METİN ÖZETLEME İÇİN CÜMLE SEÇİM METOTLARI
SENTENCE SELECTION METHODS FOR TEXT SUMMARIZATION

Aysun Güran[1], Sümeyra Nur Arslan[2], Esma Kılıç[2], Banu Diri[2]

1. Bilgisayar Mühendisliği Bölümü
Doğuş Üniversitesi
adogrusoz@dogus.edu.tr

2. Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
{snurarslan, esmakilic92}@gmail.com,banu@ce.yildiz.edu.tr

ÖZETÇE

***Özetçe* — Bu çalışmanın amacı, bir dokümandaki en önemli cümleleri seçerek ilgili dokümanın özetini çıkarmaktır. Bu amaçla 15 farklı cümle seçim metodu kullanılmıştır. Bu metotlar, 15 kadın ve 15 erkek olmak üzere, toplam 30 kişi tarafından çıkarılmış özet dokümanlarının oluşturduğu bir değerlendirme veri seti üzerinde kıyaslanmıştır. Ayrıca, bu metotlardan en başarılı olanlarının birlikte kullanılması ile elde edilen farklı özellik gruplarının başarım değerleri sergilenmiş ve analiz sonuçları paylaşılmıştır.**

Anahtar Kelimeler — metin özetleme; cümle seçim metotları; varlık ismi tanıma.

ABSTRACT

Abstract—**The aim of this work is to create text summaries by selecting the most important sentences of documents. For this aim 15 sentence selection methods are used. These methods are compared on the evaluation set created by 15 women and 15 men evaluators. The performance results of the systems that are obtained by using different sentence selecetion methods together are also analyzed and the results are shared.**

Keywords — text summarization; sentence selection methods; name entity recognition.

1. GİRİŞ

Özet, bir veya aynı konu ile ilgili birden fazla dokümandan çıkarılan ve kaynağındaki en temel bilgiyi içeren metin parçasıdır. Bir özet dokümanın bilgisayar tarafından otomatik olarak çıkartılması ise "Otomatik Metin Özetleme (OMÖ)" olarak adlandırılır. OMÖ işleminde özetlenecek doküman sayısına göre "tekli" veya "çoklu" doküman özetleme işlemlerinden bahsetmek mümkündür. Tekli doküman özetlemede bir tane kaynak doküman mevcutken, çoklu doküman özetlemede birbirleri ile ilgili olan birden fazla kaynaktan yararlanılmaktadır. Özetleme sisteminin çıktısı "yoruma" ya da "çıkarıma" dayalı olan bir özet olabilir. Yoruma dayalı olan özetlemede orijinal metindeki ifadeler kısaltılarak tekrar yazılmaya çalışılır. Çıkarıma dayalı olan özetlemede ise özetlenecek dokümandaki cümleler değiştirilmeksizin seçilmektedir. Bir özet "genel" ya da "kullanıcıya yönelik" olabilir. Bu iki kavram özetin etki alanı ile ilgilidir. Genel özet, metnin ana temalarıyla ilgili olan ayrıntılı özettir. Kullanıcıya yönelik özet ise kullanıcının yazdığı sorgu ile ilgili olan özettir. Bu çalışmada çıkarıma dayalı ve genel bir doküman özetleme sistemi üzerinde durulmuştur.

Verilen bir doküman içerisindeki en önemli cümlelerin seçilmesi adına kullanılan bazı metotlar mevcuttur. Bu metotları kullanan ve İngilizce dokümanlar üzerinde çalışan ilk yol gösterici çalışma Luhn'nın [1] çalışmasıdır. Bu çalışmada cümleler terim frekanslarına göre puanlandırılmıştır. Edmunson [2], Luhn'nın [1] çalışmasındaki kelime sıklığı bilgisine ek olarak "ipucu sözcük öbekleri", "başlık terimleri" ve "cümle konumu" gibi üç yeni özellik daha kullanmıştır. Literatürde bir cümlenin önemini tespit etmek adına sıklıkla kullanılmış olan diğer özellikler: "cümle uzunluğu", "ünlem, soru işareti ya da tırnak işareti gibi vurgu belirten bazı noktalama işaretleri", "tarih bilgisini belirten ifadeler", "önemli N-gramlar", "doküman içindeki isimler ya da nümerik karakterler" gibi özelliklerdir. Bu özelliklerin kullanımı [3]'deki tez çalışmasında ayrıntılı bir şekilde anlatılmıştır. [4-9]'da ise çıkarıma dayalı özetleme konusunda Türkçe dokümanlar üzerinde çalışılmıştır.

Bu çalışmanın örnek çalışmalardan [4],[5] farkı, Türkçe özetlerin çıkarılmasında fazla sayıda cümle seçim metodunun kullanılmış olması ve bu metotların başarımlarının değerlendirilmesinde birden fazla değerlendiricinin çıkarmış olduğu özetlerin dikkate alınmasıdır. Cümle seçim metotlarından "varlık ismi tanıma metodu" ilk kez bu çalışma ile Türkçe metinler üzerinde analiz edilmiştir. Yapılan analizler 15'i kadın, 15'i erkek olmak üzere 30 farklı değerlendiricinin özetini çıkardığı 20 doküman üzerinde yapılmıştır.

Bildirinin ikinci bölümünde çıkarıma dayalı olan metin özetlemede cümle seçimi için kullanılan metotlar anlatılmıştır; üçüncü bölümde veri seti tanıtılmış ve cümle seçim özelliklerinin bu veri seti üzerindeki başarım değerleri kıyaslanmıştır. Son olarak, sonuçlar bölümünde analizler özetlenmiş ve gelecek çalışmalardan bahsedilmiştir.

2. CÜMLE SEÇİM METOTLARI

Çalışmamızın ana mantığı doküman içerisinde geçen cümleleri farklı metotlara göre puanlandırmak ve en yüksek puana sahip olan cümleleri seçerek özet dokümanlarını oluşturmaktır.

Tablo 1: Cümle seçim metotları

Grup	Cümle Seçim Metotları
Yapısal Özellik	$ö_1$: Cümle konumu
	$ö_2$: Cümle uzunluğu
Benzeme Özelliği	$ö_3$: İlk cümleye benzerlik
	$ö_4$: Son cümleye benzerlik
	$ö_5$: Başlığa olan benzerlik
	$ö_6$: Toplam benzerlik
	$ö_7$: Ortak yakınlık sayısı
Kelime Sıklığı ve Dağılımı	$ö_8$: Dağılımsal özellik
	$ö_9$: Kelime sıklığı bilgisi
Özel Belirteçleri İçerme Durumu	$ö_{10}$: Sayısal karakter içerme
	$ö_{11}$: "?" ve "!" içerme
	$ö_{12}$: Pozitif kelimeleri içerme
	$ö_{13}$: İsim soylu kelimleri içerme
	$ö_{14}$: Varlık ismi içerme
Anlamsal Özellik	$ö_{15}$: Gizli anlamsal analiz

Puanlandırmada kullanılan metotlar Tablo 1'de belirtilen gruplar altında toplanmıştır. Metotlar hakkında ayrıntılı açıklamalar ise aşağıda belirtilmiştir:

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

ö1-Cümle Konumu: Çalışmamızda dokümanı oluşturan her bir cümleye cümlenin konumuna göre eşitlik (1) ile belirtilen skor değeri verilmiştir:

$$Skor_{ö_1}(C_i) = \frac{N - P_i}{N} \quad (1)$$

Burada N dokümandaki toplam cümle sayısı iken, P_i değeri cümlenin doküman içinde kaçıncı cümle olduğunu belirtmektedir.

ö2-Cümle Uzunluğu:
Bu özellik, dokümandaki her bir cümleye cümlenin sahip olduğu kelime sayısına göre bir skor değeri vermektedir.

ö3- İlk Cümleye olan Benzerlik:
Bu özellik dokümandaki her bir cümleye, cümlenin dokümanın ilk cümlesine olan benzerliğine göre bir skor değeri vermektedir. Cümleler arasındaki benzerlik kosinüs benzerliğine göre hesaplanmaktadır.

$$Skor_{ö_3}(C_i) = \text{kosinüs}\quad (C_i, C_{ilk}) \quad (2)$$

ö4 - Son Cümleye olan Benzerlik:
Bu özellik dokümandaki her bir cümleye, cümlenin dokümanın son cümlesine olan benzerliğine göre bir skor değeri vermektedir.

$$Skor_{ö_4}(C_i) = \text{kosinüs}\quad (C_i, C_{son}) \quad (3)$$

ö5 - Başlığa olan Benzerlik:
Bu özellik dokümandaki her bir cümleye, cümlenin dokümanın başlığına olan benzerliğine göre bir skor değeri vermektedir.

$$Skor_{ö_4}(C_i) = \text{kosinüs}\quad (C_i, baslik) \quad (4)$$

ö6 - Toplam Benzerlik:
Toplam benzerlik dokümanda bulunan bir cümlenin, diğer cümlelere olan kosinüs benzerliklerinin toplamıdır.

ö7 –Ortak Yakınlık Sayısı:
Ortak yakınlık sayısı, i ve j birbirinden farklı iken aşağıdaki gibi hesaplanır:

$$Skor_{ö_7}(C_i) = \sum_{i=1}^{N} yakinlik(C_i, C_j) = \sum_{i=1}^{N} \frac{C_i(yakinlari) \cap C_j(yakinlari)}{C_i(yakinlari) \cup C_j(yakinlari)} \quad (5)$$

C_i (yakınları) ifadesi C_i cümlesine benzerlik değeri belirli bir sınırın üzerinde olan cümleler listesinin elaman sayısıdır. C_j (yakınları) ifadesi C_j cümlesine benzerlik değeri belirli bir sınırın üzerinde olan cümleler listesinin elaman sayısıdır. C_i(yakınları)$\cap C_j$(yakınları) ifadesi ise benzerlik listesinin kesişimi olan listenin eleman sayısıdır.

ö8- Dağılımsal özellikler:
Referans [10]'nun doküman sınıflama için önerdiği dağılımsal özellikler ile bir dokümanda eşit sıklıkta geçen terimler, dokümanın içindeki yayılma durumlarına göre birbirinden farklılaştırılabilmektedir. Biz bu çalışmada, [10] ile belirtilen üç dağılımsal özelliği toplayarak elde ettiğimiz değeri metin özetlemede bir cümle puanlama metodu olarak kullandık. Referans [10]'nun önerdiği üç dağılımsal özellik "bölüm yoğunluğu", "ilk ve son konum yoğunluğu", "pozisyonların varyansı" özellikleridir. Bu özellikler hakkındaki bilgiler aşağıda belirtilmiştir:

Bölüm yoğunluğu: Bir terimin, bir doküman içindeki bölüm yoğunluğu eşitlik (6) ile hesaplanmaktadır.

$$Y_{Bölüm}(t,d) = \sum_{i=0}^{n-1} c_i > 0 \,?\, 1 : 0 \quad (6)$$

Burada c_i değeri, terim t'nin i indeksli cümle içindeki geçme sıklığı bilgisidir.

İlk ve Son Konum Yoğunluğu: Bu özellik ile bir terimin bir doküman içindeki ilk ve son konumları arasındaki fark alınmıştır. İlk ve son konum farkı özelliği aşağıdaki şekilde hesaplanır:

$$Y_{Ilk-Son}(t,d) = Son_{Gorunum}(t,d) - İlk_{Gorunum}(t,d) \quad (7)$$

$$İlk_{Gorunum}(t,d) = \min_{i \in \{0..n-1\}} c_i > 0 \,?\, i : n \quad (8)$$

$$Son_{Gorunum}(t,d) = \max_{i \in \{0..n-1\}} c_i > 0 \,?\, i : -1 \quad (9)$$

Burada c_i değeri, terim t'nin i indeksli cümle içindeki geçme sıklığı bilgisidir. n değeri ise terimin bulunduğu cümle indeks bilgisidir.

***Pozisyonların Varyansı*:**
Bu özellikte incelenen terimin yoğunluk hesabı için terimin tüm görünüşlerinin varyansları kullanılmıştır. İlk önce incelenen terimin tüm görünüşlerinin ortalaması alınmıştır. Daha sonra her bir görünüşün ortalama görünüşten farkı alınmış ve elde edilen ortalama konum bilgisi pozisyonların varyansı olarak isimlendirilmiştir:

$$Y_{PozVar}(t,d) = \frac{\sum_{i=0}^{n-1} c_i |i - merkez(t,d)|}{say(t,d)} \quad (10)$$

$$say(t,d) = \sum_{i=0}^{n-1} c_i \quad ve \quad merkez(t,d) = \frac{\sum_{i=0}^{n-1} c_i * i}{say(t,d)} \quad (11)$$

Eşitlik (6-11) ifadelerinin daha iyi anlaşılması adına referans [10]'da belirtilen sayısal örnek incelenmelidir.

Çalışmamızda bir dokümanı oluşturan her bir terime ait 3'er dağılımsal özellik değeri (Bölüm Yoğunluğu, İlk ve Son Konum Yoğunluğu, Pozisyonların Varyansı) bulunduktan sonra bu değerlerin ortalaması alınarak, her terimin toplam dağılımsal özellik değeri bulunmuştur.

$$T_{Dagitimsal}(t,d) = \left(\frac{Y_{Bölüm}(t,d) + Y_{ilk-son}(t,d) + Y_{pozVar}(t,d)}{3}\right) \quad (12)$$

Dokümanı oluşturan cümlelere puan vermek için dokümandaki her terimin $T_{Dağıtımsal}$ değerinin toplamı bulunmuş ve bu toplam cümlelere atanmıştır.

ö9 - Kelime Sıklığı Bilgisi:
Bu özellik ile doküman içerisinde yer alan her terimin frekansı hesaplanır ve cümlelere içerdiği terimlerin frekans bilgilerinin toplanmasıyla bir skor değeri verilmektedir.

ö10 – Sayısal Karakter İçerme Durumu:
Bu özellik ile cümlelere içerdikleri sayısal karakter sayısına göre bir puan verilmektedir.

ö11- "?" ve "!" İçerme Durumu:
Bir cümlenin ünlem işareti veya soru işareti ile bitmesi diğer cümlelere göre daha önemli olduğunun bir işaretidir. Buna göre eğer cümle soru işareti ya da ünlem işareti içeriyorsa cümleye bir puan verilir.

ö12- Pozitif Kelimeleri İçerme Durumu:
Bu özellik ile cümlelerin "özetle", "sonuçta", "neticede" gibi toparlayıcı kelimeleri içerip içermediği incelenir. Cümlelere içerdikleri toplam pozitif kelime sayısı kadar puan verilir.

ö13- İsim Soylu Kelimeleri İçerme Durumu:
Metinlerde yer alan isimler, metnin içeriği hakkında bilgi vermektedir. Bu yüzden metin özetleme sistemi isimlerin geçtiği cümlelere sahip olduğu isim sayısı kadar puan vermektedir. Metinler içindeki terimlerin isim olup olmadığı Zemberek yazılımı [12] kullanılarak tespit edilmiştir.

ö$_{14}$ –Varlık isimlerini İçerme Durumu

Varlık ismi tanıma, doğal dil işleme biliminin önemli alanlarından biri olup, dokümanlarda geçen isimleri kişi, yer ve organizasyon ismi olarak ayırmanın yanı sıra formül, tarih ve parasal ifadeleri de bulabilmeyi hedeflemektedir. Çalışmamızda ilk önce veri setinde bunan her bir doküman içindeki varlık isimleri tespit edilmiştir. Varlık isimleri tespit edilirken [13]'de belirtilen algoritma kullanılmıştır. Varlık isimleri tespit edildikten sonra cümle içinde geçen varlık isimleri sayısına göre her bir cümleye bir puan verilmiş ve yüksek puanlı cümleler özete eklenmiştir. Bu çalışma ile ilk kez varlık isimlerinin Türkçe doküman özetleme konusu üzerindeki etkisi analiz edilmiştir.

ö$_{15}$ –Gizli Anlamsal Analize Dayalı Puanlama

Gizli Anlamsal Analiz (GAA) sistem girdisi olarak bir metnin içeriğini almakta ve metin içindeki terimler ile cümleler arasındaki gizli anlamsal ilişkiyi ortaya çıkarmaktadır. GAA'nın sistem girdisi bir terim-cümle matrisidir (A matrisi). GAA'de A matrisinin Tekil Değer Ayrışımı (TDA) gerçekleştirilir ve matris üç çarpana ayrılır ($A=USV^T$). [11]'de belirtilen çalışmada bu çarpan matrislerine dayalı bir metin özetleme sistemi önerilmiştir. Bu çalışmada eşitlik (13)'deki, B matrisi oluşturulduktan sonra, bu matrisin hücre değerleri ile oluşturulmuş olan S_k (14) değerlerine göre her bir cümleye bir puan değeri verilmiş ve yüksek puanlı cümleler seçilerek özete eklenmiştir.

$$B = S^2 V^T \tag{13}$$

$$S_k = \sqrt{\sum_{i=1}^{r} b_{ik}^2} \tag{14}$$

3. VERİ SETİ VE PERFORMANS DEĞERLENDİRMESİ

Cümle seçim metotlarının başarım değerleri gazetelerden toplanan 20 haber dokümanı üzerinde değerlendirilmiştir. Bu veri seti ile ilgili istatistikler Tablo 2 ile belirtildiği gibidir. Bu veri setinin hazırlanmasındaki amaç cümle seçim metotlarının istikrarını göstermektir. Bu nedenle 20 dokümanı içeren değerlendirme seti 15'i kadın ve 15'i erkek olmak üzere toplam 30 farklı kişiye verilmiş ve kişilerden doküman özetlerinin %35'lik bir özetleme oranı çıkarılması istenmiştir.

Çalışmamızda cümle seçim metotlarının performansı eşitlik (15) ile ifade edilmiştir. Burada S değeri cümle seçim metotları ile seçilen cümle sayısını ifade ederken, T değeri değerlendiriciler tarafından seçilmiş olan cümle sayısını belirtmektedir.

$$performans = \frac{|S \cap T|}{|S|} \tag{15}$$

Tablo 2: Veri seti istatistikleri

Değerlendirme veri setine ait olan istatistikler	
Veri Setindeki Toplam Doküman Sayısı	20
Veri Setindeki Toplam Cümle Sayısı	201
Veri Seti İçindeki Dokümalardan en az Cümleye Sahip olan Dokümanın Cümle Sayısı	7
Veri Seti İçindeki Dokümalardan en fazla Cümleye Sahip olan Dokümanın Cümle Sayısı	13

Cümle seçim metotlarının kullanımı sırasında dokümanlardaki kelimeler Zemberek yazılımı [12] ile köklerine ayrılmış ve dokümanlar durak kelimelerinden arındırılmıştır.

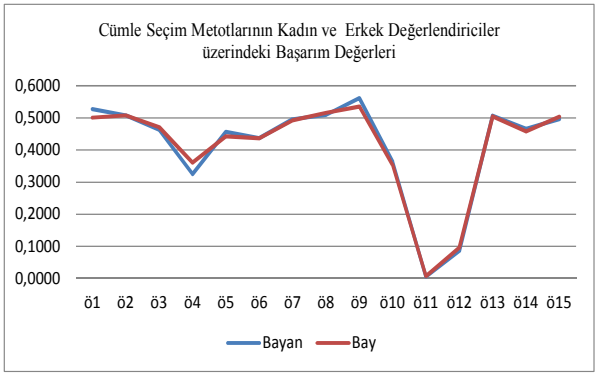
Tablo 3, çalışmamızda kullanılan cümle seçim metotlarının 30 farklı değerlendiriciye göre hesaplanan ortalama başarım değerlerini azalan sırada göstermektedir. Bu tablo aynı zamanda metot performanslarının 30 değerlendirici bazındaki standart sapma değerlerini de içermektedir.

Tablo 3: Cümle seçim metotlarının ortalama performansları ve standart sapma değerleri

Metotlar	**Ortalama Performans**	**Standart Sapma**
ö$_9$	0.547	0.072
ö$_1$	0.517	0.076
ö$_8$	0.499	0.053
ö$_{13}$	0.498	0.064
ö$_2$	0.495	0.065
ö$_7$	0.490	0.043
ö$_{15}$	0.488	0.066
ö$_3$	0.462	0.045
ö$_{14}$	0.459	0.049
ö$_5$	0.448	0.052
ö$_6$	0.431	0.034
ö$_{10}$	0.351	0.055
ö$_4$	0.337	0.054
ö$_{12}$	0.087	0.019
ö$_{11}$	0.007	0.006

Tablo 3'ten görüldüğü üzere en başarılı cümle seçim metodu "ö$_9$-kelime frekans" bilgisidir. Değerlendiriciler, sıklıkla en çok geçen kelimeleri içeren cümleleri seçmişlerdir. Bu özelliği "ö$_1$-cümle konumu", "ö$_8$-dağılımsal özellikler", "ö$_{13}$-isim soylu kelimeler" ve "ö$_2$-cümle uzunluğu" izlemektedir. Tablo 3'deki dikkat çekici noktalardan biri, metotların standart sapma değerlerinin düşük olmasıdır. Bu durum metotların farklı kişiler üzerindeki etkilerinin çok fazla değişkenlik taşımadığını göstermektedir. Tablo 3'de dikkat çeken noktalardan bir diğeri de Türkçe dokümanlar üzerindeki etkisi ilk kez bu çalışma ile analiz edilen varlık isimi tanıma metodunun (ö$_{14}$) birçok özelliği geçmesidir. Bu durum sonucunda varlık ismi tanıma özelliğinin Türkçe doküman özetleme üzerindeki etkisinin önemli olduğu söylenebilir. Varlık ismi tanıma sisteminin Türkçe dokümanlar üzerindeki başarım sonucu iyileştikçe, özelliğin metin özetleme üzerindeki etkisi daha da önem kazanacaktır.

Şekil 1, cümle seçim metotlarının kadın ve erkek değerlendiriciler üzerindeki başarım değerlerini göstermektedir.



Şekil 1: Cümle seçim metotlarının kadın ve erkek değerlendiriciler üzerindeki başarım değerleri

Şekil 1 ile belirtilen analiz sonuçlarına göre cümle seçim özelliklerinin kadın ve erkek değerlendiriciler üzerindeki ortalama başarım sonuçlarının hemen hemen aynı olduğu görülmektedir. Bu durum cinsiyetten bağımsız bir şekilde kişilerin özet çıkarırken dikkat ettikleri özelliklerin benzer olduğunu göstermektedir.

Son olarak cümle seçim metotlarını tek tek kullanmak yerine, metotların birleşimi ile elde edilen farklı grupların başarım değerleri analiz edilmiştir. Bu amaçla ortalama başarım sonucu 0.4'ün üzerinde olan 11 adet cümle seçim metodu ($ö_1$, $ö_2$, $ö_3$, $ö_5$, $ö_6$, $ö_7$, $ö_8$, $ö_9$, $ö_{13}$, $ö_{14}$, $ö_{15}$) farklı şekillerde (2'li, 3'lü vs.) birleştirilmiş ve pek çok deney yapılmıştır. Gruplar oluşturulurken cümle seçim metotlarının skor değerleri 0-1 aralığına getirilerek normalize edilmiş ve skor değerleri toplanmıştır. Başarım sonuçları değerlendirilirken cümle seçim metotlarının, üzerinde en başarılı olduğu değerlendiricinin çıkartmış olduğu özet dokümanları dikkate alınmıştır. Deneyler sonucunda elde edilen *en başarılı* gruplar Tablo 4 ile sergilenmiştir.

Tablo 4 : Birleştirilmiş metotlardan oluşan farklı sistem performansları

Özellik Grupları	**Ortalama Performans**
$ö_1$,$ö_{14}$	0.6983
$ö_1$,$ö_7$,$ö_{13}$	0.7383
$ö_1$,$ö_3$,$ö_9$,$ö_{13}$	0.7133
$ö_1$,$ö_3$,$ö_9$,$ö_{11}$,$ö_{13}$	0.7133
$ö_1$,$ö_3$, $ö_7$,$ö_9$,$ö_{11}$,$ö_{13}$	0.7008
$ö_1$,$ö_3$,$ö_7$,$ö_9$,$ö_{11}$,$ö_{13}$,$ö_{15}$	0.6967
$ö_1$,$ö_2$,$ö_3$,$ö_5$,$ö_6$,$ö_7$,$ö_8$,$ö_{13}$	0.6575
$ö_1$,$ö_2$,$ö_3$,$ö_5$,$ö_6$,$ö_7$,$ö_8$,$ö_{12}$, $ö_{13}$	0.6575
$ö_1$,$ö_2$,$ö_3$,$ö_5$,$ö_6$,$ö_7$,$ö_8$, $ö_{11}$,$ö_{12}$, $ö_{13}$	0.6575
$ö_1$,$ö_2$,$ö_3$,$ö_5$,$ö_6$,$ö_7$,$ö_8$,$ö_9$,$ö_{13}$,$ö_{14}$,$ö_{15}$	0.6675

Tablodan görüldüğü üzere en başarılı grup $ö_1$-cümle konumu, $ö_7$-ortak yakınlık sayısı ve $ö_{13}$-isim soylu kelimeleri içerme durumu metotlarının birleşiminden oluşan ikinci sıradaki gruptur. 3'lü gruplardan sonra başarım yüzdesinin düştüğü görülmektedir. Bazı durumlarda gruba yeni özelliğin katılması sistem performansını değiştirmemiştir.

Özelliklerin bireysel başarımlarını gösteren Tablo 3'e göre en başarılı ilk iki bireysel özellik $ö_1$ ve $ö_9$'dur. Ancak yapılan 2'li grup değerlendirmelerine göre bu iki özelliğin toplanması diğer ikili grupların başarım oranlarını geçememiştir. Başka bir deyişle, $\binom{11}{2}$ adet ikilinin performans değerlendirmesi sonucu en iyi başarım sonucunu veren ikili grup $ö_1$,$ö_9$ değil, Tablo 4'den görüleceği üzere $ö_1$,$ö_{14}$ olmuştur. Benzer durumlar diğer grup deneyleri için de geçerlidir. Bu durumunun sonucu olarak, ileriki çalışmalarda sezgisel yöntemlerin kullanımıyla metotların toplanması aşamasında bazı ağırlık değerlerinin kullanılması planlanmaktadır. Böylelikle bireysel performansları düşük olan özellikler ile yüksek performanslı özelliklerin belirli yüzdelerle birleştirilmesi daha yüksek başarım değerlerinin elde edilmesini sağlayabilir.

4. SONUÇLAR VE GELECEK ÇALIŞMALAR

Bu çalışmada çıkarıma dayalı olan bir metin özetleme sisteminde kullanılabilecek cümle seçim metotları incelenmiştir. Bu metotlar, 20 haber dokümanı ve bu haber dokümanlarını özetleyen 15'i kadın ve 15'i erkek olmak üzere toplam 30 kişinin çıkarmış oldukları özet dokümanını içeren bir değerlendirme veri seti üzerinde analiz edilmiştir. Analiz sonuçlarına göre 30 kişi üzerinde en başarılı olan metotlar paylaşılmıştır. Çalışmamız ile ilk kez varlık isimleri tanıma özelliğinin etkisi Türkçe dokümanlar üzerinde test edilmiş ve bu özellik ile başarılı sonuçlar elde edildiği gösterilmiştir.

Çalışmamız ile cümle seçim metotlarının bireysel performanslarının dışında bu metotların birleşimiyle elde edilen farklı grupların başarım sonuçları da analiz edilmiştir. Analiz sonuçlarına göre en başarılı olan gruplar sergilenmiştir. Bu deneylerde cümle seçim metotları eşit önem derecelerine sahip olacak şekilde birleştirilmiştir. İleriki çalışmamızda bu özellikler farklı ağırlık değerleri ile çarpılarak, özelliklerin sahip olması gereken ağırlık değerleri sezgisel metotlar ile otomatik olarak buldurulacaktır.

5. KAYNAKÇA

[1] Luhn, H.R., (1958). "The automatic creation of literature abstracts.", IBM Journal of Research Development, 2(2):159–165.

[2] Edmundson, H.P., (1969). "New methods in automatic extracting.", Journal of the Association for Computing Machinery, 16(2): 264–285.Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear", *IEEE Trans. Speech and Audio Proc*., 7(6):697-708, 1999.

[3] Güran, A. (2013), "Metin Özetleme Sistemi", Doktora Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü.

[4] Kutlu, M., Cığır, C. and Cicekli,I., (2009). "GenericSummarization for Turkish", ISCIS 2009, Kıbrıs.

[5] Uzundere, E., Dedja, E., Diri, B., Amasyalı, M.F. (2008), "Türkçe Haber Metinleri için Otomatik Özetleme", In Proceedings of ASYU 2008, Isparta, Türkiye.

[6] Pembe, C. (2011). Automated Query-Biased and Structure-Preserving Document Summarization for Web Search Tasks, PhD Thesis, Boğaziçi University, Turkey.

[7] Güran, A., Güler, N. and Bekar, E., (2011). "Automatic summarization of Turkish documents using non-negative matrix factorization ", INISTA 2011, Istanbul, Türkiye.

[8] Güran, A., Güler Bayazıt and N., Gürbüz, M.Z., (2013),"Efficient feature integration with Wikipedia based semantic feature extraction for Turkish text summarization", Turkish Journal of Electrical Engineering and Computer Science.

[9] Güran, A., Güler, N., Bekar, E. "LSA-based Turkish Text Summarization with Consecutive Words Detection", CSIE 2011, Changchun, Çin.

[10] Xue, X.B. and Zhou, Z.H., (2009). "Distributional Features for Text Categorization," IEEE Trans. Knowledge and Data Eng., 21(3): 428–444.

[11] Steinberger, J., (2007). Text Summarization within the LSA Framework, PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, January 2007.

[12] Zemberek Projesi Geliştirme Sayfaları, https://zemberek.dev.java.net/, 1 Haziran 2009.

[13] Şeker, G.A., Eryiğit. G. (2012). Initial explorations on using CRFs for Turkish Named Entity Recognition. In Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 8-15 December 2012.