

Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Paragraf Tabanlı Çıkarımsal Özetlemede Öbekleme Kullanan İki Yöntemin Kıyaslanması

Ahmet İlkay KISAYOL [a,*], Metin TURAN [b]

[a] *Bilgisayar Mühendisliği Bölümü, Fen Bilimleri Enstitüsü, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İstanbul Ticaret Üniversitesi, İstanbul, TÜRKİYE*
** Sorumlu yazarın e-posta adresi: ilkaykisayol@gmail.com*

ÖZET

Özetleme, bir bakıma metinleri kısaltma işlemidir. Bu kısaltma işlemi metinlerdeki önemli bilgileri içerecek şekilde olmalıdır. Bu çalışmanın amacı da İngilizce dilinde yazılmış makale, haber vs. gibi doküman paragraflarının içerdiği bilgi önemine göre seçilerek özetleme yapılmasıdır.

Çalışmanın ilk aşamasında doküman kümesini temsil edecek önemli kelimeler belirlenmiştir. Bu aşamada tüm dokümanlarda geçen kelimeler kök geçiş sıklıklarına göre büyükten küçüğe göre sıralanır ve belirli sayıda seçilen en sık kelimeler ile paragraf vektörü temsil edilir.

Bir sonraki aşamada, istenilen özet oranına göre paragraflar kümelere ayrıştırılır. Kümeleme algoritması olarak K-Means kullanılmıştır. Kümeler oluşturulurken başlangıç noktalarının belirlenmesi amacıyla iki farklı yöntem kullanılmıştır. Bunlardan birincisi geçiş sıklıkları hesaplanan kelimelerden ilk 10'u seçilerek bu anahtar kelimelerin en çok geçtiği paragraflar seçilir. İkinci yöntemde kullanıcının belirlediği özet oranına göre seçilecek anahtar kelime sayısı belirlenir. Daha sonra bu anahtar kelimelerin en çok geçtiği paragraflar başlangıç noktaları olarak belirlenir. Özet oluşturmada çıkarım yöntemi olarak oluşturulmuş olan her bir kümeden, kümelerin merkez noktasına Jaccard uzaklığı bakımından en yakın olan paragraf seçimi uygulanmıştır. Çıkan sonuçlar kontrol edildiğinde ikinci yöntemin daha başarılı bir sonuç verdiği gözlemlenmiştir. İkinci yönteme göre başarı oranları %20 özet oranı için %40 , %40 özet oranı için %50 ve %60 özet oranı için %71 elde edilmiştir.

Anahtar Kelimeler: *çoklu dokümanlarda özetleme, paragraf tabanlı özetleme, metin öbekleme, özellik çıkarımı*

The Two New Methodology Comparison Using Paragraph Based Inferential Abstraction

ABSTRACT

Summarization is a process of abbreviation of a text. This abbreviation should be such that it contains important information in the texts. The purpose of this study is selecting according to the importance of the information contained in the document paragraphs in articles, news, etc.

Geliş: 25/04/2018, Düzeltme: 04/06/2018, Kabul: 13/06/2018

During the first phase of the study, important words to represent the document set were identified. At this stage, the words in all the documents are sorted according to the frequency of root passage order by ascending and the most frequently selected words and paragraph vector are represented at a certain number of times.

In the next step, the paragraphs are separated into clusters according to the desired summary ratio. K-Means was used as the clustering algorithm. Two different methods were used to determine the starting points when the clusters were constructed. The first is selected from the words calculated for the first 10 pass-through frequencies, and the paragraphs most frequently passed by these key words are selected. In the second method, the number of keywords is determined according to the summary rate determined by the user. Then the paragraphs most often passed by these keywords are set as starting points. The paragraph selection that is closest to the center point of the clusters in terms of Jaccard distance is applied from each set which is constructed as a subtraction method in the summarization. When the results were checked, it was observed that the second method gave a more successful result. Success rates according to the second method were 40% for the 20% summary rate, 50% for the 40% summary rate and 71% for the summary rate.

Keywords: *multiple document summarize, paragraph base summarization, text grouping, feature extraction*

I. GİRİŞ

Özetleme, tek veya aynı konuyu içeren bir veya birden fazla dokümandan çıkarılan ve en çok bilgiyi içeren metin parçalarını bulma ve sıralama işlemidir. Bu işlemin bilgisayar tarafından otomatik olarak yapılması "Otomatik Metin Özetleme (OMÖ)" olarak adlandırılır. OMÖ işleminde özetlenecek doküman sayısına göre "tekli" veya "çoklu" doküman özetleme olarak ayırabiliriz. Tekli doküman özetlemede bir tane doküman mevcutken, çoklu doküman özetlemede birbirleri ile aynı konuyu içeren birden fazla dokümandan yararlanılmaktadır. Özetleme işlemi "yoruma" ya da "çıkarıma" dayalı olarak yapılabilir. Yoruma dayalı özetleme de, özetlenecek doküman veya dokümanlardaki ifadeler kısaltılarak yeniden yazılır. Çıkarım ile yapılan özetleme de özetlenecek doküman veya dokümanlardaki, cümleler veya paragraflar seçilerek yapılır. Bu çalışmada çoklu dokümanlar da paragraf tabanlı çıkarıma dayalı özetleme sistemi üzerinde çalışılmıştır.

Otomatik metin özetleme konusunda ilk çalışma Luhn adlı bilim adamı tarafından 1959 yılında yapılmıştır [1]. Bu çalışmada doküman özetleme işlemi 2 aşamada gerçekleştirilmiştir. Bu aşamalardan ilkinde özetlenecek olan doküman ön işlemeden geçirilir. Ön işlemede doküman içinde konu ile ilgili bir anlam ifade etmeyen etkisiz kelimeler (zamir, bağlaç vb.) temizlenir. Kalan kelimeleri ortak bir payda da buluşturmak için 6 farklı harften az olanlar ve aynı ön eke sahip kelimeler aynı sözcük olarak kabul edilir. İkinci aşamada ön işlemden geçirilerek kümelenmiş olan kelimelerden sayıca az olanlar temizlenir. Böylece, fazla sayıda geçen (yani yüksek frekanslı) kelimeler belirlenir. Bu kelimeler anahtar kelime olarak kullanılır. Bu anahtar kelimeler kullanılarak cümleler puanlandırılır. Cümleler puanların göre sıralanarak, özete en yüksek puana sahip cümleler konulur. 1969 yılında Edmundson [2], 1995 yılında Brandow, Mitze ve Rau [3] frekans tabanlı çalışmalara devam etmişlerdir. Edmundson yaptığı çalışmada, kelime sıklığına ek olarak, "ipucu sözcük öbekleri", "başlık terimleri" ve "cümle konumu" gibi üç yeni özellik daha kullanmıştır. Brandow, Mitze ve Rau, tekli dokümanlarda özetleme üzerinde çalışmışlardır. Yaptıkları özetleme sisteminde anahtar kelime seçimini cümle ağırlıklarının belirlenmesinde kullanmıştır. Dokümanlarda bulunan ilk cümleyi özete direkt eklemişlerdir. Meng

Wang [4] ve arkadaşları çoklu dokümanlar da kelime ağırlıklarını kullanarak dokümanları kendi içinde alt konulara ayırdıktan sonra bu alt konuları temsil eden cümlelerin seçimi ile özet oluşturmuştur. Jade Goldstein, Vibhu Mittal, Jaime Carbonell ve Mark Kantrowitzt[5] yaptıkları çalışma da tekli doküman üzerinde uygulanan cümle çıkarım yöntemini çoklu doküman özetlemede uygulamışlardır. Yapmış oldukları çalışmayı 1988-1992 yılları arasında Associated Press ve Wall Street Journal'da yayınlanmış olan ortalama 31 cümle içeren 200 haber metin üzerinde uygulamışlardır. Jaruskulchai [6] önerdiği yöntemde, Thai dilinde Luhn'un TF yöntemini uyarlayarak önemli kelime grubunu bulmuş ve paragraflar arası ilişki ağını oluşturmuştur. Yaptığı bu çalışmayı 3 veri kümesi üzerinde %20 ve %30 olmak üzere iki farklı özetleme oranı ile test etmiştir. %20 özet oranı için sırasıyla %50,%43 ve %40, %30 özet oranı için %55,%45 ve %48 başarı oranı elde etmiştir. Ebru Uzundere, Elda Dedja, Banu Diri ve M.Fatih Amasyalı[7] Türkçe haber metinleri için bir otomatik özetleme sistemi üzerinde çalışmışlardır. Çalışmalarında haber metinlerinin cümlelerini çeşitli özelliklere göre puanlayarak en yüksek puanlı cümlelerin seçimi ile metinlerin özetini çıkarmışlardır. Geliştirdikleri otomatik haber özetleme sisteminin performansı, kullanıcıların çıkardığı cümleler ile aynı olma olasılığı taban alınarak ölçülmüş ve oran yaklaşık olarak %55 bulunmuştur. Lloret ve Palomar [9] özetleme de cümle seçimi konusu üzerinde, kod kalite prensibini (Code Quality Principle) kullanarak, daha fazla bilgi içeren cümlelerin seçimi ile çıkarımı yapmışlardır. Yaptıkları yaklaşım ile bu yöntemin belgenin önemli cümlelerini seçmek için uygun olabileceğini ve bir özetleme sistemi oluşturulurken bu özelliğin göz önündü bulundurulmasının iyi bir fikir olabileceğini belirtmişlerdir. Min ve arkadaşları [10] çalışmalarında, Çince dilinde paragraf gruplarını oluşturmak için başlıkları, başlıkların olmadığı durumlarda ise benzerlik amacıyla önemli olarak belirledikleri kelimelerden oluşan vektörleri kullanmışlardır. Fumiyo Fukumoto ve Yoshimi Suzuki[8] çoklu dokümanlar da özetleme işini anahtar paragraf çıkarımı yöntemi ile yapmışlardır. %20 özet oranı ve 2 adet doküman ile %77,7'lik başarı oranını yakalamışlardır.

Bu çalışmamızdaki amaç, sisteme girilen metinlerin yine kullanıcı tarafından yönetilebilir olan özetleme oranına göre metinlerin paragraf tabanlı çıkarımlarının sağlanarak kullanıcıya özet sunmaktır. Bu makale de ayrıca k-means algoritması yardımıyla oluşturulan paragraf kümelerinin içerisinden paragraf seçimi yapmayı sağlayan iki farklı yöntem incelenmiş ve başarı oranları da karşılaştırılmıştır. Yapmış olduğumuz bu çalışmada k-means algoritmasında başlangıç kümelerinin tespiti için geliştirilen iki yöntem ile k-means kümeleme algoritmasının uygulanmasında farklı bakış açıları sağlamıştır.

Makalenin ikinci bölümünde çıkarıma dayalı olan metin özetlemede paragraf seçimi için kullanılan yöntem anlatılmıştır. Üçüncü bölümde veri kümesi tanıtılmış ve uygulanan yöntemlerin başarı değerleri kıyaslanmıştır. Son bölümde sonuçlardan bahsedilmiştir.

II. YÖNTEM

Bu çalışmanın amacı birden fazla dokümana ait en çok bilgiyi içeren paragrafların seçilerek kullanıcıya bu paragrafların özet olarak sunulmasıdır.

Sistem temel anlamda üç aşamadan oluşur. Birinci aşamada dokümanlar ön işlemeden geçirilir ve paragraf vektörleri oluşturulur. İkinci aşamada ilgili paragraflar organize edilerek alt-konu ilişkisi bulunan paragraf öbekleri oluşturulur. Üçüncü aşamada alt-konu öbeklerinden paragrafların seçimi ve özetleme işlemi gerçekleştirilir.

A. DOKÜMAN ÖN İŞLEME

Dokümanlar birimlere, diğer bir deyişle paragraflara, dokümanlardaki etiketler kullanılarak ayrıştırılır. Paragrafların sınırları belirlendikten sonra ön işleme adımı uygulanır.

Ön işleme adımının ilk safhasında paragraflar kelimelerine ayrıştırılır. Etkisiz kelimeler (stop words) bir dilde çok sık kullanılan ve anlamı olmayan kelimelerdir. Bu kelimeler cümlelerden temizlendikten sonra anlam olarak bir kayba sebep olmazlar, ama varlıkları sistemimizde sonuçları negatif yönde etkileyip daha isabetsiz sonuçların dönmesine sebep olabilirler. Bu yüzden İngilizce diline ait etkisiz kelimeler temizlenir. Kelime kökü, kelimenin en anlamlı parçasıdır. Ön işleme adımının ikinci safhasında ise kelimelerin kökleri bulunarak, ek almış olan kelimeler için ortak parçalar bulunmuş olur. Bu işlem kelime sıklık hesabının doğru olabilmesi için önemlidir.

İkinci adımda paragraf vektörleri oluşturulur. Paragraf vektörlerini oluşturmak için önce her bir kelime kökünün tüm dokümanlar içinde geçiş sıklıkları hesaplanır. Bu, frekansı yüksek olan kelimelerin daha fazla konu ile ilgili olduğunun sezgisel olarak varsayılmasındandır. Daha sonra, kelimeler geçiş sıklıklarına göre sıralanır. Paragraf-anahtar kelime vektörleri için bu sıralanan anahtar kelimelerden belli sayıda seçim yapılır. Bu değer birinci yöntem için ilk 10, ikinci yöntemde ise seçilen özet oranına bağlı olarak oluşacak başlangıç kümesi sayısına eşdeğer ilk kelimeler olarak belirlenmiştir. Daha sonra seçilen bu anahtar kelimeler ile dokümanlardaki her bir paragrafa ait paragraf-anahtar kelime vektörleri oluşturulur.

Tablo 1. *Örnek Paragraf-Anahtar Kelime Vektörleri.*

Paragraf	**Anahtar Kelimeler**									
	network	secur	inform	busin	comput	system	atack	data	hacker	protect
p_1	1	0	1	0	0	0	2	0	1	0
p_2	0	0	0	0	0	1	1	0	1	0
p_3	0	0	1	0	1	0	0	0	1	0

B. PARAGRAFLARIN ORGANİZE EDİLMESİ

Bu adım da K-Means [11] kümeleme algoritması kullanılarak, anahtar kelime vektörleri haline getirdiğimiz paragrafların kümelendirilmesi yapılır. K-Means algoritmasında kullanacağımız küme sayısı kullanıcı tarafından girilen özetleme oranı parametresi ile bağlantılıdır (Formül 1).

$$k = \left\lceil \frac{r}{100} * p \right\rceil \quad (1)$$

k: Küme sayısı
r: Özetleme oranı
p: Dokümanlardaki toplam paragraf sayısı

Çalışmamızda bu formülden yola çıkarak % 20 özet oranı istendiği durum için küme sayımız 20 olmuştur. Küme sayısının belirlenmesinde kullanılan bu yaklaşımın arkasında yatan ana fikir, özetin tüm doküman kümesini kapsaması için farklı alt konuların seçilebilmesidir. Kümelerin başlangıç

paragraf vektörleri (küme merkezleri) aşağıda açıklandığı gibi iki farklı yöntemle belirlenerek, sonuçlar karşılaştırılmıştır.

1.Yöntem: Paragraflar içerdiği anahtar kelimelerin toplamına göre sıralanır. Başlangıç kümesi sayısı kadar paragraf seçilir. Tablo 1'deki değerlerden yola çıkarak örnek vermek gerekirse:

$p_1 = 1 + 0 + 1 + 0 + 0 + 0 + 2 + 0 + 1 + 0 = 5$
$p_2 = 0 + 0 + 0 + 0 + 0 + 1 + 1 + 0 + 1 + 0 = 3$
$p_3 = 0 + 0 + 1 + 0 + 1 + 0 + 0 + 0 + 1 + 0 = 3$

Her bir paragraf vektörü için toplam içerdiği anahtar kelime sayısı bulunur. Tüm vektörler için bu değer bulunduktan sonra bu değere göre sıralanır. Bir önceki adımda hesaplanan başlangıç kümesi sayısına göre paragraf vektörleri başlangıç noktaları olarak seçilir.

2.Yöntem: Anahtar kelimeler tüm dokümanlarda geçme sayılarına göre sıralanır. Tablo 2'de çalışmada kullanılan verilerden bir kesit alınmıştır. Bu kesit için kelimeleri geçiş sıklıklarına göre sıraladığımız da durum "network, inform, secur, hacker" şeklinde olacaktır. Bu sıralamaya göre kelimelerin en çok geçtiği paragraflar belirlenir. Yani "network" için p_{10}, "inform" için p_6, "secur" için p_{10}, "hacker" için p_3 seçilir. Örnekte görüldüğü gibi "network" ve "inform"ı temsil eden paragraflar aynı çıktı. Böyle bir durumda yeni paragrafı seçmek için ikinci en çok geçtiği paragrafa bakılır. Bu durumda "inform" için p_2 seçilir. Son durumda başlangıç noktalarımız p_{10}, p_6, p_3 ve p_2 olmuş olur.

Tablo 2. *Anahtar Kelimelerin Paragraflarda Geçiş Sayıları.*

Paragraf	Anahtar Kelime	Geçiş Sayısı
p_1	Network	2
p_1	Secur	1
p_2	Secur	1
p_2	İnform	1
p_3	Hacker	1
p_5	Network	2
p_6	İnform	3
p_{10}	Network	5
p_{10}	Secur	3

Paragraf vektörlerinin birbirlerine uzaklıklarının hesaplanması için Öklid uzaklık formülü kullanılmıştır (Formül 2):

$$d(p,q) = \sqrt{\sum_{i=}^{n}(q_i - p_i)^2} \quad (2)$$

- p, q: Paragraf vektörleri
- p_i: p vektörüne ait kelime frekansları
- q_i: q vektörüne ait kelime frekansları

Örnek olarak Tablo 1'deki p_1 ve p_2 vektörlerini ele alalım. Bu vektörlerden p_1'in k_1 kümesinin merkez noktası olduğunu kabul edelim. Bulmak istediğimiz şey p_2 vektörünün k_1 kümesine olan uzaklığıdır. Bunun için yukarıda belirttiğimiz Öklid formülünü aşağıdaki gibi uygulayarak p_2 vektörünün k_1 kümesine olan uzaklığını bulabiliriz (Formül 3):

$$d(p_1, p_2) = \sqrt{(0-1)^2 + (0-0)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (1-2)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2} \quad (3)$$

Paragraf vektörlerinin küme merkezine olan uzaklıkları hesaplanır. Daha sonra paragraf vektörü en yakın olduğu kümeye yerleştirilir. Bu kümenin merkezi içerdiği paragraf vektörlerinin kelime frekanslarının toplamı küme içerisindeki paragraf sayısına bölünerek yeniden güncellenir. Bu işlem tüm vektörler bir kümeye yerleştirilene kadar tekrarlanır.

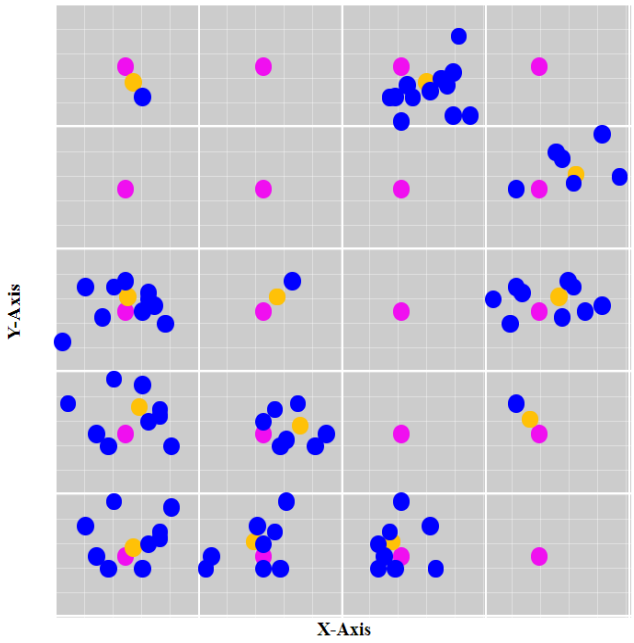
C. PARAGRAFLARIN SEÇİMİ VE ÖZET OLUŞTURMA

Tüm paragraflarımız kümelerine yerleştirildikten sonra kümelerimiz paragraf seçimi için uygun hale gelmiştir. Başlangıç küme sayısını, istenilen özet oranına göre sonuçta olarak çıkması gereken paragraf sayısı kadar belirlediğimiz her kümeden, o kümenin merkezine Jaccard uzaklığına göre en yakın olan paragraf vektörü seçilerek istenilen özet oranı oluşturulmuş olur.

Aşağıda örnek olarak gösterilen Şekil 1'deki dağılım birinci yöntem %20 özet oranı için oluşturul temsili bir şekildir. Burada pembe noktalar başlangıçta belirlenen merkez noktalarını temsil eden paragraflardır. Görüldüğü gibi 102 paragraflık bir veri kümesinden %20 oranında özet istenildiğinde 20 adet başlangıç noktası belirlenmiştir. Mavi noktalar geriye kalan diğer paragrafları temsil eder. Sarı noktalar ise kümelenmenin sonunda belirlenmiş olan merkez noktadır. Paragrafların Jaccard uzaklığını bu merkez nokta ile karşılaştırarak paragraf seçimi yapılır. Paragrafın merkez noktasına olan uzaklığını bulmak için aşağıdaki formül kullanılır (Formül 4):

$$J_g(p,\mathrm{q}) = \frac{\sum_i \min(p_i, q_i)}{\sum_i \max(p_i, q_i)} \quad (4)$$

- p: Hesaplanmış olan merkez noktasının vektörü
- q: Merkeze olan uzaklığını hesaplamak istediğimiz paragraf vektörü



Şekil 1. *Kümelendirilmiş Paragraf Vektörleri.*

III. VERİ KÜMESİ

Paragraf tabanlı çıkarım metodunun başarı değerlendirilmesi Internet üzerinden toplanan "Network Güvenliği" konusunu anlatan 19 adet doküman üzerinde yapılmıştır. Bu 19 doküman 4 farklı kişiye verilmiş ve farklı oranlar da özetlerin çıkarılması istenmiştir.

Tablo 3. *Veri Kümesi.*

Veri Kümesindeki Toplam Doküman Sayısı	19
Veri Kümesindeki Toplam Paragraf Sayısı	102

Yöntem ve özet oranına göre hata matrisi ve başarı oranları Tablo 4, Tablo 5 ve Tablo 6'da verilmiştir. Tabloda "1.Sınama, 2.Sınama, 3.Sınama, 4.Sınama" olarak sistemin başarı oranını test eden 4 farklı kullanıcının testi ifade edilmektedir.

Tablo 4. *%20 Özet Oranı İçin Hata Matrisleri.*

1.Yöntem %20 Özetleme

	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**
Yok	71	11	82	66	16	82	66	16	82	73	9	82
Var	11	9	20	16	4	20	16	4	20	9	11	20
	82	20	102	82	20	102	82	20	102	82	20	102
Hata Oranı	0,215686			0,313725			0,313725			0,176471		
Doğruluk	0,784314			0,686275			0,686275			0,823529		

2.Yöntem %20 Özetleme

	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**
Yok	73	9	82	66	16	82	69	13	82	68	14	82
Var	9	11	20	16	4	20	13	7	20	14	6	20
	82	20	102	82	20	102	82	20	102	82	20	102
Hata Oranı	0,176471			0,313725			0,254902			0,27451		
Doğruluk	0,823529			0,686275			0,745098			0,72549		

Tablo 5. *%40 Özet Oranı İçin Hata Matrisleri*

1.Yöntem %40 Özetleme												
	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	Var	Toplam	Yok	Var	Toplam	Yok	Var	Toplam	Yok	Var	Toplam
Yok	44	18	62	37	25	62	38	24	62	47	15	62
Var	18	22	40	25	15	40	24	16	40	15	25	40
	62	40	102	62	40	102	62	40	102	62	40	102
Hata Oranı	0,352941			0,490196			0,470588			0,294118		
Doğruluk	0,647059			0,509804			0,529412			0,705882		

2.Yöntem %40 Özetleme												
	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**
Yok	45	17	62	39	23	62	43	19	62	46	16	62
Var	17	23	40	23	17	40	19	21	40	16	24	40
	62	40	102	62	40	102	62	40	102	62	40	102
Hata Oranı	0,333333			0,45098			0,372549			0,313725		
Doğruluk	0,666667			0,54902			0,627451			0,686275		

Tablo 6. *%60 Özet Oranı İçin Hata Matrisleri.*

1.Yöntem %60 Özetleme												
	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**
Yok	23	19	42	20	22	42	21	21	42	18	24	42
Var	19	41	60	22	38	60	21	39	60	24	36	60
	42	60	102	42	60	102	42	60	102	42	60	102
Hata Oranı	0,372549			0,431373			0,411765			0,470588		
Doğruluk	0,627451			0,568627			0,588235			0,529412		

2.Yöntem %60 Özetleme												
	1.Sınama			**2.Sınama**			**3.Sınama**			**4.Sınama**		
	Yok	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**	**Yok**	**Var**	**Toplam**
Yok	43	17	60	26	16	42	27	15	42	20	22	42
Var	25	43	68	16	44	60	15	45	60	22	38	60
	68	60	128	42	60	102	42	60	102	42	60	102
Hata Oranı	0,328125			0,313725			0,294118			0,431373		
Doğruluk	0,671875			0,686275			0,705882			0,568627		

Hata matrisinde TN değeri, kullanıcının çıkardığı özet sonucu seçilmemesi gereken paragraflardan kaçının doğru seçildiği sayısıdır. FP değeri, kullanıcının çıkardığı özet sonucu seçilmemesi gereken paragraflardan kaçının seçilmediği sayısıdır. FN değeri, kullanıcının çıkardığı özet sonucu seçilmesi gereken paragraflardan kaçının seçilmediği sayısıdır. TP değeri, kullanıcının çıkardığı özet sonucu seçilmesi gereken paragraflardan kaçının doğru seçildiği sayısıdır.

Doğruluk değeri aşağıdaki Formül 5 ile hesaplanır;

$$(TN + FP)/Toplam \qquad (5)$$

Hata oranı değeri ise Formül 6 olarak tanımlıdır.

$$(1 - Doğruluk) \qquad (6)$$

Tablo 7. *Sistemin Başarı Oranları.*

Özet Oranı	Başarı Oranı	
	1.Yöntem	2.Yöntem
%20	%35	%40
%40	%47	%50
%60	%61	%71

Başarı oranları dikkate alındığında, özetleme oranındaki artışa paralel olarak sezgisel olarak beklendiği üzere başarı oranında da artış olduğu gözlemlenmektedir.

Tablo 8. *Yapılan Çalışmalardaki Başarı Oranları.*

Başarı Oranı	Çalışma	Veri Kümesi	Özetleme Oranı
60,80%	[12]	DUC 2006	250 Kelime
55%	[6]	Thai dilinde haber metinleri (30 doküman)	%30

Tablo 8 de paragraf tabanlı çıkarım tekniği ile oluşturulmuş bazı başarılı çalışmaların başarı oranları görülmektedir. Metin Turan [12] yaptığı tez çalışmasında DUC 2006 (Financial Times of London ve Los Angeles Times gazetelerinden her biri için seçilmiş 25 haberden oluşmaktadır.) doküman kümesini kullanarak özetleme oranını maksimum 250 kelime olarak belirlemiştir ve yakaladığı en yüksek başarı oranı %60,80'dir. Jaruskulchai, C. ve Kruengkrai[6] Thai dili üzerinde yaptıkları çalışma sonucu ulaştıkları en yüksek başarı oranı %55'dir. Bu çalışmada geliştirdiğimiz sistemin en yüksek başarı oranı %60 özetleme için %71 olmuştur.

IV. SONUÇ

Geliştirilen bu sistemde İngilizce dokümanlar üzerinde bir konuya ait birden fazla doküman içinden otomatik özet çıkarılması amaçlanmıştır. Özet çıkarılırken, kümeleme algoritmaları kullanılarak aynı bilgiyi içeren paragrafların seçimi minimuma indirilmeye çalışılmıştır. Sistem, çıkardığı özet oranı bilgisi kullanıcı tarafından değiştirilebilecek şekilde geliştirilmiştir. Bu çalışma da 19 farklı "Network Güvenliği" konusundan bahseden doküman kümesi üzerinden 4 farklı kullanıcı çeşitli oranlar üzerinden özet çıkarmıştır. Bu çıkarılan özetler sistemin başarısını karşılaştırmak için kullanılmıştır. Bu

karşılaştırma sonucu başarı oranlarının değerleri Tablo 7'de verildiği gibidir. Burada k-means algoritmasında başlangıç merkezlerini oluşturmak için belirlediğimiz iki yöntemin sonuçlarını karşılaştırdığımız zaman da ikinci yöntemin bizim sistemimiz için daha uygun olduğu görülmüştür. Özetleme oranının sistemin başarısına etkisine bakıldığında özet oranı arttığı zaman sistemin başarı oranının sezgisel olarak beklendiği gibi arttığı görülmüştür.

Gelecekte önerilen her iki yöntem üzerinde de iyileştirme çalışmaları yapılabilir. Özellikle uzun paragrafların etkilerini ortadan kaldırmak üzere normalize edilmesi ve genel paragrafların (birçok anahtar kelimeyi az sayıda içeren) ayrıştırılmasının başarım oranı üzerindeki etkileri incelenebilir. Çalışma kapsamında gerçekleştirilen yazılımın Türkçe dilini desteklenmesi sağlanabilir. Türkçe ve İngilizce dilleri arasında performans karşılaştırması yapılabilir.

V. KAYNAKLAR

[1] H. P., Lunh, "The Automatic Creation of Literature Abstracts," *IBM Journal*, pp. 159-165, 1958.

[2] H.P., Edmundson, "New Methods in Automatic Abstracting," *Journal of the ACM*, pp. 264-285, 1969.

[3] Ronald Brandow, Karl Mitze ve Lisa F.Rau, "Automatic condensation of electronic publications by sentence selection," *Information Processing and Management,* vol. 31, no. 5, pp. 675-685, 1995.

[4] Meng Wang, Xiaorong Wang, Chungui Li, "Extracting Multi-document Summarization Based on Local Topics," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tientsin, Çin, 2009.

[5] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, Mark Kantrowitzt, "Multi-Document Summarization By Sentence Extraction," NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, ABD, 2000, vol. 4, pp. 40-48.

[6] Jaruskulchai, C. ve Kruengkrai, C., "A Practical Text Summarizer by Paragraph Extraction for Thai," The Sixth International Workshop on Information Retrieval with Asian Language, Sapporo, Japonya, 2003, ss. 9-16.

[7] Ebru Uzundere, Elda Dedja, Banu Diri, M.Fatih Amasyalı, "Türkçe Haber Metinleri İçin Otomatik Özetleme," Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu'nda sunuldu, Isparta, 2008.

[8] Fumiyo Fukumoto ve Yoshimi Suzuki, "Extracting key paragraph based on topic and event detection: towards multi-document summarization," NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization, ABD, 2000, vol.4, pp. 31-39.

[9] Lloret, E. ve Palomar, M., "Challenging Issues of Automatic Summarization: Relevance Detection and Quality-based Evaluation," Informatica, vol. 34, pp. 29-35, 2010.

[10] Min, W., Zhensheng, L. ve Yuqing, G. "Study on Semantic Paragraph Partition in Automatic Abstracting System," Systems, Man and Cybernetics, Tucson, ABD, 2001, pp. 892-897.

[11] Vance Faber ,"Clustering and the Continuous k-Means Algorithm," *Los Alamos Science*, vol. 22, pp. 138-144, 1994.

[12] Metin Turan, "Özgün Paragraf Tabanlı Çıkarım Tekniği Kullanarak Otomatik Çoklu Doküman Özetleme", Doktora Tezi, Bilgisayar Mühendisliği Programı, Yıldız Teknik Üniversitesi, İstanbul, Türkiye, 2015.