Türkçe Bilgi Kaynaklı Soru Yanıtlama Dizgesi

Knowladge-Based Question and Answering System for Turkish

Pınar YAŞAR
Gtech
pinar.yasar@gtech.com.tr
ORCID: 0000-0003-3977-5834

İrem ŞAHİN
Gtech
irem.sahin@gtech.com.tr
ORCID: 0000-0003-1525-0734

Eşref ADALI
İTÜ Bilgisayar ve Bilişim Fak.
adali@itu.edu.tr
ORCID: 0000-0002-1561-8255

Öz

Soru yanıtlama dizgeleri üzerinde sürdürülen çalışmalar ya bilgi tabanlı ya da genel ağdaki (Internet) verileri kullanarak çalışmaktadır. Bu makalede tanıtılan çalışma bilgi tabanlıdır ve Türkçeye özgüdür. Altay dil ailesinin üyesi olan Türkçe sondan eklemeli bir dildir ve dil bilgisi kuralları bükünlü Hint-Avrupa dillerinin dil bilgisi kurallarından farklıdır. Bu nedenle, Türkçe yanıtlama dizgesi Türkçenin dil bilgisi kurallarına uygun olarak hazırlanmıştır. Dizge sırasıyla şu aşamalardan oluşmaktadır: Ön işleme (tümcelere ayırma, sözcüklere ayırma, hecelere ayırma, ses bilimi açısından sınama (ünlü, ünsüz uyumu, ünlü düşmesi, baş ve sondaki harf kuralları, galatlar (mumpsimus) ve yabancı sözcükler); Biçim bilimsel çözümleme; Sözcüklerin sınıflarının belirlenmesi; Söz öbeklerinin bulunması; Sözcüklerin rollerinin belirlenmesi; Tümcenin anlamlandırılması; Uygun yanıtın üretilmesi. Üretilen her yanıtın doğruluğu soru soran kişiye onaylatılmaktadır. Sorulan sorular ve bunlara ilişkin yanıtlar bir koşut derlem (parallel corpus) içinde tutulmaktadır. Dizge belli bir süre çalıştırıldıktan sonra söz konusu koşut derlem yeterli boyuta ulaşmaktadır. Bu aşamadan sonra, sorulan bir sorunun çözümlenmesine geçmek yerine koşut derlemde aranmakta ve aynı ya da benzer yanıt bulunmaya çalışılmaktadır. Benzer tümcelerin bulunması için vektör yaklaşımı kullanılmıştır. Türkçe için geliştirilen çözüm yöntemi diğer eklemeli diller için kullanılabilir niteliktedir. Koşut derlemi kendisinin üretmesi geliştirilen yöntemin önemli bir özelliğidir.

***Anahtar Sözcükler**: Bilgi tabanlı, Soru Yanıtlama Dizgesi, Biçim Bilimsel Çözümleme, Sonlu Durum Makinesi.*

Gönderme ve kabul tarihi: 05.10.2019 - 12.12.2019

Makale türü: Araştırma

Abstract

Question and answering system use either own knowledge base or data collected from Internet. In this paper a knowledge-based question and answering system for Turkish is presented. Turkish is member of Altay language family and an agglutinative language. The grammar of Turkish different then India-European languages. Therefore, this project had been developed according to Turkish grammar. System consists of several steps. These steps are: preprocessing (tokenizing of sentences, parsing a word into syllable, phonological testing (consonant softening, vowel reduction, haplology, head and tail letters), morphological testing, mumpsimus, foreign words); morphological parsing; finding of roles of words; interpretation of question and answering. In the first phase the QA system can understand question and answers. Later on, constitutes question and answer parallel corpus in other word data bases. The second phase consist of the same pre-processing then searches for the same or similar question in question database if finds the answer gets from answer database. In order to fine similar question similarity vectors approaches were used.

***Keywords**: Knowledge-Based, Question Answering, Morphological Parsing, Finite State Machine*

1. Giriş

Kamu ve özel kuruluşlar ister hizmet versinler ister ürün satsınlar üye veya müşterilerini memnun etmek için 2000 yıllarından başlayarak çağrı merkezleri oluşturmuşlardır. Kuruluşun büyüklüğüne bağlı olarak çağrı merkezinde çalışan personelin sayısı binleri geçmiştir. Günde üç vardiya çalışan bu personellerin çalışma ortamlarının oldukça sıkıcı olduğu bilinmektedir. Maliyeti düşürmek amacıyla

çağrı merkezleri ücretlerin daha düşük olduğu taşra kentlerinde hatta yabancı ülkelerde kurulmuştur. Çağrı merkezleri, günümüzde yoğunlukla e-ticaret, bankalar, oteller ve kamu kuruluşları tarafından kullanılmaktadır.

Çağrı merkezindeki insanın yerini alacak bir Soru Yanıtlama Dizgesinin (SYD) insanların çalıştığı çağrı merkezinin maliyetini çok büyük ölçüde düşüreceği açıktır. 2019 yılı için en düşük ücretin 2558,- TL olduğu ve bir vardiya süresinde bir personelin ortalama 120 isteği yanıtlayabileceği varsayımı ile bir soru yanıtlama hizmetin bedelinin yaklaşık olarak 1 TL olacağı hesaplanabilir. Bu hizmeti vermek üzere sağlanan ortam ve diğer giderler (mekan, iletişim alt yapısı, iklimlendirme, aydınlatma vb.) eklendiğinde bir hizmet bedelinin yaklaşık olarak 5,- TL'yi bulacağını söyleyebiliriz. Doğal Dil İşleme (DDİ) destekli bir SYD ile aynı hizmetin verilmesinin maliyeti genel ağ üzerinden verilen bankacılık hizmetinin maliyetine yaklaşık olarak eşit olacaktır ve bunun değeri 40 kuruş dolayındadır. Sonuç olarak hizmet maliyetinin SYD ile 12 de 1 oranında azalacağı kolayca söylenebilir.

Çağrı merkezlerinin maliyetini azaltabilmek amacıyla robot yanıtlama dizgelerinin geliştirilmesi gündeme gelmiştir. SYD olarak adlandırılan bu çalışmaların çekirdeğini DDİ oluşturmaktadır. SYD doğal olarak uygulandığı dilin özelliklerine göre tasarlanmalı ve geliştirilmelidir. Bir dil ailesi için geliştirilen bir altyapının aynı dil ailesinin üyesi olan diğer dillere uygulanabilmesi ya da uyarlanabilmesi kolay olabilmektedir ancak farklı bir dil ailesinin üyesi bir dile uygulanması kolay değildir. Bu değerlendirmenin sonucu olarak, İngilizce için geliştirilmiş olan bir SYD Türkçe SYD için uygulanabilir değildir. Bu yönde yapılan uygulamalar, gülünç sonuçlar verebilmektedir.

İngilizce için çok sayıda uygulama geliştirilmiş ve geliştirilmektedir. Ancak bu çözümler Türkçe için uygulanabilir değildir. Bunun nedeni Türkçenin eklemeli bir dil olmasıdır. Türkçe sözcükler ortalama olarak, sözcüğün anlamını değiştiren 2,85 yapım eki alabilmektedirler. Çizelge-1'de iki dil ailesi arasındaki biçim bilimsel farklılık gösterilmiştir.

Çizelge-1: Kullanılan ve Aranan Sözcükler arasındaki Farklar

İngilizce sözcük		Türkçe sözcük	
kullanılan	aranan	kullanılan	aranan
of colour	renk	rengindeki	renk
on your site	site	sitenizdeki	site
look for	look	aramaktayım	ara-

Çizelge-1'den anlaşılacağı gibi Türkçe bir sözcüğün anlamının çıkarılabilmesi için biçim bilimsel çözümlemeden geçirilmesi gerekmektedir. Bu proje kapsamında Türkçeye özgü bir SYD'nin geliştirilmesi amaçlanmıştır.

Bir bilgisayarın, sorulan bir soruyu anlaması ve buna karşılık üretmesi ile ilgili ilk örnek 1964 yılında Weizenbaum tarafından geliştirilen ELISA dizgesidir [1]. ELISA dizgesi ziyaretçinin adı öğrenerek karşılıklı yazışmaya başlamaktadır. Ziyaretçiden öğrendiği bilgileri kullanarak soru ya da yanıt tümceleri oluşturarak yazışmayı sürdürmektedir. Bu örnek Yapay Zeka (YZ) kavramına dayalı ilk örnek olarak anılmaktadır.

YZ temelli soru yanıtlama dizgeleri iki sınıfa ayrılmaktadır: Bilgi Çekme temelli BÇ-SYD ve Bilgi Kaynaklı BK-SYD [2], [3].

1.1 BÇ-SYD

BG-SYD'ler sorunun yanıtını Genelağ'da bulunan bilgileri tarayarak bulmaya çalışırlar. BÇ-SYD'nin her tür soruyu yanıtlaması beklenmez; konu sınırlamasına gidilir örneğin sanat, spor, sağlık, gezi, ev, bahçe, hayvanlar. Bilgi çekmeye dayalı dizgeler ana hatları ile aşağıda sıralanan işlemleri yerine getiriler:

- Genelağ'daki sayfaları indirme,
- Sayfa içindeki bağlantıları öğrenme ve bunların içeriklerini de indirme,
- İndirilen sayfaya ilişkin anahtar sözcükleri bulma,
- Anahtar sözcük ve sayfa bilgisini dizine yerleştirme,
- Sözcük ve sayfa bilgisini dizinleme,

Bu adımların sonunda elde edilen dizinler daha sonra arama motorları tarafından kullanılmaktadır. Genelağ'dan bilgiyi getirmek üzere kurulan yapı, ağ gezgini (crawler), dizinleyici (indexing) ve arayıcı (search engine) adı verilen üç ana bileşenden oluşur. Ağ gezgini Genelağ üzerinde sürekli olarak gezinerek web sayfalarını bulmaya ve bu sayfalardaki bilgileri toplamaya çalışır. Ağ gezgininden beklenen üç özellik sırasıyla aşağıda açıklanmıştır:

- **Kapsama alanı**: Genelağ'da bulunan web sayfalarından ne kadarını toplayabildiğinin ölçüsüdür.
- **Güncellik**: Topladığı web sayfalarının ne kadar güncel olduğunun ölçüsüdür.

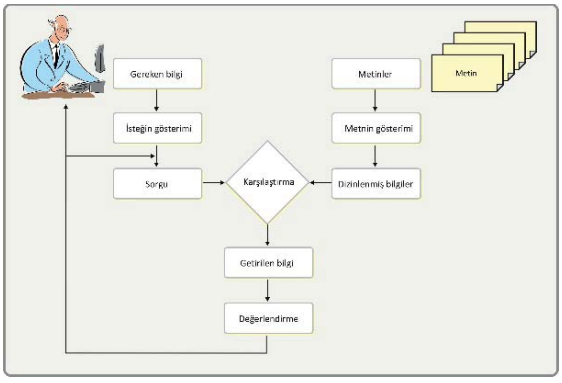
- **Değerlilik**: Bir sayfanın izlenme sıklığı o sayfanın içeriğinin değerinin bire bir karşılığı olmasa da sayfanın değeri hakkında bilgi vermektedir.

Ağ gezgininin indirdiği sayfaların içinde geçen sözcüklerden ilgili sayfayı niteleyecek anahtar sözcüklerin bulunması gerekir. Bunun için sayfa ve sayfanın bağlantıları içinde sıkça geçen sözcüklerden yararlanılarak anahtar sözcükler bulunmaya çalışılır. Bulunan anahtar sözcükler dizinleyici tarafından sözcük dizinine yazılır. Türkçe sözcükler çok sayıda yapım ve çekim eki alabilirler. Bu nedenle web sayfasında sıkça geçen sözcükler belirlendikten sonra bu sözcüklerin kök veya gövdelerinin bulunmasına çalışılır. Örneğin:

- özelliğindeki → özellik
- niteliğindeki → nitelik
- kurallarından → kural
- arananlar → ara
- çalışmaların → çalışma
- anlamlandırma → anlam

İkinci aşamada ilgili anahtar sözcüklerin geçtiği sayfanın nerede yüklü olduğu bilgisi (URL) sayfa adresleri dizinine yazılır.

Erişmek istenen bilgiye erişmeyi sorgu sözcükleri veya sorgu söz öbeklerini sağlarlar. Arayıcı (arama motoru) sorgu sözcük veya söz öbekleri ile sözcük dizininde arama yapar. Arama sonucunda ortaya çıkan sayfaları, izlenme sıklık sırasında araştırmacıya sunar. Sorgu söz öbeklerinde kullanılan mantık işleçleri VE, VEYA, DEĞİL, İZLENEN, YAKIN ve ? imidir. Şekil-1'de Genelağ üzerinden bilginin nasıl getirildiği gösterilmiştir.



Şekil-1: Genelağ üzerinden bilgi getirme süreci

1.2 BK-SYD

Sorulan bir sorunun yanıtı, soruyu yanıtlayacak olanın sahibi olduğu kaynaklarda bulunabilir. Bu tür uygulamalar doğal olarak sorulan sorunun yanıtını kendi kaynaklarını kullanarak yanıtlamaya çalışırlar. BK-SYD'lere aşağıdaki uygulamalar örnek olarak verilebilirler:

- Hava durumu ile ilgili soruları yanıtlama,
- Spor karşılaşmalarının sonuçlarını söyleme,
- Otelde konaklama isteyenlere yardımcı olur.
- Uçak bileti alacaklara uçuş bilgilerini sunma,
- e-ticarette müşterilerin isteklerini öğrenip en uygun ürünü sunma.

Çizelge-2'de bilgi kaynaklı soru yanıtlama dizgelerinde karşılaşılabilecek olası sorulardan örnekler verilmiştir.

Çizelge-2: Bilgi Kaynaklı Soru Yanıtlama Dizgeleri için Olası Soru Örnekleri

Soru	Yanıt
Türkiye Azerbaycan maçının sonucu ne?	2 -1
23 Nisan günü otelinizde boş oda var mı?	Evet var
Yarın hava nasıl olacak?	Bulutlu, yağmurlu ve 19 derece
Tek taşlı yüzük bakıyorum. Sizde var mı?	Üzgünüz, tek taşlı yüzük bizde yok
Alaska'ya uçuşunuz var mı?	Alaska'ya uçuşumuz yok.
Kuğugölüne iki bilet istiyorum 19 Mayıs akşamı Ön sıralardan lütfen	Hangi gösteri Yer seçiminiz? Üçüncü sıradan veriyorum.

Bir bilgi kaynaklı soru yanıtlama dizgesini oluşturabilmek için birbirinin koşutu olan soru ve yanıt veri tabanlarının oluşturulması gerekir. İlgili konuda sorulabilecek sorular ve bunlara karşılık düşecek yanıtları hazırlamak oldukça emek gerektiren bir çalışmadır. Konu kapsamı çok dar olan soru yanıtlama dizgeleri için olası tüm soruları ve karşılıklarını içeren koşut veri tabanı hazırlanabilir.

Bu makalede sunulan dizge BK-SYD'dir ve proje hakkında ayrıntılı bilgiler Bölüm 3'te verilecektir.

2. Yakın Çalışmalar

Eş zamanlı bir SYD Ojokad ve ark. tarafından önerilmiştir [4]. Örgüt temelli soru yanıtlama dizgesi olarak adlandırılan bu çalışma, "soruyu işleme" ve "yanıtlama süreci" gibi iki kısımdan oluşmaktadır. Sorulan sorunun kalıbı için bir kısıtlama kuralları getirilmemiştir. Yanıtlama süreci için öncelikle sorunun sözcükleri sınıflanarak etiketlenmiştir. Soru konuları veri tabanındaki ilgili veri takımlarından elde edilmekte ve anahtar sözcükler ile karşılaştırılmaktadır.

Benzerlik için Levenshtein yöntemi kullanılmaktadır. Söz konusu proje İngilizce için geliştirilmiştir.

Ghobadi ve ark. tarafından tanıtılan soru yanıtlama dizgesi DDİ üzerinde kurulmuştur [5]. Projenin amacı bir e-ticaret sitesinde müşteri isteklerini yanıtlamaktır. Sorulan bir soru öncelikle anahtar sözcüklere ayrıştırılmakta ve tanımlamalar ortaya çıkarılmaktadır. Yanıt tümce hazırlanırken bu tanımlamalardan yararlanılmaktadır.

Bir diğer çalışma Azevedo ve ark. tarafından gerçekleştirilmiştir [6]. Bu çalışmada önce soru anlaşılmaya çalışılmakta, daha sonra bu soruya en uygun yanıt bulunmaktadır. Bu çalışma uygulama alanına özgü olarak hazırlanmıştır. Çalışmanın aşamaları şöyle verilmektedir: DDİ yöntemleri kullanılarak sorunun anlamlandırılması; Yanıt için gerekli bilgilerin elde edilmesi; Yanıt için bilgi kaynağının kullanılması.

Cui ve ark. şablon temelli bir çalışma gerçekleştirmişlerdir [7]. Bu çalışmada sorular bir şablon üzerinde temsil edilmektedir. Bilgi kaynağı, soru ve yanıt derlemleri kullanılmaktadır. Şablona uygun olarak dizge soruları yanıtlamaktadır.

Paul ve ark. basit İngilizce ile yazılmış ve içinde gerçek bilgiler bulunana belgeler için bir model oluşturmuşlardır [8]. Kısa hikayelerden oluşan bu belgelerin yanıtları kısa tümcelerdir. Modeli oluşturmak üzere altı görev tanımı yapılmıştır ve bu görevler şöyle adlandırılmıştır: adılı çözme, bağımlılığı ortaya çıkarma, çizgede sözcük köküne erişme ve çizge varlıkları arasındaki ilişkiler, çizge oluşturma, soru çizgesinin oluşturulması ve çizge üzerinde yanıtın araştırılması.

3. Gerçekleştirilen Proje

Bu makalede tanıtılan proje bir BK-SYD projesidir ve Türkçe için geliştirilmiştir. BK-SYD daha önce açıklandığı gibi, sorulan sorunun yanıtını oluştururken kuruluşa özgü verilerden yararlanır. Projenin başarımını ölçebilmek adına sanay bir e-ticaret sitesi oluşturulmuştur. Bu site giyim ile ilgili ürünleri pazarlayacak biçimde kurgulanmıştır. Giyim ile ilgili ürünler pazarlayan bir e-ticaret sitesinin soru yanıtlama konusunda vereceği hizmetler şunlardır:

- Ürün araması
- Ödeme
- İade veya değiştirme
- Gönderme
- Şikayet

Bu işlemlerden en karmaşık ve zor olanı ürün aramasıdır. DDİ açısından en ilginç olanı da bu konudur. Ürün arama ile ilgili olası sorular ve yanıtları Çizelge-3'te verilmiştir:

Çizelge-3: Olası soru ve yanıtlar

Soru	sıra	Yanıt
Mavi renkli, kısa kollu, erkek spor gömleğiniz var mı?	1	Evet, aradığınız özellikte gömleğimiz var.
	2	Kısa kollu erkek spor gömleğimiz var.
	3	Erkek spor gömleğimiz var.
	4	Spor gömleğimiz var.
	5	Gömleğimiz var.
	6	Üzgünüz aradığınız özellikte bir ürünümüz yok

Müşterinin soruduğu sorunun yanıtı hazırlanırken doğal olarak kuruluşun ürün veri tabanına bakılmaktadır. Bu durumu göstermek üzere Çizelge-3'te altı farklı yanıt örneği verilmiştir. Aranan ürüne tam uyan bir ürün olduğunda ilk yanıt, tam uyum olmadığında diğer yanıtlar verilecektir. Gerçekleştirilen proje aşağıdaki aşamalardan oluşmaktadır:

- Ön işleme
- Biçim bilimsel çözümleme
- Sözcüklerin rollerinin belirlenmesi
- Sorunun anlamlandırılması,
  - Yanıtın oluşturulması
  - Paralel derlemin oluşturulması

3.1 Ön İşleme

Bir soruyu yanıtlamadan önce, sorunun ayıklanması ve yazım yanlışlarından arındırılması gerekir. Bu amaçla aşağıda sıralanan işlemler gerçekleştirilmiştir:

- Tümceleri ayırma,
- Yazım yanlışlarını düzeltme,
- Türkçe abecede bulunan özel harflerin eksikliğini giderme,
- Kısaltmaların açılması,
- Galatların düzeltilmesi.

Bazı sorular birden fazla tümceden oluşabilir. Müşterilere, sorularını tek tümce biçiminde sormaları istenemez. Birden çok tümceden oluşan sorularda, sorgu özelliği olan tümce bulunmaya çalışılmıştır.

3.1.1 Ayrıştırma

İleride yapılacak çalışmalar için tümce sözcüklere ayrılmıştır. Bu amaçla Pyton dili ile birlikte sağlanan NLTK aracı kullanılmıştır. Anılan programın tümceleri ve sözcükleri ayırmadaki başarımı %99 olarak ölçülmüştür. Şekil-2'de tümce ve sözcüklerin parçalanmasına ilişkin sonuçlar gösterilmiştir.

```
In [3]: from nltk.tokenize import word_tokenize, sent_tokenize
text1 = """Algı ve ifadelerimiz arasındaki bağlantıların neler olduğu bulunabilir.
Bu programı ne olduğunu farkedip, istediğimiz şekilde değiştirdiğimizde,
istediklerimize düşündüğümüden daha kısa zamanda ulaşabiliriz.
Bu değişimin algılarda, kullandığımız dil ve davranışlarda bütünlük içinde değiştirilmesi gerekmektedir.
Yaşadığımız tecrübelerden ortaya çıkan stratejiler farkında olmadan kullanıldığı için üst yapıda yapılan bir değişim,
değişimi ortaya çıkartmayacaktır. """
stoken = sent_tokenize(text1)
for i in stoken:
    tokens=word_tokenize(i,'turkish')
    print(tokens)

['"""', 'Algı', 've', 'ifadelerimiz', 'arasındaki', 'bağlantıların', 'neler', 'olduğu', 'bulunabilir', '.']
['Bu', 'programı', 'ne', 'olduğunu', 'farkedip', ',', 'istediğimiz', 'şekilde', 'değiştirdiğimizde', ',', 'istediklerimize',
'düşündüğümüden', 'daha', 'kısa', 'zamanda', 'ulaşabiliriz', '.']
['Bu', 'değişimin', 'algılarda', ',', 'kullandığımız', 'dil', 've', 'davranışlarda', 'bütünlük', 'içinde', 'değiştirilmesi',
'gerekmektedir', '.']
['Yaşadığımız', 'tecrübelerden', 'ortaya', 'çıkan', 'stratejiler', 'farkında', 'olmadan', 'kullanıldığı', 'için', 'üst', 'yapı
da', 'yapılan', 'bir', 'değişim', ',', 'değişimi', 'ortaya', 'çıkartmayacaktır', '.']
```

Şekil-2: Tümceleri ve sözcükleri parçalama programının sonuçları

3.1.2 Yazım Yanlışlarını Düzeltme

Türkçenin fonetik bir dil olması nedeniyle bir sözcüğün nasıl yazıldığının bilinmesine gerek yoktur. İngilizcenin önemli sorunlarından biri olan bu konuda, Türkçe sorular için yapılması gereken çok fazla işlem yoktur ancak bazı müşterilerin sözcükleri yanlış yazabilecekleri varsayılabilir. Örneğin, bir ya da birkaç harfi eksik ya da fazla ya da yanlış yazabilir. Sıkça karşılaşılan bir sorun harflerin sırasını şaşırmadır. Yazım yanlışları için şu örnekler verilebilir: *kelbekler, kellebekler, kelabekler, kelebelker.*

Yazım yanlışlarını düzeltmek üzere bir algoritma geliştirilmiştir. Bu algoritmanın ana hatları ve aşamaları şöyledir:

1. ***Sözcüğü hecelerine ayırma***: Sözcük Türkçenin kurallarına uygun olarak hecelere ayrılır.
2. ***Ses bilimi sınaması:*** Sözcüğün Türkçenin ses bilimi kurallarına uyup uymadığı sınanır. Bu kurallar sırasıyla şunlardır.
   - Ünsüz yumuşaması
   - Ünsüz benzeşmesi
   - Ses düşmesi
   - Sözcük başı ve sonunda bulunabilecek harfler

   Yukarıda belirtilen her ses olayı için gerekli olan program yazılmıştır. Bu sınamadan geçemeyen bir sözcüğün doğru yazılmadığına karar verilir.

   Türkçeye özgü ç, ı, ğ, ö, ü gibi harfleri barındırmayan tuş takımlarını kullananlar, Türkçe sözcükleri yanlış yazmaktadırlar. Örneğin *kirmizi, yesil, gordum*,

   Türkçe harfleri içermeyen sözcüklerin yazım yanlışlarını bulmak için sık kullanılan sözcükler sözcesinden yararlanılmıştır. Sözcüğün doğrusunu bulmak üzere Levenshtein benzerlik yönteminden yararlanılmıştır.

   Günümüzde, özellikle gençler mesaj yazarken kısaltma kullanmayı yeğlemektedirler. *Örneğin mrb, slm, tşk.*

   Kısaltmaları açabilmek amacıyla bir kısaltmalar çizelgesi hazırlanmıştır. Bir kısaltmanın karşılığı bu çizelgeden bulunarak açık biçime dönüştürülmüştür.

   Bir başka durum özen göstermeden yanlış kullanılan sözcüklerdir. Özensiz yazım biçimleri yaygın olarak kullanılmaktadır ve insanlar bunların yazılışını doğru saymaktadırlar. Örneğin *gelcem, gitcem.*

   Özensiz yazılan sözcükleri düzeltebilmek amacıyla 4.110 sözcükten oluşan bir özensiz sözcükler çizelgesi hazırlanmıştır. Bu çizelgede sözcüğün yanlış ve doğru biçimi yer almaktadır.

   Yabancı sözcükler de soru tümcesinde yer alabilmektedir. Özellikle yabancı marka yazımları böyle olmaktadır. Yabancı sözcükler için hazırlanan çizelgede 220 sözcüğe yer verilmiştir.

   Ses bilgisi ile çözülen tüm durumlar için gerekli olan algoritmalar ve bunları gerçekleştirecek programlar proje kapsamında tasarlanmış ve hazırlanmıştır.

3. ***Sözcüğün sınıfının belirlenmesi:*** Bir sözcüğün sınıfını belirlemek üzere izlenen yöntem ana hatları ile şöyledir: Türkçede yüklemin tümce sonunda bulunduğu varsayılarak, tümce sonundaki sözcüğün yüklem olup olmadığına, daha önce hazırlanmış olan eylem sözcesinde araştırılarak karar verilmiştir. Yüklem belirlendikten sonra tümce içindeki sözcüklerin sınıfları bulunmaya çalışılmıştır. Bunun için Coşkun'un çalışmasında açıklanan kurallardan yararlanılmıştır [9].
4. ***Biçim bilimsel sınama:*** Yazım yanlışını bulmak için kullanılan son aşama sözcüğün biçim bilimsel çözümleyiciden geçirilmesidir. Çözümleme sonunda erişilen kök sözcük ad ve eylem sözcelerinde araştırılır. Eğer bu sözcelerde bulunuyor ise sözcüğün doğru

yazıldığı sonucuna varılır. Örneğin çözüm sonucu şöyle ise:

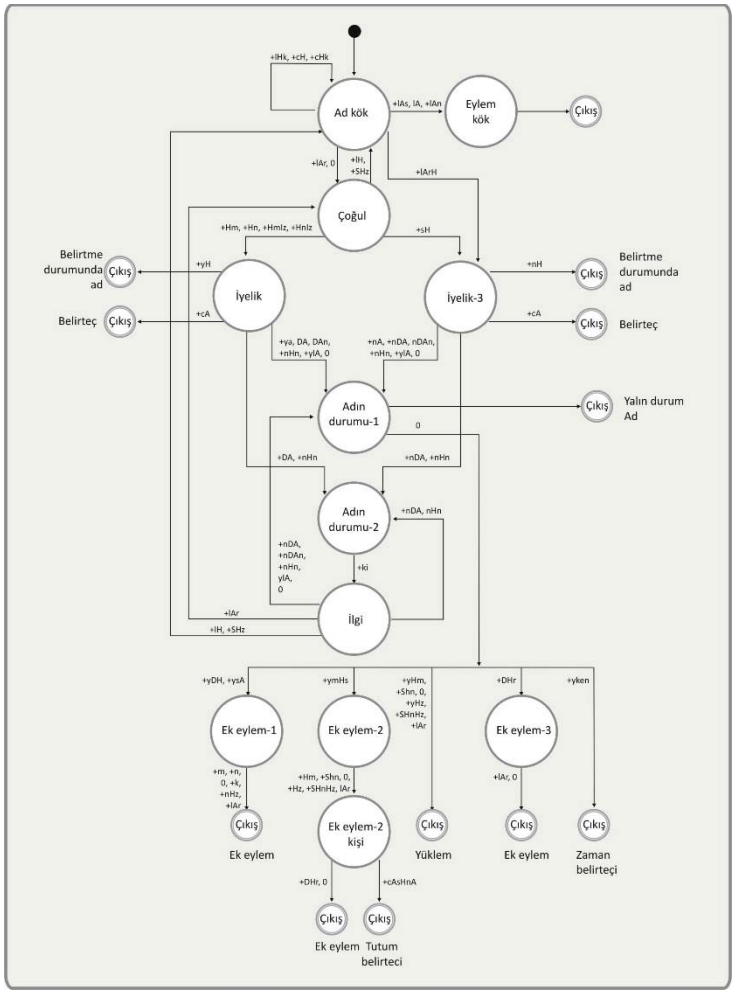
Kelebeği →kelebek+i

sözcüğün doğru yazıldığına aşağıdaki gibi ise yanlış yazıldığına karar verilir.

Kelbeği →kel+beği

Biçim bilimsel çözümleyicinin ürettiği çözümlerdeki belirsizlikleri gidermek için kural temelli yöntemler kullanılmıştır. Bu yöntemlerin gereksinimi olan sözceler hazırlanmıştır. Aynı sözceler sözcük sınıflarının belirlenmesinde de kullanılmıştır. Hazırlanan sözceler aşağıda sıralanmıştır. Sözcelerin yanında verilen sayılar barındırdıkları sözcüklerin sayısını göstermektedir.

- Ad (12.636)
- Eylem (462)
- Önad (11.311)
- Belirteç (2.316)
- Kısaltmalar (105)
- Sık Kullanılan sözcükler (44.960)
- Ünlem (189)
- Bağlaç (55)
- Adıl (99)
- İlgeç (33)



Şekil-3: Türkçe adların biçim bilimsel çözümlemesi için Sonlu Durum Makinesi

3.2 Biçim Bilimsel Çözümleme

Bir sözcüğün köküne ulaşmak için biçim bilimsel çözümleyiciden geçirilmesi gerekmektedir. Bu konuda yapılmış çalışmalar bulunmaktadır [10], [11]. Söz konusu çalışmalar yaklaşık %97 başarıma sahiptir ve olası tüm olasılıkları ortaya çıkarmaktadırlar. Projenin özelliği gereği ilk üç çözüm yeterli görülmüştür.

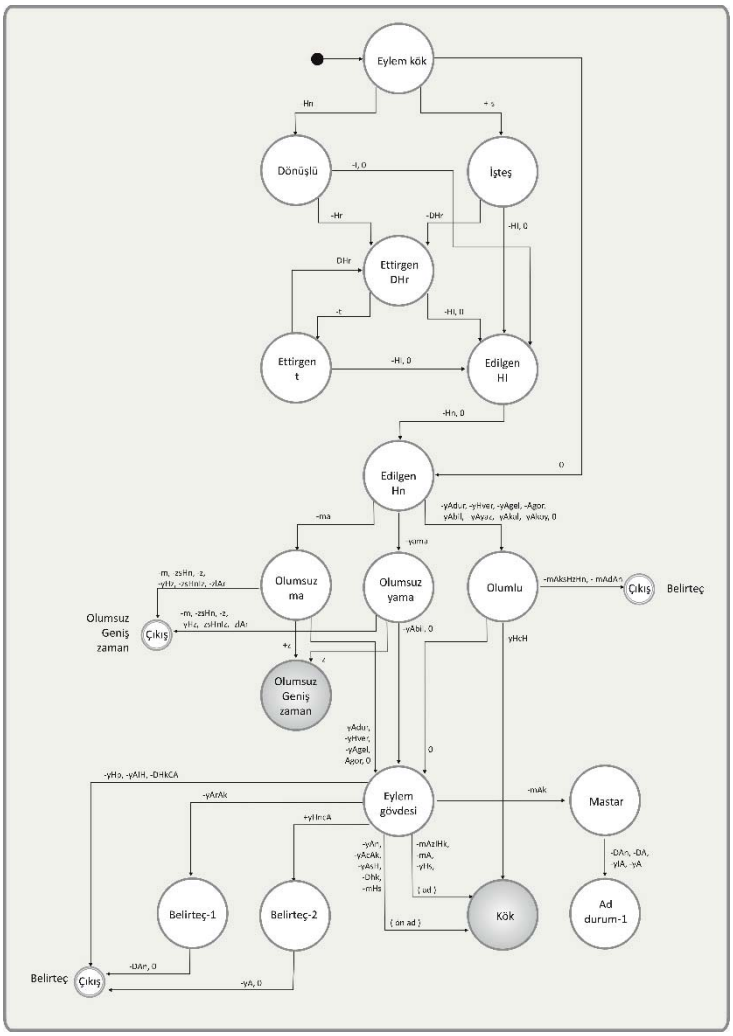
Proje kapsamında biçim bilimsel çözümleyici iki nedenle gerekli olmuştur: 1º Yazım yanlışlarını bulmak ve 2º soru tümcesi içindeki bir sözcüğün sınıf ve rolünü bulmak için. Bu amaçla ad ve eylemler için Şekil-3 ve Şekil-4'te gösterilen algoritmalar kullanılmıştır.

3.3 Sözcüklerin Rollerinin Belirlenmesi

Türkçe tümceler incelendiğinde 4 temel biçiminde olduğu görülmektedir:

- Basit tümce
- Sıralı bağımsız tümce
- Ortak nesneli tümce
- Ortak öznel, tümce

Proje kapsamında derlenen örnek soru tümcelerinin büyük bir çoğunluğunun basit ve sıralı basit tümceler olduğu görülmüştür. Türkçe sözcüklerin yapısı Özne Tümleç Yüklem (ÖTY) biçimindedir. Ancak Türkçe tümce yapılarının esnek olduğu da bilinmektedir.



Şekil-4: Türkçe eylemlerin biçim bilimsel çözümlemesi için Sonlu Durum Makinesi

Biçim bilimsel çözümleme sonucunda sözcüklerin sınıfları elde edilmiştir. Bu bilgileri kullanarak her sözcüğün tümce içindeki rolü (yüklem, belirteç, nesne ve özne) belirlenmiştir. Yüklem yalın ya da birleşik eylem olabilir. Belirteç genellikle yüklemin önünde yer alır; nesne yalın ya da ad öbeği biçiminde olabilir; özne açık ya da kapalı (gizli) olabilir.

Basit ve sıralı bağımsız tümce örnekleri Şekil-5 ve Şekil-6'da gösterilmiştir.

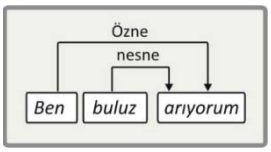
Bir soru tümcesi içindeki sözcüklerin rollerinin bulunmasında izlenen yol aşağıda açıklandığı biçimdedir:

1. Çift önad ve çift belirteçler bulunur.

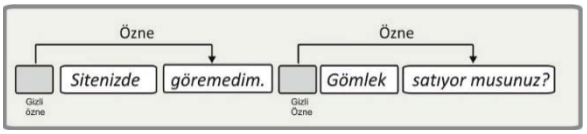
2. Sözcüklerin sınıfları (ad, eylem, önad, belirteç, adıl, ünlem, bağlaç, yabancı sözcük) belirlenir. Bu aşamada bazı ad, eylem, önad ve belirteçlerde belirsizlikler ortaya çıkmaktadır. Ortaya çıkan belirsizlikler Türkçenin dil bilgisi kuralları ile giderilmiştir.

3. Gövde sözcüklerin sınıflarını bulmak üzere her gövde incelenmiş ve sözcelerde aranmıştır. Sonuç olarak gövdenin sınıfı belirlenmiştir. Bu aşamada da bazı ad, eylem, önad ve belirteçlerde belirsizlikler ortaya çıkmaktadır. Ortaya çıkan belirsizlikler Türkçenin dil bilgisi kuralları ile giderilmiştir.

4. İlk üç adım tamamlandığında bir sözcüğün sınıfı belirlenmiş olmaktadır. Bir sözcüğün soru tümcesi içindeki sınıfının belli olması, soru tümcesinin anlamını ortaya çıkarmaktadır.

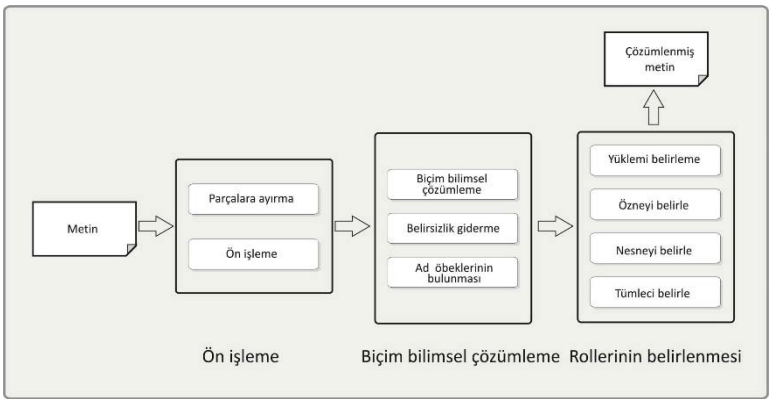


Şekil-5: Basit tümce yapısı



Şekil-6: Sıralı bağımsız tümce yapısı

Bir soru tümcesinin çözümlenmesine ilişkin akış Şekil-7'de gösterilmiştir.



Şekil-7: Bir soru tümcesinin çözümlenmesi

3.4 Sorunun Anlamlandırılması, Yanıtın Oluşturulması ve Koşut Derlemin Oluşturulması

Bir soru tümcesinde, yüklem müşterinin isteğini belirler ve nesne istediği şeyi belirler. Bu açıdan bu iki tümce ögesi önemlidir ve sorunun anlamını yüklenir. Bu nedenle, çalışmada önce yüklem ve nesnenin bulunmasına çalışılmıştır. Örneğin

S: *Mavi renkli, küçük boy pamuk kadın bluzunuz var mı?*

Yüklem: var mı? ve

Nesne: mavi renkli küçük boy kadın buluzu

Yanıt yüklem ve nesneye dayalı olarak oluşturulur. Bu arada ilgili ürünün var olup olmadığı işletmenin veri tabanından öğrenilir. Buna göre yanıt şunlardan biri olabilir:

Y: *Kadın bluzumuz var.*

Y: *Pamuk kadın buluzumuz var.*

Y: *Küçük boy kadın bluzumuz var.*

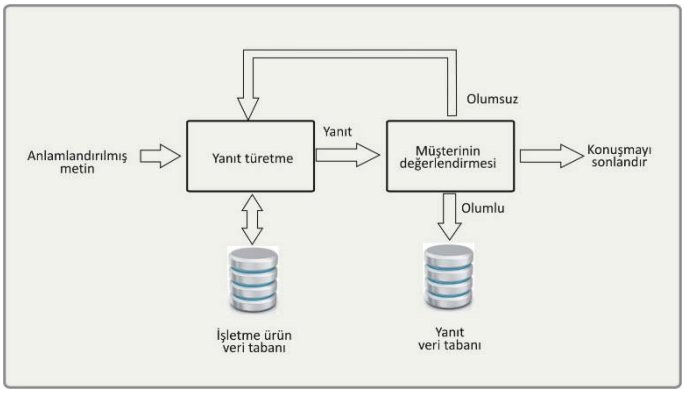
Y: *Mavi renkli küçük boy kadın buluzumuz var.*

Soru yanıtlandıktan sora müşteri işletmenin web sitesinde ilgili sayfaya yönlendirilir ve ardında şöyle bir soru yöneltilir:

Aradığınız ürünü bulabildiniz mi?

Müşterinin yanıtlaması için "Evet" ve "Hayır" düğmeleri sağlanır. Müşterinin yanıtı "Hayır" ise isteğini farklı bir tümce ile yazması istenir. Böylece müşterinin memnuniyeti sağlanmaya çalışılır ayrıca geliştirilen dizgenin soruyu anlama başarımı ölçülür. Bir sorunun anlamlandırılması ve yanıtın oluşturulmasına ilişkin yapı Şekil-8'de gösterilmiştir.

Şekil-8'de gösterilen yapı yeterince uzun süre çalıştırıldığında, koşut veri tabanı oluşturulmuş olmaktadır.



Şekil-8: Bir soru tümcesinin çözümlenmesi

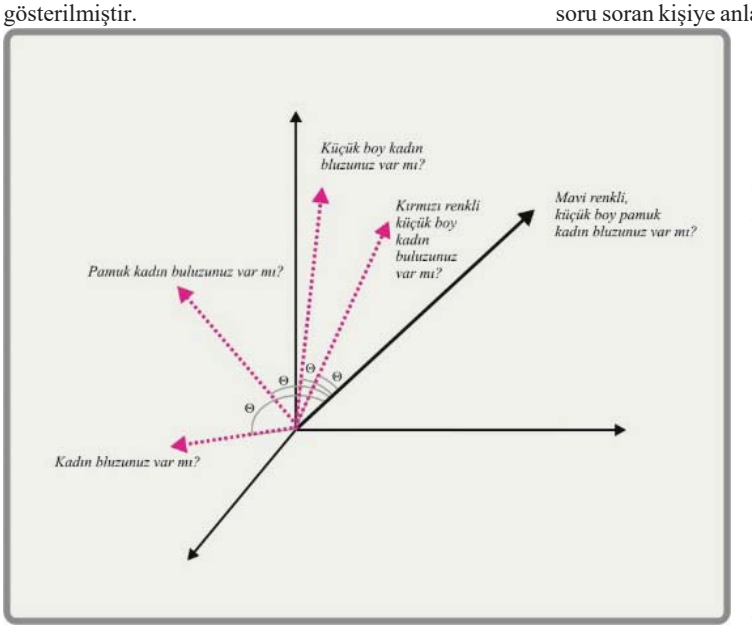
Koşut veri tabanı aslında iki veri takımından oluşmaktadır: Soru ve bu soruya verilmiş doğru yanıt. Koşut veri tabanı yeterli olgunluğa ulaştığında, soru yanıtlama dizgesinin konuyu yeterince öğrendiği kabul edilir. Bu noktadan sonra sorulan bir sorunun çözümlenmesi ve anlamlandırılmasına gidilmeden, benzer bir sorunun daha önce sorulmuş olup olmadığı koşut veri tabanının sorular kısmında araştırılır. Bu araştırma sırasında sorunun aynısı ya da benzeri bulunmaya çalışılır.

Sorulan bir sorunun aynı ya da benzerini bulmak üzere vektör uzayı yaklaşımı kullanılmıştır. Bu yöntemde sorulan bir soru içinde geçen her sözcük bir vektör ile temsil edilmektedir. Benzer şekilde, soru veri tabanındaki sözcükler de vektör biçiminde temsil edilmektedir.

$\vec{s} = (b_{1,1}, b_{1,2}, \ldots\ b_{1,n})$ sorulan soru

$\vec{b} = (a_{1,1}, a_{1,2}, \ldots\ a_{1,n})$ veri tabanındaki soru

İki vektör arasındaki benzerlik iç çarpım veya kosinüs benzerliği yöntemleri ile bulunabilmektedir. Şekil-9'da Kosinüs yöntemi ile yapılan bir karşılaştırma gösterilmiştir.



Şekil-9: Sorulan ve veri tabanındaki soruların karşılaştırılması

4. Sonuçlar ve Öneriler

Türk dili özelinde geliştirilen bu çalışmada ağırlıklı olarak Türkçenin dil bilgisi kurallarından yararlanılmıştır. Bir soru tümcesinin tümcelere ayrılması için hazır bir araç kullanılmış ve bu aracın başarımı proje için yeterli bulunmuştur.

Yazım yanlışlarını bulunması ve giderilmesi için Türkçenin ses bilimi özellikleri ve biçim bilimsel özelliklerinden yararlanılmıştır. Bu amaçla, heceleme ve biçim bilimsel çözümleme programları hazırlanmıştır. Ayrıca diğer ses özelliklerini de sınayan programlar hazırlanmıştır. Ön işleme adını verdiğimiz bu aşamanın başarımı %97'yi aşmaktadır ve bu başarım proje için yeterlidir.

Soru tümcesinin ögelerinin ve bu ögelerin rollerinin bulunması için özgün algoritmalar geliştirilmiştir. Algoritmalar kural temellidir. 2000 tümce üzerinde yapılan denemeler geliştirilen algoritmaların başarılı olduğunu göstermiştir. Ancak bu ölçümlerin daha büyük veri kümeleri üzerinde yapılması daha sağlıklı sonuçlar verecektir.

Sorunun anlamlandırılması ve buna karşılık yanıtın üretilmesi kural temelli olarak geliştirilmiştir. Soru tümcesinin yüklem ve nesnesine bağlı olarak bir tümce oluşturulmaktadır. Tümce oluşturulurken Türkçenin dil bilgisi kuralları gözetilmiştir. Böylece soru soran kişiye anlamlı ve düzgün bir tümce ile yanıt verilmesi başarılmıştır.

Müşteriye verilecek yanıt hazırlanırken, kuruluşun veri tabanında nesne araştırılmıştır. Türetilen yanıt, aranan ürünün bire bir eşinin bulunmasından en uzak benzerinin bulunmasına kadar değişkenlik göstermektedir. Böylece müşteriye var ya da yok yanıtı vermek yerine yakın ürünler önerilebilmektedir.

Geliştirilen dizge, temelde bir sorunun anlaşılması ve buna karşılık düzgün bir yanıtın hazırlanmasını amaçlamaktadır. Doğal olarak bu işlemler zaman almaktadır. Yapılan denemelerde geçen sürenin 1 saniye dolayında olduğu ölçülmüştür. Daha hızlı bir çözüm için, koşut derlem uygulaması denenmiştir. Koşut derlem zaman içinde dizge tarafından üretilmektedir. Koşut derlemin boyutu büyüdükçe daha başarılı olacağı açıktır. Makale hazırlandığında kullanılan koşut derlem içinde yaklaşık 10.000 tümce yer almaktadır.

Gerçekleştirilen dizge bir ürünün aranması ve buna karşılık verilmesi biçiminde denenmiştir. Buna ek olarak müşteri şikayeti, değiştirme ve iade, ödeme işlemleri ile ilgili kısımlar üzerinde çalışılmaktadır.

5. Kaynaklar

[1] J. Weizenbaum, *Computer and Human Reason: From Judgment to Calculation*. Pp. XII, 300 San Francisco, California. W. H. Freeman Co, 1976

[2] X. Li. D. Roth, *Learning Question Classifiers,* COLING '02 Proceedings of the 19th international conference on Computational linguistics - Volume 1 Pages 1-7, 2002

[3] X. Li. D. Roth, *Learning Question Classifiers: The Role of Semantic Information,* Natural Language Engineering 1 (1): 000–000. Cambridge University Press. 2004

[4] B. Ojokoh, P. Ayokunle, *Online Question Answering System*, International Journal of Computer Science Research and Application 2013, Vol. 03, Issue. 03, pp. 02-09

[5] A. Ghobadi, T. M. Rahgozar, *A knowledge-based question answering system for B2C eCommerce,*

Knowledge-Based Systems, Volume 21, Issue 8, December 2008, Pages 946-950, Elsevier

[6] R. P. de Azevedo, M. J. V. Pereira, P. R. Henriques, *DSL Based Automatic Generation of Q&A Systems*, WorldCIST'19 2019, AISC 930, pp. 460–471, 2019 Springer Nature Switzerland

[7] W. Cui, Y. Xiao, H. Wang, Y. Song, S. Hwang, W. Wang, *KBQA: Learning Question Answering over QA Corpora and Knowledge Bases*, Proceedings of the VLDB Endowment, Vol. 10, No. 5, 2017

[8] P. J. Paul, S. Amaran, K. S. Kumar, U.M. Prakash, *Question Answering System Using the Approach Of NLP*, International Journal of Pure and Applied Mathematics, Volume 117 No. 7 2017, 445-458, 2017

[9] N. Coşkun, *Türkçe Tümcelerin Ögelerinin Bulunması,* Yüksek Lisans Tezi, İTÜ Fen Bilimleri Ens. 2013

[10] K. Oflazer, *Two-level Description of Turkish Morphology*, Bilkent University

[11] G. Eryiğit, *Biçimbilimsel Çözümleme,* TBV Bilgisayar Bilimleri ve Mühendisliği Dergisi sayı: 6, 2012

.