
Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Anlamlı ve Benzer Olmayan Türkçe Metinler Üretmek için N-Gram Yöntemi ile İstatistiksel ve Kural Tabanlı Yaklaşımın Birlikte Kullanımı

Combining Statistical and Rule-Based Approach with N-Gram Method to Generate Meaningful and Dissimilar Turkish Texts

Mehmet Ali Kutlugün [1*], Yahya Şirin [2]

[1,2] İstanbul S. Zaim Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi Bilgisayar Müh. Bölümü, İstanbul, TÜRKİYE
Sorumlu Yazar / Corresponding Author *: mehmet.kutlugun@std.izu.edu.tr

Geliş Tarihi / Received: 26.07.2019 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 23.12.2019 *DOI: 10.21205/deufmd.2020226504*
Atıf şekli/ How to cite: KUTLUGÜN, M. A., ŞİRİN, Y.(2020). Anlamlı ve Benzer Olmayan Türkçe Metinler Üretmek için N-Gram Yöntemi ile İstatistiksel ve Kural Tabanlı Yaklaşımın Birlikte Kullanımı. DEUFMD, 22(65), 343-352.

Öz

Metin üretme çalışmaları, mevcut cümlelerin analiz edilerek yeni bilgi çıkarımlarının yapıldığı, varolan bir bilgiden yola çıkarak bununla bağlantılı anlamlı bilgilerin elde edildiği sistemlerdir. Bu sistemler, özellikle internet üzerinden yapılan aramalarda girilen cümlelerin türetilerek kullanıcılara arama sonuçları ile ilgili daha anlamlı sonuçlar döndürebilmek gibi kolaylıklar sağlarlar. Bir metin üretici geliştirmek için, doğal dilin kaynaklarını tanımlayacak bir dilbilim teorisine ve bu kaynakları bilgisayar ortamında işleyecek bir yazılım aracına ihtiyaç vardır. Bu çalışmada, kaynak veri kümesindeki cümlelerden sınıf tabanlı n-gram modeli kullanılarak Türkçe dil yapısına uygun anlamlı ve yeni cümleler oluşturulması hedeflenmiştir. Cümle üretimini gerçekleştirmek için trigram modeli önerilmiş ve bu model kural tabanlı yaklaşım ile birlikte kullanılmak için geliştirilmiştir. Bu çalışmada kullanılan yöntem ile diğer kullanılan yöntemlerden farklı olarak, üçlü kelime grupları şeklinde ayrılan gruplardan belirlenen kurallar çerçevesinde, art arda ekleme yöntemiyle anlamlı ve farklı cümleler üretilmesi gerçekleştirilmiştir. Böylece kaynak metin dosyasındaki kelime ya da kelime gruplarından ilişkili olduğu grup sayısı kadar farklı cümleler birbirine bağlanarak yeni metinler oluşturulmuştur.

Anahtar Kelimeler: *Doğal dil işleme, Doğal dil üretme, Metin işleme, Metin üretme, İstatistiksel dil modelleme*

Abstract

Text generation studies are the systems which new knowledge inferences are made by analyzing the existing sentences and meaningful information is obtained from an existing knowledge. These systems provide convenience to users to return more meaningful results related to search results, especially on internet searches. To develop a text generator, there is a need for a linguistic theory to define the sources of natural language and a software tool to process these resources in computer environment. In this study, it is aimed to generate meaningful new Turkish sentences using class-based n-gram model from the sentences in the source data set. Trigram model has been proposed to generate sentences and this model has been developed for use with rule based approach. Unlike

other methods, the method used in this study produced meaningful and different sentences with the successive addition method within the framework of the rules determined from the groups divided into triple word groups. Thus, new texts were generated by connecting different sentences from the word or word groups in the source text file as much as the number of the groups that associated with.

Keywords: *Natural language processing, Natural language generating, Text processing, Text generation, Statistical language modeling*

1. Giriş

İnsan ile bilgisayar arasındaki etkileşim giderek artan bir öneme sahiptir. İnternet ortamında mevcut milyarlarca sayfalık bilgi yayımlanmaktadır. Büyük boyutta yazılı veya sözlü metinlerin otomatik olarak çözümlenmesine yönelik tekniklerin geliştirilmesi, çok büyük boyutta bulunan bu metinlerin sayısal ortamda işlenmesine olanak vermektedir. Böylece, insan emeğiyle daha uzun sürede yapılacak bazı çalışmalar, bilgisayarlar aracılığıyla çok kısa sürede yapılabilmektedir. Bu kolaylıklarla birlikte, doğal dil işleme (DDİ) çalışmaları yeni bir boyut kazanmıştır[1]. DDİ uygulamaları ile karakter belirleme, metin doğrulama, metin okuma, otomatik metin çevirisi, konuşma tanıma, yazılan bir kelimeye göre bir sonraki kelimenin tahmin edilmesi, bilgi bulup çıkarma, metin özetleme, soru yanıtlama, doğal dilde anlamlı cümle ve metin üretme, büyük boyuttaki metin içinde ifade arama ve yabancı dil eğitimi gibi çalışmalar yapılabilmektedir[2].

Bu çalışmada, anlamlı Türkçe yeni metinler üretmek için tekrar eden kelime gruplarından yararlanılmıştır. Kaynak veri kümesinden elde edilen bu kelime grupları, doğal dilin bazı temel kuralları ile işlenerek anlam bütünlüğüne sahip yeni metinler üretilmiştir.

1.1. Metin üretme

Doğal Dil Üretimi (DDÜ), metin planlama ve metin üretme olmak üzere iki bölüme ayrılmaktadır. Metin planlama kısmında, kavramsal girdilerden metnin anlamsal tanımı üretilir. Daha sonra metin üretme sistemi, bu anlamsal tanımları gerçek bir metne dönüştürür[3]. Bir metinden farklı yeni metinler üretme, farklı bir konuda benzer metinler yazma veya aynı başlığı içeren yeni bir metin oluşturma gibi amaçlar için kullanılmaktadır[4]. Bu sayede, yapılandırılmış veri üzerinden yorum çıkarma gerçekleştirilebilmektedir.

Bir metin üretici geliştirmek için, doğal dilin kaynaklarını tanımlayacak bir dilbilim teorisine ve bu kaynakları bilgisayar ortamında işleyecek bir yazılım aracına ihtiyaç vardır[2]. Doğal diller arasındaki dilbilgisi ve gramer yapısı büyük ölçüde farklı olduğundan, oluşturulacak sistem, dili en uygun biçimde temsil edebilmelidir. İlgisiz sonuçları filtrelemek için ilgi düzeyinin doğrulanması gereklidir. Metin üretme işlemi, hem söz dizim kuralları ve morfolojiye uygun olmalı hem de üretilen yeni metinler arasında tutarlılık bulunmalıdır.

Metin üretme işlemleri, istatistiksel ve kural tabanlı olmak üzere iki biçimde gerçekleştirilebilmektedir. Kural tabanlı yaklaşımlar, sözcüğün, cümlenin ve metnin özelliklerini kullanırlar[5]. Bu şekilde dil'e ait biçimbilimsel, sözdizimsel ve anlambilimsel özellikler kullanılarak kurallar meydana getirilir[6]. İstatistiksel yöntemler ise, bir çeşit gözlenme sıklığı bilgisini temel alarak verilen bir derlemdeki eşdizimlerden faydalanmaktadırlar[7].

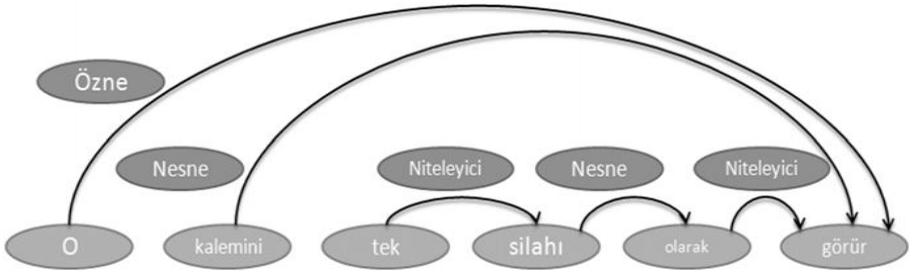
Bu çalışmada üretilen metinlerin tutarlılığını sağlamak amacıyla n-gram modelinden yararlanılmıştır. İki veya daha fazla sözcüğün birlikte gözlenme sayısını temel alan bu model uygulanırken, derlemde yan yana bulunan sözcüklerin gözlenme sıklıkları ölçülmektedir. Doğal dil'e ait sözdizimsel kurallar da dikkate alınarak kurallı bir yapı oluşturularak yeni cümleler üretilmektedir. Özetle, kural tabanlı yaklaşım ile istatistiksel yaklaşım uygun şekilde birleştirilerek birlikte kullanılmıştır.

2. Önceki Çalışmalar

Bu uygulama alanında farklı dillerde, istatistiksel ve kural tabanlı birçok yöntem kullanılmıştır. Mocan [5] ilköğretim düzeyindeki öğrencilerin okuduklarını anlama düzeylerinin geliştirilmesi amacıyla, metin işleme ve anlamaya dayalı soru soran bir sistem tasarımı üzerinde çalışmıştır. Eğitim alanında, ilköğretim düzeyindeki öğrencilerin anlama düzeylerinin ölçülmeye çalışıldığı bu çalışma,

girilen metinleri; eklerine ayırma, yan cümleler içeren karmaşık cümleleri basit cümleciklere ayırarak aralarındaki bağlantıları tespit etme ve basit cümleleri öğelerine ayrılarak bu öğelerin türlerine göre belli sorular üretme aşamalarından oluşmaktadır. Uygulama gerçek anlamda mantıksal bir düşünme sistemine sahip değildir. Ancak verilen metin içerisinde kuralların belirtilmesi ile sorular üretebilmektedir. Ayrıca bir sözcüğün aynı anda hem isim, hem sıfat, hem de zamir olarak kullanıldığı durumlarda Türkçe'nin esnekliğinden dolayı yanlış sorular üretebilmektedir. Sistem, sadece soru üretme üzerine kurulduğundan dolayı geniş kapsamlı ele alınmamıştır.

Bilgin ve Amasyalı [7], Türkçe için verilen giriş kümesindeki kelimelerden önerdikleri Şartlı Rasgele Alanlar (Conditional Random Fields-CRF) modeli ile istatistiksel dizilim sınıflandırmasına dayalı etiket atama sistemi üzerinde çalışmışlardır. Aday etiketlerden hesapladıkları olasılık dağılımındaki yüksek değerli etiket dizilimini seçerek kelimeleri gruplara ayırmışlardır. Kullanılan sistemin dile bağımlı yapısından dolayı bazı diller için CRF'nin performansı daha düşük çıkmıştır.



Şekil 1. Bağlılık grafiği [7].

Adalı ve Erenler [8], Türkçe için otomatik metin ve konuşma üretim sistemi üzerinde çalışmışlardır. Çalışmalarında, bir hasta veri tabanından alınan verileri Türkçe dilinin kurallarına uygun şekilde dizerek anlamlı ve dilbilgisi kurallarına uygun raporlar oluşturmayı hedeflemişlerdir. Bunun için gramer yapısının oluşturulmasında ilişkisel veri tabanından yararlanmışlardır. Ancak, çalışmada kullanılan yöntem için oluşturulan gramer yapısı, mevcut veri tabanı ile sınırlı kalmaktadır.

Özdemir ve Amasyalı [9], bilgi parçaları arasından anlamsal ilişkilerin tutulduğu bir veri tabanından yararlanarak yeni metinler üretme üzerine çalışma yapmışlardır. Veri tabanındaki birbiriyle kelime bazında benzerliği olmayan ancak anlamca birbirine yakın olan metinlerin bulunması hedeflenmiştir. Ancak, gramer yapısı esnek olmayan, sadece eşanlamlı veya benzer kelimelerin yer değiştirmesi ile çalışma sonuçlanmaktadır. Sistemin performansı, kullanılan veri tabanına oldukça bağımlı olduğundan veri tabanı içerik zenginleştirmesi gerekmektedir.

Metin ve Karaoğlan [10] Türkçe bir derlemde eşdizim şeklindeki sözcük grupları üzerinde çalışarak, otomatik olarak eşdizim belirlenmesi için istatistiksel yöntem tekniğini uygulamışlardır. Bu uygulamada tam bir cümle oluşturmaktan ziyade, deyim, bileşik fiiller, söz öbekleri ve tamlamalar şeklindeki eşdizimler üzerine çalışılmıştır. Çalışmada, kural tabanlı yöntemler dikkate alınmadan sadece istatistiki yöntemler ile eşdizim belirlenmesine çalışılmıştır. Ayrıca sözcükler arasında kurulan bağlar ölçülmediği için anlamlı tam cümleler üretilememiştir.

Burcu Can [11], yapmış olduğu çalışmada, DDİ alanında günümüzde oldukça popüler bir yöntem olan Uzun Kısa Süreli Bellek Ağları (Long-short Term Memory Networks-LSTM)[12] kullanarak daha çok sözcük anlamının önemli olduğu uygulamalar için bir Türkçe kök bulma yöntemi önermiştir. Ancak önerilen yöntem tamamen kök bulma ile sınırlandırılmıştır.

Türkçe dışındaki bazı dillerde ise; Brown ve arkadaşları [13] istatistiksel teknikler ile sınıf-bazlı n-gram modeli kullanarak İngilizce kelimeleri farklı sınıflara sözdizimsel veya semantik olarak gruplama çalışması gerçekleştirmişlerdir. Bu çalışma, tam ve

anlamlı bir cümle oluşturmaktan ziyade, sözdizimsel veya anlamsal olarak kelime gruplarını bir küme altında toplama ile sınırlandırılmıştır.

Mairesse ve arkadaşları [14] sözdizimsel açıklama içermeyen, semantik kavramlardan bir biçim oluşturmak için, alana özgü veriler kullanarak dinamik Bayes ağını kullanan istatistiksel bir İngilizce dil üretici oluşturmuşlardır. Anlamsal olarak ilişkilendirilen yığınlardan yararlanarak oluşturulan cümleler her ne kadar farklı olsa da, cümle sıralaması dikkate alınmadığından önemli ölçüde benzerlikler içermektedir.

Uchimoto ve arkadaşları [15] anahtar kelime veya başlıklardan cümleler üretmek için bir yöntem sunmuşlardır. İki aşamadan oluşan bu yöntemde geliştirdikleri model ile bilgi boşluğu veya kayıp kelimeleri bağımlılık ağaçları biçiminde ele alarak aday metin yapısı oluşturmuşlardır. Sözcükler arasındaki bağımlılık bilgilerinin yanı sıra n-gram bilgisini de dikkate alarak verilen anahtar kelimeler ile uygun cümleler elde etmişlerdir. Çalışmada daha fazla bilgiyi göz önüne alarak kurallar oluşturmak yerine, kısıtlı üretim kuralları ile çalışıldığı için, hem doğal olmayan hem de gramer olarak uygun olmayan cümleler oluşmaktadır.

Tan Jiwei ve arkadaşları [16], bir belgenin önemli cümlelerini tanımlayan ve daha sonra benzer cümlelerden başlık üretmek için hiyerarşik cümle özetleme modelini önermişlerdir. Ancak bu çalışma ile kabaca oluşturulan özet cümlelerden yararlanarak başlık bilgisi elde edilmeye çalışılmıştır. Çalışmada en fazla 50 kelimeden ve bir veya iki cümleden oluşan özet cümleler oluşturulmuştur. Ancak bu özet cümleler tüm belgenin anlaşılmasına dayanan tamamen uçtan uca bir yöntem sunmamaktadır.

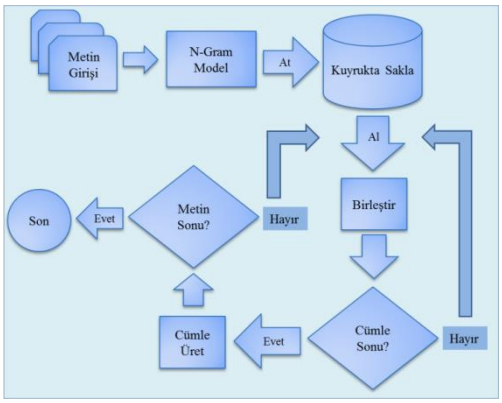
Bauer ve arkadaşları [17] farklı veri tabanlarından gelen verileri birbirine bağlayan bir algoritma ile veri ilişkilerini, korelasyonları, etkileri ve özel bazı özellikleri tanımlayarak, kural tabanlı farklı türlerde doğal dil metinleri üreten bir sistem geliştirmişlerdir. Bu sistemde kural temelli yaklaşımın DDÜ için uygun olduğunu ve eldeki problemleri güvenilir ve hızlı bir şekilde çözebileceğini ileri sürmüşlerdir. Ancak, dillerin sözdizimini açıklayan farklı kural setleri aracılığıyla ifade edilebileceğini vurgulayarak, farklı dillerde kullanılabilen ve cümleleri birden çok dilde ifade edebilecek kadar esnek bir makine oluşturmanın temelde teknik zorluk olduğunu belirtmişlerdir.

Bu konuda yapılan çalışmalarda görüleceği üzere, sadece bir yöntemi kullanarak doğal dil çözümlemesi yapmak cümle üretmek için yeterli değildir. Tek başına kural tabanlı yöntem kullanmak, dilbilgisi kurallarını yazmanın zorluğu nedeniyle kabul görmemektedir. Bunun yanında, dilin biçimbilimsel yapısının genel olarak sistematik olmasından ötürü, bu konuda kural tabanlı yaklaşım ağırlık kazanmaktadır[18]. Bu çalışmada, iki yöntem birlikte kullanılarak Türkçe metin üretme işlemi gerçekleştirilmiş, bu sayede sadece tek bir kalıba bağlı kalmadan, değişken uzunluklarda cümleler elde edilmiştir. Oluşturulacak cümleler için n-gram modelinin sağladığı kelime gruplarının birlikte bulunma avantajı ile çok detaylı kural tanımlamaları yapılmasına gerek kalmamıştır. Kelimeleri ayrı ayrı ele almak yerine, kelime gruplarının birbirleri ile olan ilişkisinden faydalanılarak belirli kurallar çerçevesinde kelimeleri birleştirmek, sonuca ulaşmak için yeterli görülmüştür. Bunun yanında, kullanılan derlemin boyutuyla kıyaslandığında üretilecek yeni cümleler, daha az sayıda olabilmektedir. Bunun sebebi, belli bir anlamsal bütünlük içerisinde cümlelerin birleştirilmek istenmesi gösterilebilir. Kelime gruplarının istatistiklerine göre bigram veya trigram kullanılması nispeten konu uyumu sağlamaktadır. Üretilen cümleler takip eden gruplar ile uygun şekillerde birleştirildiğinde, aynı konu üzerinde devam eden okunaklı metinler ortaya çıkmaktadır.

3. Deney Sistemi

3.1. Amaç

Çalışmanın amacı, kaynak veri kümesindeki cümleleri n-gram şeklinde kelime gruplarına ayırmak ve istatistiksel yaklaşım ile kural tabanlı yaklaşımı uygun biçimde sentezleyerek, Türkçe anlamlı yeni cümleler oluşturmaktır. Bunun için, her üçlü kelime grubunun cümle öğeleri sınıf-bazlı olarak ele alınmış ve kaynak cümleler ile aynı olmayacak şekilde, yeni anlamlı cümleler elde edilmiştir. Önerilen çalışmaya ait süreç diyagramı Şekil.2'de gösterilmiştir.



Şekil 2. Süreç Diyagramı

3.2. Veri kümesi

Veri kümesi için, Yıldız Teknik Üniversitesi Kemik Doğal Dil İşleme Grubunun sağladığı '42 bin haber' isimli veri kümesinin 'Genel' kategorisindeki 6673 adetten oluşan haber içerikli metinleri kullanılmıştır[19].

4. Deneysel Yöntem

N-gram modeli ile bigram ve trigram kelime gruplarının sıklığı hesaplanarak, azalan sırada yığınlarda toplanmıştır. Her bir kelimenin ilişkili olduğu sınıf (öğe grubu) belirlenerek yeni anlamlı ve tam cümleler üretilmiştir. Üretilen bu yeni cümlelerin hiçbiri diğerleriyle aynı olmayacak şekilde tasarlanmıştır.

4.1. N-gram modeli

İstatistiksel DDİ'de iki veya daha fazla sözcüğün birlikte gözlenme sayısını temel alan bu yöntem uygulanırken derlemde yan yana bulunan (n-gram) sözcüklerin gözlenme sıklıkları ölçülmektedir[18]. Derlemdeki bu sözcük grupları sıklık değerleri azalan sırada listelenir, bu liste oluşturulacak cümleler için aday kelime gruplarını içermektedir. Listede yüksek sıklık değerine sahip olan adayların uygun eşdizimler olduğu kabul edilmektedir[10]. Çoğunlukla n-gramlar 1-gram (unigram), 2-gram (bigram), 3-gram (trigram) şeklinde kullanılmaktadır. Trigram modeli, bir alandaki olası cümlelerin büyük bir bölümünü kaplayan yeterli eğitim verisi varsa oldukça güçlüdür[20]. Bu yöntemdeki en büyük problem gerçek eşdizimleri, diğer sözcük birliklerinden ayıran eşik değerin belirlenmesi aşamasıdır. Yöntemin dezavantajı ise çok sıklıkla gözlenen eşdizim adayları içinde işlevsel bazı kelimelere (bu, ve, şey ,vb.) rastlanmasıdır[21]. Bu sebeple sıklık değerine bağlı olarak hazırlanan listelerde sözcük türü filtresi gibi çeşitli filtreler uygulanmaktadır[22].

Tablo 1. Trigram sıklık tablosu örneği.

Sıra	Trigram Kelime Grubu	Sıklık
1	Meydana gelen depremden	4
2	gelen depremden sonra	4
3	depremden sonra tsunami	2
4	sonra tsunami uyarısı	2
5	tsunami uyarısı yapılmıştı.	2

4.2. Türkçe cümle yapısı

Türkçe, cümle öğelerinin sıralanışı açısından bir Özne-Nesne-Yüklem (Subject, Object, Verb-SOV) sıralı dil olarak kabul edilebilir. İsimler durum sonekleri almadığında SOV dizimi varsayılır. Ancak Türkçe, cümle öğelerinin sırasını değiştirmemize olanak tanımaktadır[2]. Bu çalışmada, genel gramer kuralı olarak SOV sırası dikkate alınarak cümleler üretilmiştir.

4.3. Kelime türünün belirlenmesi

Bir kelimenin türü ya da diğer bir deyişle ait olduğu sınıf; İsim (Noun), Fiil (Verb), Zamir (Pronoun), Sıfat (Adjective), Zarf (Adverb), Edat (Preposition), Bağlaç (Conjunction) olabilir. Bazı kelimeler bu sınıflardan sadece birine ait iken, bazıları birden fazlasına ait olabilir. Kelimenin türünün belirlenmesi işlemi, bir kelimenin cümle içindeki konumu incelenerek, o kelime için doğru konuşma bölümü (Part-of-Speech-POS) tespit edilerek gerçekleştirilebilir[5].

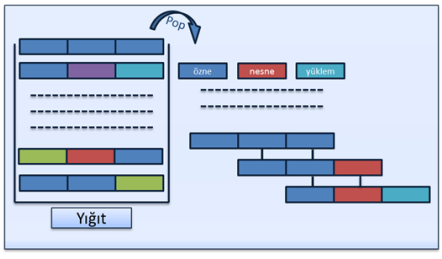
Tablo 2. Tespit edilen kelime türleri örneği.

Sayı	Kelime	Tür
1	hesaplamak	Eylem
2	hesap	Ad
3	ziyaret	Ad
4	dilek	Ad
5	getirdi	Eylem (çekilmiş)

4.4. Malt ayrıştırma

Malt Parser, Joakim Nivre tarafından İsveç'te Vaxjo Üniversitesinde gerçekleştirilmiş bir ayrıştırma algoritmasıdır. Ötele-İndirge (shift-reduce) ve ayrıştırma modeli olarak Destek Vektör Makineleri (Support Vector Machine-SVM) kullanan bir programdır. Dilden bağımsız yapısıyla yüksek doğruluk değerlerine ulaşabilir. Ötele-İndirge algoritması, genelde cümleyi soldan sağa doğru, iki farklı veri yapısından faydalanarak ayrıştırır. Yığında işlenmekte olan sözcükler tutulurken, kuyrukta işlenmek üzere bekleyen sözcükler tutulur. Ayrıştırıcı her adımda üç hareketten birini uygular. Bu hareketler; Öteleme, Soldan Sağa Bağla ve Sağdan Sola Bağla şeklindeki durumlardır. Öteleme işleminde kuyrukta bekleyen elemanın yığına itilmesi demektir. Bu önceki kelimeyle bir bağlantı oluşturulamadığı ya da yığının boş olduğu durumlarda gerçekleşir. Yığındaki eleman ile sırada bekleyen eleman arasında sağa doğru bir bağ varsa Sağa Bağlama işlemi gerçekleşir. Eğer sola doğru bir bağ varsa Sola Bağlama işlemi gerçekleşir[23]. Ayrıştırma modeli, yığının en üstündeki ve kuyruğun en başındaki elemana bakarak bir sonraki hareketin ne olacağına karar verir[7].

Bu çalışmada kullanılan yöntem ile kaynak cümlelerdeki üçlü kelime grupları mevcut bütün alternatifleri ile hesaplanarak uygun çözümlemeler elde edilebilmeye çalışılmıştır. Tüm uygun trigram ifadeler bir yığında toplanarak, en sık geçen en üstte yer alacak şekilde sıralanmıştır. Yığından çıkarılan (pop) ifadeye bağlanabilecek en uygun alternatif sondan iki bağlı trigram veya sondan tek bağlı trigramlar şeklinde rasgele seçilerek farklı cümleler üretilmiştir. Bu işleme yığında bulunan uygun trigram ifadeler bitene kadar devam edilmiştir. Böylece yığındaki uygun ifadelerin sayısı kadar cümle oluşturulabilmiştir.



Şekil 3. Soldan sağa bağlama ile birleştirme

Cümleleri birleştirmek için uygun bigram veya trigram ifadeler soldan sağa bağla kuralı ile ele alınmış ve metinsel ifadeler elde edilmiştir.

5. Deneysel Çalışma

Uygulama Java platformunda gerçekleştirilmiş olup, veri kümesinden elde edilen metinler içinden bigram ve trigram şeklinde kelime grupları elde edilmiştir. Bu gruplar azalan frekans sırasında bir yığında tutularak, öğelerin kontrolü gerçekleştirilmiştir. Öğeleri tam bir cümle oluşturan trigram ifadeler yığından atılmıştır. Öğelerinden bir veya birkaçı eksik olan ifadelerden sayıca en fazla geçenler ile yeni cümleler oluşturulmuştur. Cümlenin başında kullanılacak ifadeler için, büyük harf ile başlayan adaylardan seçim yapılmıştır. Nokta işareti ile biten en uygun tamamlayıcı ifadeler ile cümleler sonlandırılmıştır. Bu ifadelerden birbirini tamamlama ölçülerine göre kısa veya uzun cümleler üretilebilmiştir. Örneğin; "Meydana gelen depremden etkilendi." şeklinde, kısa formda cümleler üretilebildiği gibi, "Meydana gelen depremden sonra tsunami uyarısı yapılmıştı." şeklinde, daha uzun cümleler de üretilebilmektedir.

Uygulamaya ait tüm kurallar aşağıda açıklanmıştır. Bu kurallar uygulanmadan önce veri kümesinden ayrı ayrı elde edilen 3-Gram ve 2-Gram ifadeler tespit edilerek azalan şekilde sıralanmıştır. Kural tabanlı yapının uygulanması sonucu üretilecek metin en fazla 5 cümleden oluşmaktadır.

Cümle başı 3-Gram ifade seçimi: Elde edilen 3-Gram ifadeler içinden büyük harf ile başlayanlar cümle başı olarak değerlendirilmiştir. En sık geçenden başlanarak yeni cümleler oluşturulmaya çalışılmıştır. Üretilen yeni cümlelerin orjinal metinde geçmemesine dikkat edilmiştir. Cümle başı için seçilen bir 3-Gram ifade nokta işareti ile bitmemelidir. Ör. "Başbakan uçağa bindi." ifadesi tam bir cümle (tüm öğelere sahip) olduğu için yığından çıkarılmıştır. Ayrıca rakamsal ifadeler ile başlayan cümleler kapsam dışında tutulmuştur. Örneğin; "30 Ağustos zafer ..." ifadesi ile cümleye başlanmamıştır.

İzleyen (niteleyici) 3-Gram ifadelerin seçimi: İzleyen ifadelerden tam bir cümle oluşması için, ifadenin son veya son iki kelimesi aynı olanlar seçilmiştir. Bu işleme soldan sağa bağla kuralı dikkate alınarak devam edilmiştir. Böylece

Tablo III'te görüleceği gibi iki grup arasında ortak bir bağ kurulmuş olmaktadır.

Tablo 3. İzleyen ifade örneği.

	Trigram Grupları	Sıklık
Cümle Başı İfade	Meydana **gelen depremden**	4
İzleyen ifade	**gelen depremden** sonra	3

İzleyen ifadeler nokta ile bitebilir. Bu şekilde en az 4 kelimeden oluşan bir tam cümle elde edilmiş olur. Örneğin "Bilgisayar dünyasına yeni katıldı." ifadesi, öncelikle büyük harf ile başlayan "Bilgisayar dünyasına yeni" trigram ifadesi ve "dünyasına yeni katıldı." izleyen ifadesi ile tamamlanmıştır. Nokta işaretine ulaşıldığı için bir tam cümle elde edilmiş olmaktadır. Özetle, cümle başı büyük harf ile başlayıp noktaya kadar devam eden diğer trigramların eklenmesi ile cümle üretilmektedir. Bu işlem, sondaki iki kelime seçilerek yapılabileceği gibi, sadece son kelime seçilerek de yapılabilir. Aynı cümle başı trigram ifadesi ile başlayıp sondan iki ya da tek kelime bağlı izleyen ifadelerin seçimi ile aşağıdaki gibi cümleler üretilebilmektedir.

Tablo 4. Sondan iki bağlı ifadeler

Sıra	Trigram Grupları	Sıklık
1	Meydana **gelen depremden**	4
2	**gelen depremden** sonra	4
3	depremden sonra tsunami	2
4	sonra tsunami uyarısı	2
5	tsunami uyarısı yapılmıştı.	2

Tablo 5. Sondan tek bağlı ifadeler

Sıra	Trigram Grupları	Sıklık
1	Meydana gelen **depremden**	4
2	depremden sonra tsunami	2
3	tsunami uyarısı yapılmıştı.	2

Diğer cümlelerin üretimi:

İlk üretilen cümleden sonra devam edecek diğer cümleler kendinden önceki cümleler ile bağ kurabilecek şekilde, trigram veya bigram ifadelerden oluşmaktadır. Bu işlem iki yöntem ile gerçekleştirilmiştir. Birinci yöntemde iki cümle trigram ifade ile birleştirilebilir. Tüm metin için bu şekilde sıkı bir bağ kurarak üretilmesi oldukça zordur. Çünkü, üretilen bir cümlenin tamamlayıcısı veri kümesinde bulunamayabilir. Örneğin yukarıdaki örnekte verilen cümle bir başka cümle ile aşağıdaki gibi birleştirilebilmektedir.

…tsunami uyarısı yapılmıştı. → uyarısı yapılmıştı. **Yetkililer** ...

…tsunami uyarısı yapılmıştı. → uyarısı yapılmıştı. **Uyarılardan** ...

cümlelerinden iki farklı metin oluşacak şekilde birleştirilebilmiştir. Bu şekilde bağlanan cümleler arasında kuvvetli bir anlamsal bağ meydana gelmekte, akıcı ve anlamlı bir metin elde edilmektedir. İzleyen cümle şu şekilde devam edebilir:

…yapılmıştı. **Uyarılardan haberi** → **Uyarılardan haberi** olmayan...

İkinci yöntemde ise cümleleri birleştirmek için bigramlar tercih edilmiştir. Bu yöntem ile daha kuvvetli bir bağ kurulmasa da daha fazla sayıda farklı metin üretilmiş olur. Aşağıdaki örnek ifadede görüldüğü gibi ilk cümlenin son ifadesi ile izleyen cümlenin cümle başı tek bağ ile birleştirilmiştir.

... tsunami uyarısı **yapılmıştı**. (Cümle sonlu Trigram ifade) → …**yapılmıştı**. Yetkililer (Bigram ifade)

Yeni cümle, bundan sonra trigram ifadeler ile devam edebilir. Bu yöntemde, iki cümle birbiri ile bağlanırken bigram ifadeden yararlanılmaktadır. Örneğin;

…**yapılmıştı**. Yetkililer (bigram) → Yetkililer **tüm uyarılara** (trigram) → **tüm uyarılara** rağmen → uyarılara rağmen vatandaşları... şeklinde sondan iki bağ ile kuvvetli bir bütünlük içinde devam edebilir. Ya da;

…**yapılmıştı**. Yetkililer (bigram) → Yetkililer tüm **uyarılara** → **uyarılara** rağmen vatandaşları… şeklinde sondan tek bağ ile devam edebilmektedir. Kısacası cümledeki bağ, ortak tek veya iki sabit kelime ile sağlanmaktadır.

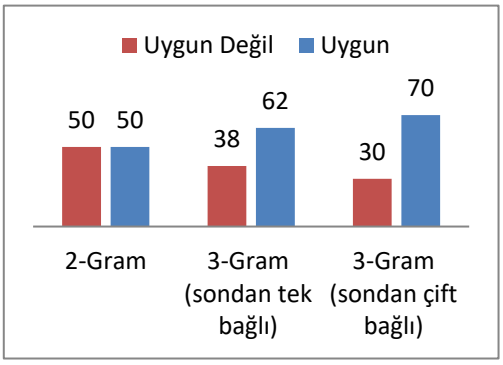
6. Deneysel Sonuçlar

Bu çalışmada üretilen cümleler, sınıflama ölçeği (nominal scale) ile gramatik açıdan;

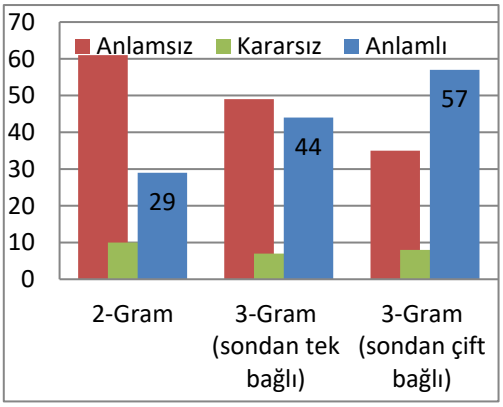
"uygun/uygun değil" ve anlam açısından; "anlamlı/anlamsız" ölçütleri ile birbirinden farklı 100 adet katılımcıya iki seçenekli anket uygulaması şeklinde sunulmuştur. Sistemin başarımı, verilen 'uygun ve 'anlamlı' sayılarının ortalaması alınarak her yöntem için üretilen 25 adet cümle sonuçlarından elde edilen tutarlılık yüzdesi ile ölçülmüştür. Deneysel sonuçlara ait tablo ve grafikler aşağıda verilmiştir.

Tablo 6. Başarım değerlendirme tablosu

Yöntem	Gramatik Uygunluk Oranı (%)	Anlamlılık Oranı (%)
2-Gram	50	39
3-Gram (sondan tek bağlı)	38	51
3-Gram (sondan iki bağlı)	70	65



Şekil 4. Gramatik uygunluk sonuçları



Şekil 4. Anlamlılık sonuçları

Sonuçlar incelendiğinde, 2-Gram yöntemi ile cümle üretmenin mümkün olmadığı, ancak bazı tekrarlar için kullanılmasının faydalı olacağı düşünülebilir. 3-Gram yöntemlerinde ise, kuvvetli bağ kuruldukça aşamalı olarak artacak şekilde hem gramatik olarak uygun, hem de daha anlamlı cümleler üretilebildiği görülmektedir.

Üretilen bazı cümleler gramer olarak doğru olmasına rağmen, anlam açısından uygun olmayan cümleler de üretilmiştir. Örneğin "Rusya dışişleri bakanı Sergey Lavrov NATO ilişkileri konularında açıklamalarda bulundu." kaynak cümlesinden, "Bulgaristan dışişleri bakanı Sergey Lavrov NATO ilişkileri konularında açıklamalarda bulundu." şeklinde bir cümle üretilebilmektedir. Bu cümle her ne kadar gramer kuralları açısından doğru bir cümle olsa da anlam bakımından doğru değildir. Bu ve benzeri üretilen yanlış çözümlemeler anlambilim ile ilgili olduğundan mevcut çalışma kapsamında yanlış olarak değerlendirilmemiştir.

Üretilen cümlelerden katılımcılar tarafından hem gramatik olarak uygun, hem de anlamlı olarak değerlendirilen bazı örnekler aşağıda verilmiştir.

Tablo 7. Üretilen Örnek Cümleler

Müzakerelerde yeni fasılların açılmasını beklediklerini söylediler.
Bunların hepsinin değerini bilmek lazım.
Bugün tüm güne yayılan bir grevden bahsediyoruz..
Kurulacak hükümetin bu konudaki çalışmalarını hızlandırma imkanımız olacaktır.
Kılıçdaroğlu : Tabii buradan alacağımız derslerle geleceğimizi inşa ediyoruz.
Başkan Akaydın'ın soruşturmayı yönlendirdiği öğrenildi.
Erdoğan , cumartesi , mahkemede önemli bilgilere ulaşıldığını söyleyebilirim.

7. Sonuç

Bu çalışmada kullanılan yöntem ile diğer kullanılan yöntemlerden farklı olarak, üçlü kelime grupları şeklinde ayrılan gruplardan belirlenen kurallar çerçevesinde, art arda ekleme yöntemiyle anlamlı ve farklı cümleler üretilmesi gerçekleştirilmiştir. Bunun için temel olarak istatistiksel yaklaşım uygulanmış, kısmi ve basit düzeyde kural tabanlı yaklaşımdan yararlanılmıştır. Bigram ifadeler ile yapılan metin üretme işleminde, sayıca daha fazla yeni cümle üretilmesine rağmen, anlam açısından uygunluk seviyesinin daha düşük olduğu gözlenmiştir. Sondan iki bağlı trigram ifadeler kullanılarak daha az sayıda cümle üretilse de, daha kuvvetli anlamsal bağlar ile yeni cümleler üretilmiştir. Bu yöntem ile sistem başarımının daha yüksek seviyede olduğu tespit edilmiştir. Kelime gruplarının orjinal durumları ile direk seçilmesi çok sayıda cümlenin birbirine bağlanmasını zorlaştırmaktadır. Ek alan ifadeler için ayrı kurallar uygulayarak üretilen cümlelerin çeşitliliği artırılabilir. Ancak, her ilave kural için sistemin çalışma süresi buna paralel olarak artacaktır.

Sonraki Çalışmalar

Mevcut çalışmada, hazır bir derlem içinde uygun biçimlerdeki kelime grupları ile yeni metinler oluşturulması hedeflenmiştir. Ancak, çalışma sadece bu derlem ile kısıtlı olduğundan üretilen metin sayısı da kısıtlı olmaktadır. Sonraki çalışmalarda, farklı derlemler için aynı kurallar geliştirilmeye çalışılacaktır. Özellikle, Latin dil ailesine uygun diller için kullanılabilecek olan bu yöntem, yapay sinir ağları veya derin sinir ağları yardımıyla, anlamlı ve daha fazla konu bütünlüğü olan, birbirinden farklı, çok sayıda metin üretilebilir. Özellikle son zamanlarda LSTM yönteminin cümle üretme konusundaki başarımı bilinmektedir. Tüm bu yöntemlerin uygulanmasında, sistemin eğitilmesi için geçen süre ayrıca dikkate alınmalıdır. Hangi durumlarda daha hızlı ve daha verimli sonuç elde edildiği tespit edilmelidir.

Teşekkür

Çalışmada verdikleri destekten ötürü Türk Hava Yolları Ar-Ge grubuna, anketi dolduran tüm katılımcılara ve kaynaklarından faydalandığım tüm akademik yayın yazarlarına, katkılarından ötürü teşekkür ederim.

Kaynakça

[1] Akalın, Ş.H. 2007. Türk Dünyasında Bilgisayar Destekli Dil Bilimi Çalışmaları ve Türk Dil Kurumu. 38. Uluslararası Asya ve Kuzey Afrika Çalışmaları Kongresi (ICANAS), 10-15 Eylül, Ankara, 17-22.

[2] Oflazer, K. ve Bozşahin, H.C. 2006. Türkçe Doğal Dil İşleme. Çukurova Üniversitesi Türkoloji Araştırmaları Merkezi.

[3] Korkmaz, T. 1996. Turkish Text Generation with Systemic-Functional Grammar. Master's Thesis, Bilkent University, Department of Computer Engineering and Information Science, Ankara.

[4] Gündoğdu, Ö.E. ve Duru, N. 2016. Türkçe Metin Özetlemede Kullanılan Yöntemler. 18. Akademik Bilişim Konferansı, Adnan Menderes Üniversitesi, 30 Ocak-5 Şubat, Aydın.

[5] Kazkılınç, S. 2013. Türkçe Metinlerin Etiketlenmesi. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

[6] Mocan, Z. 2005. Metin İşleme:Soru Soran Bir Sistem Tasarımı. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

[7] Bilgin, M. ve Amasyalı, M.F. 2017. Dependency parsing with stacked conditional random fields for Turkish. Journal of the Faculty of Engineering and Architecture of Gazi University, 32(2), 385-392.

[8] Adalı, Ş. ve Erenler, Y. 2003. Türkçe için Okuma Fonksiyonlu Otomatik Metin Oluşturma Sistemi. Elektrik-Elektronik-Bilgisayar Mühendisliği 10. Ulusal Kongresi, İstanbul Sayfa:484-487.

[9] Özdemir, C.B. ve Amasyalı, M.F. 2010. Hayat Bilgisi Veritabanı Kullanılarak Otomatik Cümle Üretimi. XV. Türkiye'de İnternet Konferansı, 2-4 Aralık, İstanbul, cilt.1 s.1-4.

[10] Metin, Senem Kumova, and Bahar Karaoğlan. 2010. Collocation extraction in Turkish texts using statistical methods. International Conference on Natural Language Processing. Springer, Berlin, Heidelberg.

[11] Can, Burcu. 2019. LSTM Ağları ile Türkçe Kök Bulma. *Bilişim Teknolojileri Dergisi*, *12*(3), 183-193.

[12] Hochreiter, S., Schmidhuber, J. 1997. "Long Short-Term Memory", Neural Computation 9 (8), pp. 1735–1780.

[13] Brown, P.F. et al. 1990. Class-Based N-Gram Models of Natural Language. In Proceedings of the IBM Natural Language ITL, Paris, pp. 283–298.

[14] Mairesse, F. et al. 2010. Phrase-based Statistical Language Generation using Graphical Models and Active Learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), Uppsala.

[15] Uchimoto, K. et al. 2002. Text Generation from Keywords. COLING '02 Proceedings of the 19th International Conference on Computational Linguistics, Taipei.

[16] Tan, J. et al. 2017. From neural sentence summarization to headline generation: a coarse-to-fine approach. 26th International Joint Conference on Artificial Intelligence (IJCAI-17), 19-25 August, Melbourne-Australia, p.4109-4115.

[17] Bauer, A. et al. 2015. Rule-based Approach to Text Generation in Natural Language-Automated Text Markup Language. (ATML3), Challenge+ DC@ RuleML.

[18] Kutlugün, Mehmet Ali, and Şirin, Yahya. 2018. Turkish meaningful text generation with class based n-gram model. 26th Signal Processing and Communications Applications Conference (SIU). IEEE. DOI: 10.1109/SIU.2018.8404801.

[19] Yıldırım, O. ve Atık, F. 2013. Kişisel Gazete, Bitirme Projesi. Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, İstanbul.

[20] Erdogan, H. 2001. Speech Recognition for a Travel Reservation System. International Conference on Artificial Intelligence (IC-AI'2001), 25-28 Jun, Las Vegas-Nevada.
[21] Manning, C.D. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.
[22] Justeson, J.S. and Katz, S.M. 1995. Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. Computational Linguistics.
[23] Eryiğit, Gülşen, Joakim Nivre, and Kemal Oflazer. 2006. The incremental use of morphological information and lexicalization in data-driven dependency parsing. International Conference on Computer Processing of Oriental Languages. Springer, Berlin, Heidelberg.