TÜRKÇE HABER METİNLERİ İÇİN OTOMATİK ÖZETLEME

Ebru UZUNDERE[1] Elda DEDJA[2] Banu DİRİ[3] M.Fatih AMASYALI[4]

[1,2,3,4]Elektrik-Elektronik Fakültesi Bilgisayar Bölümü
Yıldız Teknik Üniversitesi, Beşiktaş, İSTANBUL

Email: ebruuzundere@hotmail.com elda.dedja@gmail.com banu@ce.yildiz.edu.tr mfatih@ce.yildiz.edu.tr

Özet

Günümüzde internet dünyasının gelişmesi hayal edilemeyen bir bilgi fazlalığını da ortaya çıkarmıştır. Bunun sonucunda bilgisayarların yardımıyla Otomatik Metin Özetleme (Automatic Text Summarization) sistemleri geliştirilmektedir. Özet çıkarma işlemini başarılı bir şekilde gerçekleştirmenin yolu, bu işlemin insanlar tarafından yapılmasıdır, ancak her haberin, makalenin, vd. elle özetlenmesi oldukça zordur. Bu çalışmada, özeti çıkarılacak metnin cümleleri çeşitli özelliklerine göre puanlanmış, kullanıcının istediği özetleme oranına göre en yüksek puanlı cümleler seçilerek metnin özeti çıkarılmıştır. Geliştirdiğimiz otomatik haber özetleme sisteminin performansı, sistem ve kullanıcıların özet olarak çıkardığı cümlelerin aynı olma olasılığı alınarak ölçülmüş ve yaklaşık %55 olarak bulunmuştur.

1. Giriş

Internet'in kullanımı ve buna paralel olarak içerdiği bilgi miktarının her geçen gün artması ile istenilen bilgiye hızlı erişmek önem kazanmıştır. Otomatik metin özetleme, arama motorlarında ve diğer bazı sistemlerde çok kullanışlı olabilir. Binlerce sayfa içerisinden aranılan konu hakkında en önemli bilgileri hızlı ve güvenilir bir şekilde elde etme şansı sunar.

Temel olarak iki çeşit metin özetleme yaklaşımı mevcuttur [1]. Bunlar:

Cümle Seçerek Özetleme: Özetlenecek metindeki önemli cümleleri, istatistiksel metotlar, sezgisel çıkarımlar veya bunların ikisinin birleşimiyle seçerek özetleme yapar.

Yorumlayarak Özetleme: Özetleme işi, özetlenecek metnin akıllıca yorumlanması ile yapılır. Bu özetlemede orijinal metindeki ifadeler akıllı bir şekilde kısaltılarak tekrar yazılmaya çalışılır. Örneğin "Ahmet elmadan, portakaldan ve armuttan nefret eder" ifadesi "Ahmet meyveden nefret eder" şeklinde özetlenir. Bu tip özetleme yapabilmek için çok geniş sembolik kelime bilgisine ihtiyaç duyulur.

Otomatik metin özetleme sistemlerinde bir ya da birden fazla dokümanın özeti çıkarılabilir. Bu çalışmada tek bir dokümanın özetlenmesi üzerine çalışılmıştır. Ancak sergilenen yaklaşım çoklu dokümanlara da uygulanabilir.

Otomatik metin özetleme konusunda ilk çalışma Luhn adlı bir bilim adamı tarafından 1959 yılında yapılmıştır [2]. Bu çalışmada Luhn, özet çıkarmak için kelimelerin cümleler içinde kullanılma frekanslarından yararlanmıştır ve en çok kullanım frekansına sahip kelimelerin o yazı hakkındaki en önemli görüşleri verdiği önsezisinde bulunmuştur. Daha sonra 1969 yılında Edmundson [3], 1989 yılında da Salton [4] bu konuyla ilgili çalışmalarda bulunmuştur. Edmundson yaptığı çalışmada, kelime frekanslarına ek olarak, ipucu veren ifadeleri (Sonuç olarak, Özetle, Bu makale gösteriyor ki,...), konu başlığını içeren kelimeleri, cümlenin bulunduğu yeri, özetleme işlemi sırasında yeni özellikler olarak almıştır. Bu eski yaklaşımların temelindeki düşünceler modern metin özetleme araştırmalarında hala kullanılmaktadır [1]. Altan [5] Türkçe ekonomi haber makalelerinin özetlerinin çıkarılmasında metindeki cümleleri çeşitli özelliklerine göre puanlamış, Pembe de [6] arama motorlarında kullanılmak üzere metnin iç hiyerarşisini de inceleyen bir özetleme sistemi geliştirmiştir. Jung ve çalışma arkadaşları da [7] ardışık cümlelerin birleştirilmesiyle oluşturulan sözde cümlelerin puanlanması tabanlı bir çalışma yapmışlardır.

Makalenin ikinci bölümünde metin özetleme süreçleri, üçüncü bölümde geliştirilen sistemin yapısı ve dördüncü bölümde de deneysel sonuçlardan bahsedilmiştir.

2. Metin Özetleme Süreçleri

Lin ve Hovy'ye [8] göre metin özetlemenin yapılabilmesi için üç farklı aşamanın gerçekleştirilmesi gerekir. Bunlar konunun belirlenmesi, yorumlama ve üretmedir.

2.1. Konunun Belirlenmesi

Konu belirleme aşamasının amacı parçadaki en önemli konuların belirlenmesini sağlamaktır. Bunu sağlamak için kelime frekanslarının hesaplanması, cümlenin bulunduğu yerin incelenmesi, ipucu veren ifadelerden yararlanılması gibi teknikler kullanılır. Bazı yazı tiplerinde, yazının başlığı, yazının ilk cümlesi gibi kritik pozisyonlar yazıyla ilgili en önemli konuları barındırabilirler. "Özetle", "En önemlisi", "Sonuç olarak" gibi ipucu veren ifadeler yazıyla ilgili önemli noktaları gösteren işaretler olabilir. Ayrıca çok sıkça kullanılan kelimeler, edat veya belirteç olmadıkları sürece, içinde bulundukları cümlelerin önemli olduklarını gösterebilirler.

2.2. Yorumlama

Yoruma dayalı olan bu teknikte, karıştırma ve kaynaştırma yapılarak birbiriyle ilgili olan cümleler, daha genel cümleler ile ifade edilebilirler. Örneğin, "Ali masaya oturdu, menüyü okudu, yemeğini yedi ve gitti" cümlesinin yorumlanmış hali "Ali restorana gitti" olacaktır.

2.3. Üretme

Üretme aşaması ise metnin özetlenmiş olan son çıktısının üretilmesidir. Bu aşama en basitinden çok karmaşığına kadar çeşitli üretme metotları içerir. Kullanılabilecek bazı metotlar:

- Birinci aşamada seçilen sözlerin veya cümlelerin özet çıktısına eklenmesidir (extraction).
- En çok kullanılan anahtar kelimelerin veya yorumlanan düşüncelerin özet çıktısına eklenmesidir (topic list).
- İki veya daha fazla cümlenin birbirine bağlanmasıdır (phrase concatenation).
- Cümle üretecinin kaynaşan fikirleri veya birbiriyle ilgili düşünceleri giriş olarak alıp, yeni cümleler üretmesidir (sentence generation).

3. Otomatik Özet Çıkarma Sistemi

Otomatik Metin Özetleme, verilen bilgi kümesinden en önemli bilgilerin kullanıcıya sunulması işlemidir [8]. Geliştirilen sistem çeşitli gazetelerin web sitelerinden ya da çevrimiçi haber sitelerinden elde edilen haberlerin özetlenmesini amaçlamaktadır.

3.1. Puanlamada Kullanılan Özellikler

Geliştirilen bu yöntemde ilk olarak metin, cümlelere ayrılır ve *1,2,…,n* şeklinde indekslenir. Cümlelere ayırma işlemi geliştirilen algoritma ile kod içinde otomatik olarak yapılmaktadır. Daha sonra her cümle, önceden belirlenmiş 13 özelliğe göre incelenir. Bu özellikler sırasıyla:

Başlık-Title: Cümlenin, başlıkta ve varsa alt başlıklarda geçen kelimeleri içerip içermediği incelenir.

Yüksek Frekans- High Frequency: Metin içerisinde yer alan her kelimenin frekansı hesaplandıktan sonra, bir frekans listesi oluşturulur. Yüksek frekanstan düşük frekansa doğru sıralanmış olan bu liste, metindeki her bir kelimeyi ve geçme sıklığını temsil eder. Bu listeye *ve, veya, ile, için,* vd. gibi tek başına bir anlamı olmayan kelimeler dahil edilmemiştir. Liste oluşturulduktan sonra en yüksek frekansa sahip kelimeler (listenin %10'u) özetleme işleminde dikkate alınmaktadır. Cümlenin, bu en yüksek frekanslı kelimeleri içerip içermediği incelenir.

Yer-Location: Cümlenin metin içerisindeki yerine bakılır. Özellikle giriş veya sonuç paragrafında yer alan cümleler incelenir, çünkü buralarda yer alan cümleler özet için önem taşımaktadır.

Anahtar Kelimeler- Key Words: Sistemin özet çıkarırken önemli olduğu düşünülen anahtar kelimelerin kullanıcıdan alınması düşünülmüştür. Örneğin, bir ekonomi haberinin özeti çıkarılmak istenildiğinde kullanıcı dışarıdan *piyasa, döviz* gibi kelimeleri girebilmektedir. Sistem bu kelimelerin geçtiği cümlelere yüksek puan vermektedir.

Özel İsimler- Uppercase: Haber metinlerinde yer alan özel isimler, haberin içeriği hakkında bilgi vermektedir. Bu yüzden sistem özel isimlerin geçtiği cümlelere önem vermektedir.

Pozitif Kelimeler- Positive Words: Cümlelerin *özetle, sonuç olarak, sonuçta, neticede* gibi toparlayıcı kelimeleri içerip içermediği incelenir.

Negatif Kelimeler- Negative Words: *Çünkü, ancak, öyleyse* gibi kelimeleri içeren cümleler konu hakkında ayrıntılı bilgi veren cümlelerdir ve bu cümlelerin özet metinde yer alması gerekmez.

Eşdizimli Kelimeler- Collocation: Cümlenin anlamını pekiştiren eşdizimli kelimelerin cümle içerisinde yer alıp almadığı incelenir.

Sayılar- Numbers: İçerisinde rakam bulunduran cümleler haber metinlerinde önem taşıdığından cümlenin rakam içerip içermediği incelenir.

Çift Tırnak İşareti- Quotation Mark: Bu işareti içeren cümleler alıntı cümleler olup, haber metinlerinde bu tip cümleler önem taşımaktadır.

Bitiş İşareti- Ending Mark: Bir cümlenin ünlem işareti veya soru işareti ile bitmesi diğer cümlelere göre daha önemli olduğunun bir işaretidir.

Ortalama Uzunluk- Average Length: Metinde yer alan cümlelerin ortalama uzunluğu hesaplanır. ±1 ortalama uzunluğa sahip cümleler önemlidir.

Gün Ay- Day Month: Gün veya ay ismi içeren cümleler incelenir. Tarih bilgisi içerdiğinden bu tip cümleler önem taşır.

3.2. Cümlelerin Seçimi

Metin içerisinde yer alan cümleler mevcut 13 özelliğe göre incelendikten sonra cümlenin toplam puanı hesaplanır. Bunun için tüm özelliklere bir ağırlık değeri atanmalıdır. Çalışmamızda ağırlık değerleri sezgisel olarak belirlenmiş olup, Tablo 1'de gösterilmiştir. Örneğin, ***başlık*** özelliğinin ağırlığı 20, ***pozitif kelimeler*** özelliğinin ağırlığı da 15 olsun. İlgili cümle, başlıkta geçen iki kelimeyi ve pozitif kelimeler listesinde yer alan kelimelerden bir tanesini içermiş olsun. Bu durumda cümlenin puanı (20*2)+(15*1)=55 olarak hesaplanacaktır.

Her cümlenin puanı hesaplandıktan sonra, kullanıcı tarafından verilen özetleme yüzdesine göre, en yüksek puana sahip cümleler özet metinde yer almaktadır.

3.3. Sistemin Özellikleri

Geliştirilen sistem kullanıcıya iki seçenek sunmaktadır. Birincisi Normal Özet çıkarma, ikincisi cümlelerin puanlanmasında kullanılan özelliklerin ağırlıklarının değiştirilebildiği Ayrıntılı Özet çıkarmadır. Normal Özet çıkarma, kullanıcının özetleme oranını, anahtar kelimeleri girebileceği ve dosya yükleyebilmesini sağlayacak alanlardan oluşmaktadır. Ayrıntılı Özet çıkarma ise bunlara ek olarak tüm özelliklerin ağırlık değerlerinin değiştirebileceği bir alana daha sahiptir. İleriki çalışmalarda bu ağırlık değerlerinin belirlenmesinde, kullanıcıların verdikleri özetlerden yararlanılarak optimize edilmiş ağırlık değerlerinin kullanılması planlanmaktadır. Ayrıntılı Özet çıkarma özelliği ile kullanıcıya esneklik sağlayan sistem, kullanıcının arzu ettiği özet metni elde etmesini de kolaylaştırmıştır.

Geliştirilen sistem çeşitli aşamalarda ***Zemberek*** [9] programından yararlanmaktadır. Örneğin başlık ile metindeki kelimelerin kökleri ***Zemberek*** programı yardımıyla elde edilir. Elde edilen köklerin cümledeki anlamına bakılmaksızın karşılaştırma yapılır. ***Zemberek*** programının kullanıldığı bir diğer alan metin içerisinde geçen özel isimlerin bulunduğu aşamadır. Kelimeler ***Zemberek*** programına gönderilerek özel isimler bulunur ve cümlenin puanı arttırılır.

Sistem pozitif, negatif ve eşdizimli kelimelerin tespitini, çeşitli kaynaklardan elde edilen ve daha önceden veri tabanına eklenen değerlerle karşılaştırarak yapmaktadır.

4. Deneysel Sonuçlar

Bu çalışmada, farklı haber sitelerinden alınan 10 adet haber metni kullanılmıştır. Bu haber metinlerinin her biri 15 farklı test kullanıcısına verilmiş ve cümle seçimi şeklinde haber metinlerinin özetinin çıkarılması istenmiştir. Daha sonra aynı haber metinleri, geliştirdiğimiz sisteme verilmiş, özetleme oranı parametresine de kişinin özet olarak çıkardığı cümle sayısı kadar özetin elde edilebileceği bir değer verilerek sonuç alınmıştır. Örneğin, haber metni 10 cümle içeriyorsa ve kullanıcı özet olarak 3 cümle seçmişse, sisteme bu dosya yüklendiğinde özetleme oranı %30 olarak verilmektedir.

Sistem test aşamasında Normal Özet çıkarma özelliğini kullanmıştır. Böylelikle varsayılan parametreler kullanılmış ve adil bir şekilde sistemin başarısı ölçülmeye çalışılmıştır. Tablo 1'de sistem tarafından kullanılan özelliklerin ağırlık değerleri verilmektedir.

Sistemin performansı, her bir metin dosyasının özetleme başarısı ayrı ayrı bulunup, ortalaması alınarak hesaplanmıştır (Tablo 2, sütun S %). Özeti çıkarılacak olan bir metin dosyasının başarısı, ilgili metin dosyasının özetini çıkaran her test kullanıcının özet olarak aldığı cümleler ile, sistemin özet olarak kabul ettiği cümlelerle aynı olan cümle sayısının, kullanıcının çıkardığı özet cümle sayısına oranlarının toplamının ortalaması alınarak hesaplanır. Her bir metinden elde edilen ortalama başarı ve sistemin ortalama başarısı Tablo 2'de gösterilmiştir.

Tablo 1: Özelliklerin ağırlık değerleri

Özellikler	**Ağırlık değeri**
Başlık	20
Yüksek Frekans	10
Giriş	20
Sonuç	2
Anahtar Kelimeler	8
Özel İsimler	3
Pozitif Kelimeler	15
Negatif Kelimeler	-20
Eşdizimli Kelimeler	4
Sayı	3
Çift Tırnak İşareti	2
Bitiş İşareti	2
Ortalama Uzunluk	10
Gün Ay	5

Metin içerisindeki cümle sayısının özetleme başarısına etkisi incelendiğinde, beklenildiği gibi genelde cümle sayısı az olduğunda sitemin başarısı yüksektir.

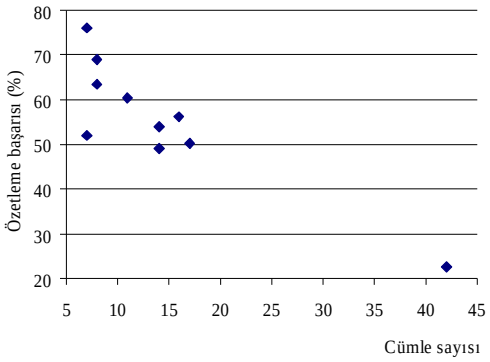
Şekil 1, cümle sayısı ile sistemin özetleme başarısının değişimini göstermektedir. Sistemin başarısını yükselten parametrelerden biri de özetleme oranıdır. Bu değer ne kadar büyükse

başarı da o kadar yüksek olmaktadır. Bu ilişki Şekil 2'de görülmektedir.

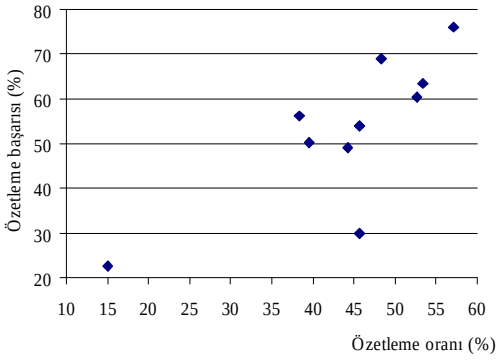
Tablo 2: Sistemin başarısı
(No: haber metninin numarası, Cs: haber metnindeki cümle sayısı, Öo: özet oranı, Öss: özet oranının standart sapması, S%: sistemin başarısı, Bss: başarının standart sapması)

No	Cs	Öo	Öss	S %	Bss
1	14	0,44	1,82	49,20	19,15
2	17	0,40	2,52	50,20	14,61
3	11	0,53	2,18	60,40	15,62
4	14	0,46	2,03	54,07	13,78
5	7	0,57	1,13	76,13	18,60
6	8	0,48	1,85	68,87	14,40
7	8	0,53	0,59	63,40	13,20
8	16	0,38	2,26	56,07	15,09
9	7	0,46	0,68	52,0	18,29
10	42	0,15	3,09	22,73	12,77
Ortalama	14,40	0,44	1,81	**55,31**	15,55

Sonuç olarak, her bir haber metni ve kullanıcı için sistemin ürettiği sonuçlar karşılaştırıldı ve sonuçların ortalaması alınarak sistemin performansı yaklaşık %55 olarak ölçüldü.



Şekil 1: Cümle sayısının başarıya etkisi



Şekil 2: Özetleme oranının başarıya etkisi

Tablo 2 incelendiğinde özetleme oranının, cümle sayısı ve özetleme oranının standart sapmasıyla ters orantılı, cümle sayısının da özetleme oranının standart sapmasıyla doğru orantılı olduğu görülmüştür.

5. Sonuç

Geliştirilen sistem, Türkçe haber metinlerinin otomatik olarak özetinin çıkarılmasını amaçlamaktadır. Özetin çıkarılmasında 13 özellik belirlenmiş ve bu özelliklere sezgisel bir yaklaşımla ağırlık değerleri verilmiş olup, bu değerlerin değiştirilebilme imkanı kullanıcıya sunulmuştur. Sistemin sonucuna etkisi olan özetleme oranı parametresi ise yine kullanıcının seçimine bırakılmıştır. Çalışmamızda 10 farklı haber metninin 15 kullanıcı tarafından özeti çıkarılmış, daha sonra sistemin verdiği özet metinleri ile karşılaştırması yapılarak başarı yaklaşık %55 olarak elde edilmiştir.

Cümle sayısının ve özetleme oranının sistemin performansına etkisi incelendiğinde uzun metinlerde performansın düştüğü, özetleme oranı arttırıldığında performansın yükseldiği görülmüştür.

Gelecek çalışmalarımızda ilk hedef olarak, veri setinin genişletilmesi ve kullanıcılardan istenen özet bilgisinin kendilerine farklı özetleme oranları verilerek toplanmasıdır. İkinci aşamada eldeki veri setinin bir bölümünü kullanarak, çeşitli öğrenme algoritmalarıyla sistemin özet parametrelerini kendisinin belirlemesi düşünülmektedir.

Sistemde kullanılan 13 adet özelliğin tek tek incelenerek hangilerinin sisteme olumlu hangilerinin olumsuz etkisinin tespiti de gelecek çalışmalarımızdaki hedeflerimizdendir. Böylelikle sistemin performansına olumsuz etki eden özellikler belirlenebilir ve sistemden çıkarılabilir.

Bir diğer hedefimiz ise cümle ayırma algoritmasının geliştirilmesidir. Verilen metni cümlelerine ayırmak için geliştirilen sistemler incelenerek en optimum yolun bulunması ve sisteme uygulanması sistemin cümlelere ayırma performansını artıracaktır.

6. Kaynaklar

[1] http://people.dsv.su.se/~hercules/papers/FarsiSum.pdf

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

[3] H.P., Edmundson, "New Methods in Automatic Abstracting", *Journal of the ACM*, Vol.16(2), p:264-285, 1969
[4] G. Salton, J. Allan, C. Buckey, A. Singhal, "Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts", *Science*, 264, p:1421–1426, 1989
[5] Z. Altan, "A Turkish Automatic Text Summarization System", *Proceedings of the Artificial Intelligence and Applications*, 2000
[6] C. Pembe, T. Güngör, "Automated Query-biased and Structure-preserving Text Summarization on Web Documents", *INISTA*, 2007
[7] W. Jung, Y. Ko, and J. Seo, "Automatic Text Summarization Using Two-Step Sentence Extraction", *AIRS 2004*, LNCS 3411, pp. 71 – 81, 2005
[8] E. Hovy, C.Y., Lin, "Automated Text Summarization in SUMMARIST", *Annual Meeting of the ACL Proceeding of a Workshop*, 1998
[9] http://code.google.com/p/zemberek/