak bilinmekte ve gelecekteki popüler QA araştırma konularından biri olarak gösterilmektedir. Bu tür QA sistemlerinde, çoklu dil destekleyen dil modellerinin kullanılarak QA modellerinin oluşturulması, farklı dillerdeki kelime vektörlerinin aynı uzayda oluşturulması, otomatik çeviride kullanılacak makine çeviri sistem kalitelerinin artırılması gibi çeşitli yöntemler uygulanarak farklı dildeki sorguları destekleyen modeller üzerindeki performanslar artırılabilir. Ayrıca, farklı doğal dillerdeki aynı anlamdaki kelimeler için tek bir kelime vektörünün kullanılması, farklı dil destekleyen KB'lerin oluşturulması gibi yöntemler uygulanarak farklı doğal dillerdeki sorgular üzerinde daha iyi sonuçlar elde edilebilir.

7. SONUÇ (CONCLUSION)

Son zamanlarda, derin öğrenme yöntemlerinin ve büyük ölçekli veri kümelerinin geliştirilmesiyle metin tabanlı QA sistem performanslarında büyük ölçüde iyileşme sağlanmıştır. Günümüzde, birçok QA veri kümesindeki insan performansı geliştirilen birçok son teknoloji modeller ile aşılmıştır. Bundan dolayı, son dönemlerde QA sistemleri araştırmacıların dikkatini çekmektedir. Özellikle büyük ölçekli QA veri kümeleri yüksek hesaplama kapasitesi gerektirdiğinden gerek donanım alanındaki gerekse de LSTM, GRU gibi derin öğrenme tabanlı teknikler, dikkat mekanizması, kolektif öğrenme, dil modelleri kural tabanlı ve makine öğrenmesi gibi klasik yaklaşımlara göre umut verici sonuçlar vermektedir.

Son dönemlerdeki QA sistemlerinde elde edilen yüksek performansların bir diğer sebebi olan BERT, XLnet, DistilBERT, Roberta, Metinden Metne Transfer dil modeli (T5) ve Reformer gibi önceden eğitilmiş dil modelleri sıfırdan model eğitimi yaklaşımı yerini ince ayarı model tasarım yaklaşımına bırakmıştır. Böylece, daha hızlı modeller eğitilebilmekte ve öğrenilmiş olan ilgili bir görevden QA sistemlerine öğrenmenin aktarılması yoluyla sistem performansında iyileşmeler sağlanmaktadır.

Literatürü incelediğimizde bahsedilen ilerlemenin çoğunun İngilizce veya İspanyolca gibi kaynak açısından zengin dillerde olduğu görülmektedir. Yeterli verinin bulunmadığı dillerde verinin sınırlı kullanılabilirliği nedeniyle diller arası veya çok dilli yaklaşımlar dile özgü modellere göre daha iyi bir yaklaşım olarak görülmektedir. Ayrıca, önceden eğitilmiş dil modellerinin düşük kaynaklı sistemler için performans üzerinde olumlu etkisi bulunmaktadır.

Bu çalışmada literatürdeki son yıllarda yer alan QA sistemleri veri kümeleri ve modelleri incelenmiş bu modellerin kullandıkları yöntemlere odaklanılmıştır. Bu sistemlerin arkasında kullanılan yöntemler dört kategoride ele alınmış QA sistemleri üzerinde nasıl uygulandığı özetlenmiştir. Ayrıca, bu çalışmada QA sistem modelleri üzerindeki en son gelişmeler ve eğilimler hakkında

yardımcı olacak fikirler ve gelecekteki QA sistemleri için açık araştırma alanları belirlenmiştir. Gelecekteki araştırma alanları olarak yeterli veriye sahip olmayan diller üzerindeki sistemler, birden fazla dil üzerinde çalışabilen sistemler, çok sayıda bilgi kaynağının kullanılmasının gerekli olduğu sistemler ve karşılıklı konuşmaya dayalı diyalog sistemleri öne çıkmaktadır.

KAYNAKLAR (REFERENCES)

[1] D. Kapashi, P. Shah, "Answering Reading Comprehension Using Memory Networks", **Stanford Deep Learn. NLP Course,** 2015.

[2] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation", **Conference on Empirical Methods in Natural Language Processing, Katar**, 1532-1543, 2014.

[3] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching Word Vectors with Subword Information", *Trans. Assoc. Comput. Linguist.*, *5*, 135-146, 2017.

[4] A. Kumar *et al.*, "Ask me anything: Dynamic memory networks for natural language processing", **33rd International Conference on Machine Learning,** NewYork, A.B.D., 1378-1387, 2016.

[5] C. Xiong, V. Zhong, R. Socher, "Dynamic coattention networks for question answering", **5th International Conference on Learning Representations**, Toulon, Fransa, Nisan 2017.

[6] D. Golub and X. He, "Character-level question answering with attention", **EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing**, Texas, A.B.D. 2016.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge", **Proceedings of the ACM SIGMOD International Conference on Management of Data, A.B.D,** 1247-1250, 2008.

[8] F. M. Suchanek, G. Kasneci, G. Weikum, "Yago: A core of semantic knowledge", **16th International World Wide Web Conference**, Alberta, Kanada, 697-706, 2007.

[9] Y. Chen, L. Wu, M. J. Zaki, "Bidirectional attentive memory networks for question answering over knowledge bases", **2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Minneapolis, A.B.D, 2019.

[10] D. Sorokin and I. Gurevych, "Modeling Semantics with Gated Graph Neural Networks for Knowledge Base Question Answering", *arXiv preprint arXiv:1808.04126*, 2018.

[11] W. T. Yih, M. W. Chang, X. He, J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base", **53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing,** 1, 1321-1331, 2015.

[12] L. Su, T. He, Z. Fan, Y. Zhang, M. Guizani, "Answer Acquisition for Knowledge Base Question Answering Systems Based on Dynamic Memory Network", *IEEE Access*, *7*, 161329-161339, 2019.

[13] M. Wasim, D. Waqar, D. Usman, "A Survey of Datasets for Biomedical Question Answering Systems", *Int. J. Adv. Comput. Sci. Appl.*, *8*(7), 484-488, 2017.

[14] O. Kolomiyets and M. F. Moens, "A survey on question answering technology from an information retrieval perspective", *Inf. Sci. (Ny).*, *181*(24), 5412-5434, 2011.

[15] D. Diefenbach, V. Lopez, K. Singh, P. Maret, "Core techniques of question answering systems over knowledge bases: a survey", *Knowl. Inf. Syst.*, *55*(3), 529-569, 2018.

[16] J. Weston *et al.*, "Towards AI-complete question answering: A set of prerequisite toy tasks", **4th International Conference on Learning Representations, ICLR**, San Juan, Puerto Rico, 2016.

[17] G. Yiğit, M. F. Amasyalı, "Ask me: A Question Answering System via Dynamic Memory Networks," **2019 Innovations in Intelligent Systems and Applications Conference, ASYU**, İzmir, Türkiye, 1-5, 2019.

[18] Z. Yang *et al.*, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering", **2018 Conference on Empirical Methods in Natural Language Processing**, Brüksel, Belçika, 2369-2380, 2018.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, "SQuad: 100,000+ questions for machine comprehension of text", **EMNLP Conference on Empirical Methods in Natural Language Processing**, Texas, A.B.D, 2383-2392, 2016.

[20] P. Rajpurkar, R. Jia, P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD", *arXiv preprint arXiv:1806.03822*, 2018.

[21] S. Reddy, D. Chen, C. D. Manning, "CoQA: A Conversational Question Answering Challenge", *Trans. Assoc. Comput. Linguist*, 249-266, 2019.

[22] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, S. Fidler, "MovieQA: Understanding stories in movies through question-answering", **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, A.B.D**, 4631-4640, 2016.

[23] A. Trischler *et al.*, "NewsQA: A Machine Comprehension Dataset", *ACL*, 191-200, 2017.

[24] J. Welbl, P. Stenetorp, S. Riedel, "Constructing Datasets for Multi-hop Reading Comprehension Across Documents", *Trans. Assoc. Comput. Linguist.*, *6*, 287-302, 2018.

[25] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension", **ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)**, Vancouver, Canada, 1, 1601-1611,2017.

[26] J. Berant, A. Chou, R. Frostig, P. Liang, "Semantic parsing on freebase from question-answer pairs", **EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference,** Washington, A.B.D, 1533-1544, 2013.

[27] T. Nguyen *et al.*, "MS MARCO: A human generated MAchine reading COmprehension dataset", *CEUR Workshop Proceedings*, 2640-2660, 2016.

[28] A. Bordes, N. Usunier, S. Chopra, J. Weston, "Large-scale Simple Question Answering with Memory Networks", *arXiv:1506.02075*, 2015.

[29] K. Papineni, S. Roukos, T. Ward, W. Zhu, "BLEU : a Method for Automatic Evaluation of Machine Translation", *Comput. Linguist.*, 311-318, 2002.

[30] R. D. Banker and S. M. Datar, "Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation", *J. Account. Res.*, 27(1), 21-39,1989.

[31] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries", *Proc. Work. text Summ. branches out (WAS 2004)*, 74-81, 2004.

[32] Q. Xiao, X. Chang, X. Zhang, X. Liu, "Multi-Information Spatial-Temporal LSTM Fusion Continuous Sign Language Neural Machine Translation", *IEEE Access*, 8, 216718-28,2020.

[33] J. Cheng, F. Zhang, X. Guo, "A Syntax-Augmented and Headline-Aware Neural Text Summarization Method", *IEEE Access*, 2020.

[34] K. Palasundram, N. Mohd Sharef, K. A. Kasmiran, A. Azman, "Enhancements to the Sequence-to-Sequence-Based Natural Answer Generation Models", *IEEE Access*, 8:218360-71, 2020.

[35] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Comput.*, 9(8), 1735-80, 1997.

[36] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,", arXiv:1412.3555, 2014.

[37] C. Tan, F. Wei, N. Yang, B. Du, W. Lv, M. Zhou, "S-Net: From answer extraction to answer synthesis for machine reading comprehension", **32nd AAAI Conference on Artificial Intelligence, AAAI,** Louisiana, A.B.D, 32(1), Şubat, 2018.

[38] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, P. Blunsom, "Reasoning about entailment with neural attention", **4th International Conference on Learning Representations, ICLR,** San Juan, Puerto Rico, 2016.

[39] Y. Hao *et al.*, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge", **ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers),** Vancouver, Kanada, 1, 221-231, 2017.

[40] M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, "Bidirectional Attention Flow for Machine Comprehension", arXiv:1611.01603, 2016.

[41] W. Wang, N. Yang, F. Wei, B. Chang, M. Zhou, "Gated self-matching networks for reading comprehension and question answering", **ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers),** Vancouver, Kanada, 1, 189-198, 2017.

[42] J. Weston, S. Chopra, A. Bordes, "Memory networks", **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings,** San Diego, A.B.D, 2015.

[43] C. Xiong, S. Merity, R. Socher, "Dynamic memory networks for visual and textual question answering", **33rd International Conference on Machine Learning, ICML,** NY, A.B.D., 2397-2406, 2016.

[44] T. Yu, J. Yu, Z. Yu, Q. Huang, Q. Tian, "Long-Term Video Question Answering via Multimodal Hierarchical Memory Attentive Networks", *IEEE Trans. Circuits Syst. Video Technol.*, 2021.

[45] F. Ma *et al.*, "Long-term memory networks for question answering", in *CEUR Workshop Proceedings*, 1986, 7-14, 2017.

[46] Y. Shen, P. Sen Huang, J. Gao, W. Chen, "ReasoNet: Learning to stop reading in machine comprehension", **in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,** Anchorage, A.B.D, 1047-1055, 2017.

[47] L. Dong, F. Wei, M. Zhou, K. Xu, "Question answering over freebase with multi-column convolutional neural networks", **ACL-IJCNLP 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing,** 1, 260-269, 2015.

[48] K. Xu, S. Reddy, Y. Feng, S. Huang, D. Zhao, "Question answering on freebase via relation extraction and textual evidence," **54th Annual Meeting of the Association for Computational Linguistics,** Berlin, Almanya, 1, 2326-2336, 2016.

[49] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, "The graph neural network model", *IEEE Trans. Neural Networks*, 20(1), 61-80,2009.

[50] Y. Xiao, G. Zhou, "Syntactic edge-enhanced graph convolutional networks for aspect-level sentiment classification with interactive attention", *IEEE Access*, 157068-157080, 2020.

[51] N. de Cao, W. Aziz, I. Titov, "Question answering by reasoning across documents with graph convolutional networks", **in NAACL HLT Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference,** Minneapolis, 1, 2306-2317, 2019.

[52] Y. Cao, M. Fang, D. Tao, "BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering", **in NAACL HLT Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,** Minneapolis, 1, 357-362, 2019.

[53] E. KARTAL *et al.*, "Bir Öğrenciyi Üstün Zekâlı ve Yetenekli Olarak Aday Göstermek İçin Doğru Soruları Sormak: Bir Makine Öğrenmesi Yaklaşımı", *Bilişim Teknol. Derg.*, 13(4), 2020.

[54] A. Özgür and H. Erdem, "Saldırı Tespit Sistemlerinde Kullanılan Kolay Erişilen Makine Öğrenme Algoritmalarının Karşılaştırılması", *Bilişim Teknol. Derg.*, 5(2), 41-48, 2012.

[55] A. Ittycheriah, M. Franz, W.-J. Zhu, A. Ratnaparkhi, R. J. Mammone, "IBM's Statistical Question Answering System," in *Proceedings of TREC-9 Conference*, 2000.

[56] B. F. Green, A. K. Wolf, C. Chomsky, K. Laughery, "Baseball: An automatic question-answerer", **in Proceedings of the Western Joint Computer Conference: Extending Man's Intellect, IRE-AIEE-ACM 1961,** 219-224, 1961.

[57] M. Banko, E. Brill, S. Dumais, J. Lin, M. Way, "AskMSR: Question answering using the worldwide Web", **Proc. AAAI Spring Symp. Min. Answers,** 7-9, 2002.

[58] Z. Zheng, "AnswerBus question answering system", **InHuman Language Technology Conference (HLT)**, 27, 2002.

[59] S. Stoyanchev, Y. Song, W. Lahti, "Exact phrases in information retrieval for question answering", 9–16, 2008.

[60] K. Zhang and J. Zhao, "A Chinese question-answering system with question classification and answer clustering", **7th International Conference on Fuzzy Systems and Knowledge Discovery**, A.B.D., 6, 2692-2696, 2010.

[61] M. Al-Shenak, K. M. O. Nahar, K. M. H. Halawani, "Aqas: Arabic question answering system based on svm, svd, and lsi", *J. Theor. Appl. Inf. Technol.*, 97(2), 681-91, 2019.

[62] S. Ilhan, N. Duru, Ş. Karagöz, M. Sağır, "Metin Madenciliği ile Soru Cevaplama Sistemi", ***Elektron. ve Bİlgisayar Mühendisliği Sempozyumu***, Bursa, 26-30, 2008.

[63] M. F. Amasyalı and B. Diri, "Bir Soru Cevaplama Sistemi: BayBilmiş", *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 1(8), 2016.

[64] L. Yang and L. Song, "Contextual Aware Joint Probability Model Towards Question Answering System", *arXiv:1904.08109*, 2019.

[65] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding", *Advances in Neural Information Processing Systems*, 2019.

[66] W. Wang, M. Yan, C. Wu, "Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering", **56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)**, Melbourne, Avustralya, 1, 2018.

[67] W. Cui, Y. Xiao, H. Wang, Y. Song, S. W. Hwang, W. Wang, "KBQA: Learning question answering over QA corpora and knowledge bases", *VLDB Endowment*, *10*(5), 565-576, 2016.

[68] K. Xu, S. Zhang, Y. Feng, D. Zhao, "Answering Natural Language Questions via Phrasal Semantic Parsing", in *Natural Language Processing and Chinese Computing*, Berlin, Heidelberg, 333–344, 2014.

[69] X. Liu, Y. Shen, K. Duh, J. Gao, "Stochastic answer networks for machine reading comprehension", **56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)**, 1, 1694-1704, 2018.

[70] D. Weissenborn, G. Wiese, L. Seiffe, "FastQA: A Simple and Efficient Neural Architecture for Question Answering", arXiv:1703.04816, 2017.

[71] C. Tan, F. Wei, N. Yang, B. Du, W. Lv, M. Zhou, "S-Net: From Answer Extraction to Answer Generation for Machine Reading Comprehension", **32nd AAAI Conf. Artif. Intell. AAAI**, Louisiana, A.B.D, Şubat 2017.

[72] L. Song, Z. Wang, M. Yu, Y. Zhang, R. Florian, D. Gildea, "Exploring Graph-structured Passage Representation for Multi-hop Reading Comprehension with Graph Neural Networks", *arXiv:1809.02040*, 2018.

[73] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", **NAACL HLT Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,** Minneapolis, 2019.

[74] D. Chen, A. Fisch, J. Weston, A. Bordes, "Reading Wikipedia to answer open-domain questions", **55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers),** Vancouver, Kanada, 1, 1870-1879, 2017.

[75] Z. Wang, H. Mi, W. Hamza, R. Florian, "Multi-Perspective Context Matching for Machine Comprehension", *arXiv: 1612.04211*, 2016.

[76] K. Xu, Y. Lai, Y. Feng, Z. Wang, "Enhancing key-value memory neural networks for knowledge based question answering", **NAACL Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,** Minneapolis, 1, 2937-2947, 2019.

[77] A. Abujabal, M. Riedewald, M. Yahya, G. Weikum, "Automated template generation for question answering over knowledge graphs", **26th International World Wide Web Conference,** İsviçre, 1191-1200, 2017.

[78] S. Hakimov, C. Unger, S. Walter, P. Cimiano, "Applying semantic parsing to question answering over linked data: Addressing the lexical gap", **International Conference on Applications of Natural Language to Information Systems**, 103-109**,** 2015.