Araştırma Makalesi / Research Article

Metin Özetlemesi için Düğüm Merkezliklerine Dayalı Denetimsiz Bir Yaklaşım

Cengiz HARK[1*], Taner UÇKAN[2], Ebubekir SEYYARER[2] , Ali KARCI[3]

[1]Bitlis Eren Üniversitesi, Enformatik Bölümü, Bitlis
[2]Van Yüzüncü Yıl Üniversitesi, Bilgisayar Teknolojileri Bölümü, Van
[3]İnönü Üniversitesi, Bilgisayar Mühendisliği Bölümü, Malatya
(ORCID:0000-0002-5190-3504) (ORCID: 0000-0001-5385-6775)
(ORCID: 0000-0002-8981-0266) (ORCID:0000-0002-8489-8617)

Öz
Cümle seçerek özetleme çalışmaları kapsamında birçok farklı yaklaşım mevcuttur. Bu çalışmada tek dokümanlı çıkarıcı metin özetleme için yeni ve denetimsiz bir süreç önerilmektedir. Çalışma kapsamında metin dokümanları çizgelerle temsil edilmektedir. Sunulan yaklaşım temel olarak metinleri temsil eden çizgeleri kullanmakta ve cümlelere yönelik bir ağırlıklandırma önermektedir. Önerilen sürecin farklı düğüm ağırlıklandırma yöntemlerini kullanarak önemli düğümleri belirlenmesi, önerilen özetleme sisteminin cümle puanlandırma aşamasını oluşturmaktadır. Son olarak bu çalışma kapsamında metin özetleme amaçlı önerilen yaklaşımın, açık erişimli metinler ve bu metinlere ait özetleri içeren Document Understanding Conference (DUC-2002) veri seti üzerindeki performansı ROUGE değerlendirme metrikleri kullanılarak hesaplanmıştır. Yapılan deneysel çalışmalar sonucunda önerilen özetleme sisteminin geleneksel çizge tabanlı yaklaşımlar ile rekabet edebilir ölçüde performans değerleri ortaya koyduğunu göstermektedir. Önerilen özetleme yaklaşımı ile elde edilen ROUGE-2 metriğinin Duyarlılık, Kesinlik ve F-Skor değerleri sırasıyla 0.17068, 0.15772, 0.16383 olarak hesaplandı. Ayrıca sunulan bu basit ve etkili yöntemin dilbilimsel bir süreç izlememesi oldukça önemlidir.

Anahtar kelimeler: Çıkarıcı Metin Özetleme, Düğüm Merkezliği, Cümle Sıralama, ROUGE Metrikleri.

An Unsupervised Approach Based on Node Centralization for Text Summarization

Abstract
There are many different approaches to summarizing through sentence selection. In this study, a new and unsupervised process is proposed for single document extractor text summarization. Within the scope of the study, text documents are represented by graph. Our method basically uses the graph representing texts and proposes a weighting for sentences. Using the different node weighting methods of the proposed process, identifying important nodes and thus important sentences constitutes the sentence grading stage of the proposed summarizing system. Finally, within the scope of this study, the proposed approach for text summarization was calculated by using Document Understanding Conference (DUC-2002) evaluation metrics in the ROUGE dataset containing open access texts and summaries of these texts. As a result of the experimental studies, the proposed summarization system shows the performance values that can compete with the traditional graph-based approaches. It is also important that this simple and effective method does not follow a linguistic process.

Keywords: Extractive Text Summarization, Node Centrality, Sentence Ranking, ROUGE Metrics.

1. Giriş

Ham veri, bir olgu hakkında birbirleri ile olan ilişkileri henüz tam olarak ortaya konmamış bilgi toplulukları olmakla birlikte sayısal formatlarla ifade edilebilen, taşınabilen dizeler olarak da

*Sorumlu yazar: chark@beu.edu.tr
Geliş Tarihi:22.05.2019, Kabul Tarihi: 08.08.2019

tanımlanabilmektedir. Bu ham verilerin bilinen bir amaç doğrultusunda anlamlı birlikteliklerinin elde edilmiş hali ise bilgidir. Her geçen gün büyüyen veri miktarı yeni yaklaşımlar ile çözülmeyi bekleyen problemlere kapı açmaktadır. İşlenmeyi bekleyen dijital veriler çok büyük oranda belirli bir formatı olmayan ham metinlerdir. Sözü edilen verilerin anlamlı ve faydalanılabilir veri kaynaklarına dönüşmesi için üzerinde çalışılarak analiz edilmeleri gerekmektedir [1, 2]. Bilgiye erişim zamanının kısaltılarak kabul edilebilir erişim sürelerinin elde edilmesi için yeni yöntemlerin geliştirilmesi gerektiği kadar verinin analiz edilmesi de gerekmektedir [3, 4].

Otomatik metin özetleme, bu analiz yöntemleri arasında yer almaktadır. Son yıllarda araştırmacılar metin dokümanlarının özetlenmesi amacı ile yeni algoritma ve yöntemler geliştirmek için çalışmaktadırlar. Zira uzun metin belgelerinin insanlar tarafından özetlenmesi hem çok zor hem de gerçekten fazla zaman gerektirmektedir. İş dünyası, akademi, sağlık gibi birçok alanda özetleme oldukça gerekli görülmektedir. Araştırmacılar için metin özetleme hala önemli bir çalışma alanıdır. Radev [5] özeti, özgün metin veya metinlerdeki önemli bilgileri taşıyan, çoğunlukla daha kısa olan metinler olarak adlandırmaktadır. Erkan [6] 2004'de metin özetlemeyi, kullanıcı için faydalı bilgiler sağlayan belirli bir metnin sıkıştırılmış bir versiyonunu otomatik olarak oluşturma işlemi olarak tanımlamaktadır. 2019'da Joshi [7] otomatik metin özetleme çalışmalarını, son kullanıcılar için hızlı bir şekilde anlaşılabilir olması amacı ile uzun metin belgelerini sıkıştırılmış bir şekilde temsil etmeyi amaçlayan önemli bir Doğal Dil İşleme süreci olarak tanımlamaktadır.

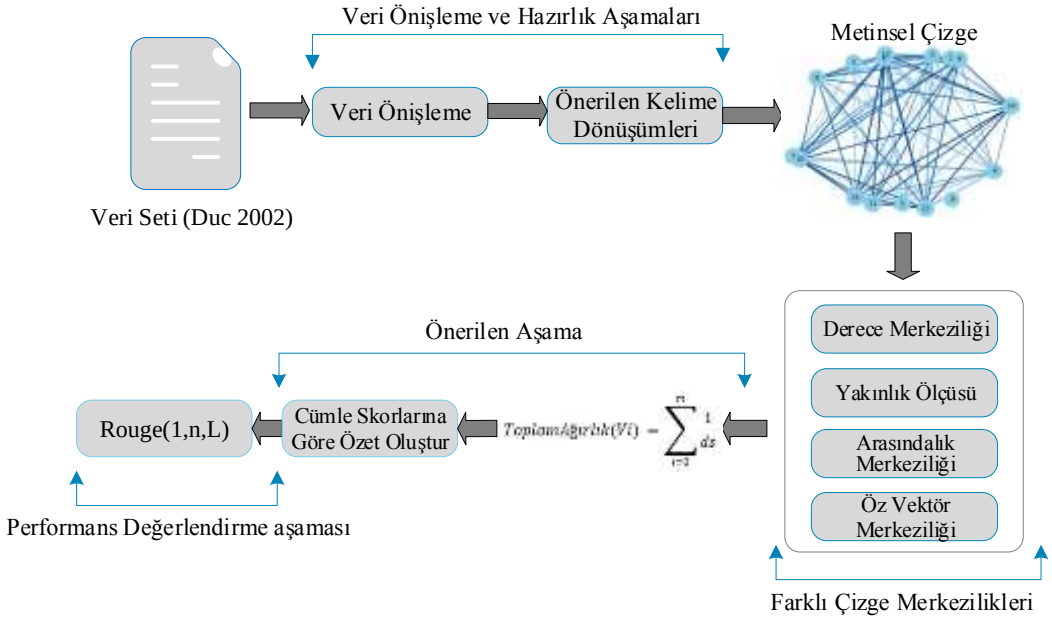
Bir özetin taşıması istenen bilgi içeriği kullanıcı tarafından özel olarak belirlenebilir. Bu şekilde konu-odaklı özetler (topic-oriented summarization) kullanıcının ilgilendiği konuya odaklanır ve belirtilen konu ile ilgili metindeki bilgileri çıkarır. Genel Özetleme (Generic summarization) ise, orijinal metnin genel konu içeriğini koruyarak, mümkün olduğunca fazla bilgiyi kapsamaya çalışır [6, 8]. Metin özetleme çalışmalarında temel anlamda iki farklı yaklaşım kullanılmaktadır. Bu yöntemler çıkarıcı özetleme (extractive summarization) sistemleri ve yorumlayıcı özetleme (abstractive summarization) sistemleri olarak kategorize edilebilmektedir. Metin belgesinden önemli cümleleri seçmek yolu ile özetleme yapmak çıkarıcı özetleme olarak adlandırılmaktadır. Skor değeri atanan cümle parçacıkları hakkında bir önem düzeyi belirlenmiş olur. Belirlenen bir üst limit ile sınırlandırılarak özetler elde edilir. Yorumlayıcı özetleyici sistemlerde süreç tamamen farklıdır. Bu yaklaşımda cümleler belirli adımlar ile yorumlanarak tekrar ifade edilir. Paragraflar veya cümleler gibi metin içerisinde yer alan kısımlar dilbilimsel metotlar kullanılarak ana metinden elde edilir. Bu metot gerçek metni anlamaya yorumlamaya ve daha az kelime kullanarak özlü bir şekilde elde etmeye dayanan bir yöntemdir [7, 9]. Yorumlayıcı özetleme sistemleri çıkarıcı özetleme sistemine göre daha karmaşık adımlardan oluşmaktadır. Ayrıca özetleme sistemleri, verilen girdi belgelerinin sayısına bağlı olarak, tekli doküman özetleme (single-document summarization) ve çoklu doküman özetleme (multi-document summarization) olarak da sınıflandırılabilmektedir [10].

Son zamanlarda metin özetlemeye yönelik olarak etkili çizge tabanlı yaklaşımlar öne sürülmüştür. Mihalcea TextRank [11] adını verdiği yaklaşımında çizge bazlı cümle skorlama yöntemi ile metinleri özetleme yoluna gitmiştir. Benzer şekilde LexRank yaklaşımı ile Erkan [6] cümlelere puan vererek çıkarıcı özetleme gerçekleştirmiştir. Yine Parveen [12] Egraph'ı tanıtmıştır.

Bu çalışmada denetimsiz çizge tabanlı, genel, tekli ve çıkarıcı özetlemeye odaklanılmaktadır. Önerilen metot, metinde yer alan cümlenin önemini, çizgedeki düğümlerin öneminden yola çıkarak belirlemekte ve merkezlik ölçülerini doğrudan çizge ilişkilerine bağlı olarak hesaplamaktadır. Çizge teorisi literatüründe yer alan düğüm merkezlik ölçüm değerleri hesaplanmakta ve elde edilen puanların birleştirilmesi ile düğüm önem sıralaması oluşturulmaktadır. Bu bağlamda bir düğümün merkezlik değeri yani düğümün önemi, düğümün derecesi, diğer düğümlere olan yakınlığı, kendisi dışındaki düğümlerle sahip olduğu en büyük mesafesi vb. düğüme ait birçok özellik ile ifade edilmektedir. Farklı yöntemlerle merkezi düğümlerin belirlenmesi ile düğümleri temsil eden cümlelerin belirlenmesi önerilen özetleme sisteminin bu adımını oluşturmaktadır. Hesaplanan düğüm merkezlik ölçütlerinden elde edilen puanlar ile bir cümle seçim stratejisi ve hatta cümle sıralaması ortaya konmaktadır. Çıktı özeti, önceden tanımlanmış bir uzunluk ile sınırlandırılmış (bu çalışma için 200 kelime) en üst puanları alan cümleleri seçerek üretilmektedir.

2. Önerilen Yöntem

Metin özetleme için önerilen modelin süreçlerine dair blok diyagramı Şekil 1'de gösterilmektedir. Bu çalışmada, verilen metinlerden uygun özetler çıkarılması amacıyla çizge tabanlı genel, çıkarıcı ve tek belgeli bir özetleme yöntemi sunulmuştur. Çalışmanın ayrıntılı blok diyagramı Şekil 1'de adım adım gösterilmektedir. Önerilen otomatik özetleme yöntemi temelde üç ana aşamadan oluşmaktadır. İlk aşamada bir takım önişlemler yapılmakta, yazarlar tarafından geliştirilen metin ön işleme aracı kullanılarak yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi işleme alınmasının önüne geçilmektedir. İkinci aşamada cümleler arasındaki ilişkiler şekilsel ve matematiksel olarak temsil edilmektedir. Üçüncü ve son aşamada çizgeye ait farklı düğüm merkezlik yaklaşımları uygulanarak önemli düğümlere ve dolayısı ile önemli cümlelere ulaşılmaktadır.



Şekil 1. Önerilen özetleme modelinin şematik taslağı

2.1. Metin Önişleme Hazırlık

Yoğun veri kümelerinin pek azı yapılandırılmış formatlara sahip veri kümeleridir. Yapılandırılmış veri kümeleri tablolar halinde satır ve sütunlarla ifade edilebilen veya etiketlere sahip veri kümeleridir. Bu çalışmada da üzerinde çalışılacak veriye belirli bir yapısallık kazandırılması aşamaları oldukça karmaşık ve zaman alıcı aşamaların geride bırakılması ile mümkün olmuştur. Belirli bir bütünlük taşımayan, buna karşın bir yapısallık kazandırılması gereken veri kümeleri için bazı önişlemler zorunludur. Sisteme girdi olarak kullanılacak metinlerin üzerinde çalışılabilecek biçimlere aktarılması ayırt ediciliği olmayan gereksiz verilerden soyutlanmış olması sistemin başarımını arttıracağı öngörülmektedir. Zira bahsedilen yoğun kümeler doğal dil kullanılarak oluşturulmuşlardır [13]. Metin özetleme sürecinde, üzerinde çalışılacak metinler herhangi bir önişlem sürecine katılmadan kullanılmamaktadır. Durak kelimeler (zamirler, edatlar, bağlaçlar, vb..) özetleme öncesinde veri setinden uzaklaştırılması gereken ve ayırt edici özelliği olmayan ifadelerdir. Bu nedenden ötürü verileri, üzerinde çalışılabilecek formatlara aktararak metin özetleme işleminin gerçekleştirilebilmesi için cümleler üzerinde temsil gücü olmayan durak kelimeler orijinal veri kümelerinden çıkarılmaktadır. Böylece özetleme sürecindeki işlem yükü azaltılmaktadır [1]. Sözü edilen bu normalizasyon adımları NET framework 4.7'de C# kullanılarak geliştirilmiştir. Bu normalizasyon adımları tek başına veri önişleme aşamasında yeterli olmamaktadır. Bu adımlara ilaveten yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi değerlendirilmesi, cümleler arasındaki bağlantı ve ilişkilerin tespitini zorlaştırmaktadır. Sezgisel olarak

sözü edilen problemin, özetleme sonrasında elde edilen başarımı ciddi bir şekilde etkileyeceği öngörülerek özetleme öncesinde kelimeler ve ifadeler üzerinde bir dönüşüm sağlanmıştır. Sözü edilen dönüşüm Net framework 4.7 kullanılarak gerçekleştirilmiştir. Tablo 1, durak kelimeler çıkarıldıktan sonra elde edilen kelimeler üzerinde gerçekleştirilen dönüşüme dair bazı örnekler içermektedir.

Tablo 1. DUC 2002 Veri Seti'nde yer alan d070f isimli örnek metin dokümanının geliştirilen yazılım aracı ile gerçekleştirilen bazı dönüşümlerini içeren tablo

Ön İşleme yazılımından önce	Ön İşleme yazılımından sonra
Lawyer, Lawmaker	Law
powerful	power
husband's	husband
Russian	Russia
Chilean	Chile
newspaper	news

Bazı kelimeler üzerinde hiçbir değişiklik gerçekleştirilmemektedir. Geliştirilen yazılım aracının herhangi bir değişikliğe gitmediği kelimeler metinde yalnızca bir adet bulunan kelimelerdir. Dolayısı ile anlamsal bakımından benzeştiği veya yakın anlamlı olduğu bir ifade ile değişimi mümkün olamamaktadır. Buna karşılık Tablo 1'de yer alan "Lawyer" ve "lawmaker" gibi kelimeleri, tüm metinde bulunan anlamlı en yalın halleri ile yani "Law" kelimesi ile değiştirilmiştir. Ekli ifadelerin yerini alan "Law" kelimesi yine metnin tamamında aranmış ve belirlenmiştir. Böylelikle sözü edilen ekli kelimelerin farklı anlamlar taşıyormuş gibi çizgeye aktarılmalarının önüne geçilmektedir. Özetle değişikliğe uğrayan kelimenin yerini alacak olan en yakın ve anlamlı kelime gerçekleştirilen yazılım tarafından yine metin içinde aranmakta ve değişim yapılmaktadır.

Bu çalışmada açıklandığı gibi ham ve günlük konuşma dili özelliklerini taşıyan metinler, metin önişleme ve hazırlık süreci ile evvela gereksiz ve istenmeyen verilerden arındırılmıştır. Daha sonra geliştirilen yazılım aracı kullanılarak benzer veya çok benzer anlamlar taşıyan ifadelerin çizgeler oluşturulurken farklı anlamlar taşıyor gibi algılanmasının önüne geçilmiştir. Bu sayede cümleler arasındaki bağlantıların ve ilişkilerin tespitinin kolaylaştırılması hedeflenmiştir.

2.2. Metinsel Çizge

Özellikler ve bu özellikler arasındaki ilişkilerin şekilsel olarak temsil edilmesi ile problemlerin çizgeler ile modellenmesi gerçekleştirilebilmektedir. Çizgeler, kavramsal olarak düğümlerden ve düğümler arası ilişkileri temsil eden ayrıtlardan oluşmaktadır. Düğümler ve ayrıtlar iki sonlu kümedirler. Düğümler, çizgenin temsil ettiği topluluğu oluşturan esas bireylerdir. Ayrıtlar ise, esas bireyler arasındaki söz konusu ilişkilerdir. Genel olarak çizge $G=(V,E)$ ile gösterilmektedir. Düğümler kümesi $V=\{v_1, v_2,...,v_n\}$, ayrıtlar kümesi ise $E=\{e_1,e_2,...e_n\}$ $(E \subseteq V \mathbf{X} V)$ dir. $e_i=\{v_j,v_j+1\}$ ayrıtı için uç düğümler v_j ve v_j+1 düğümleridir. v_j ve v_j+1komşu düğümleri için $e_1=(v_j,v_j+1) \in E$ ve $(v_j+1,v_j) \in E$ ise, bu ayrıtlar yönlendirilmemiş ayrıtlardır. Bu tarz ayrıtların oluşturduğu çizgeler yönlendirilmemiş çizgelerdir. Eğer v_j ve v_j+1 komşu düğümleri $e_1=(v_j,v_j+1) \in E$ ve $(v_j+1,v_j) \notin E$ ise, bu tür ayrıtlar yönlendirilmiş ayrıtlardır. Yönlendirilmiş ayrıtların oluşturduğu çizgeye yönlendirilmiş çizge adı verilmektedir [2, 14, 15]. Bir $G=(V,E)$ çizgesi üzerinde $V=\{v_1, v_2,...,v_n\}$ düğümleri arasındaki ilişkileri temsilen iki düğüm arasındaki ayrıtlar $E=\{e_1,e_2,...e_n\}$ eksi olmayan ağırlıklar taşıyor ise, bu tarz çizgeye ağırlıklı çizge adı verilmektedir [16]. Bu çalışmada metinleri temsilen oluşturulan çizgeler yönlendirilmemiş ve ağırlıklı çizgelerdir. Metinsel çizgelerin temsil edilmesi amacıyla literatürde farklı yaklaşımlar söz konusudur. Tam bir cümle veya bir cümlecik düğümler ile temsil edilebilirken, düğümler arasındaki kesişim, birlikte oluşumlar gibi farklı ilişki şekilleri ise, ayrıtlar ile temsil edilebilmektedirler.

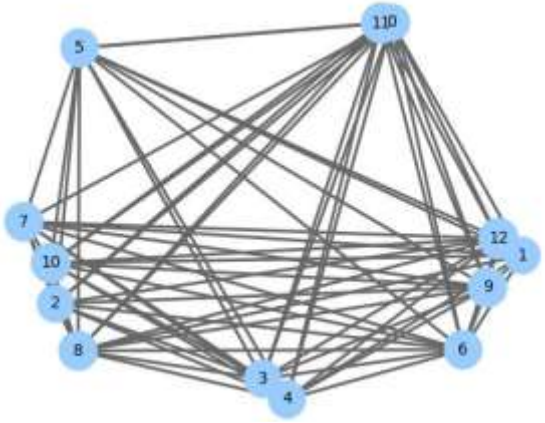
Genel olarak bir çizgede yer alan düğümlerin metinleri temsili noktasında kavramlar, terimler, cümleler, paragraflar ve belgeler düzeyinde karşılıkları olabilmektedir. Düğümler arasındaki ilişkileri temsil etmesi için ise anlamsal ilişkiler, kelime birliktelikleri çerçeve modeli gibi farklı yaklaşımlara sahip temsil tiplerine rastlanmaktadır.

Bu çalışmada metinleri temsil eden çizgeler oluşturulmuştur. Temsili çizgelerdeki sonlu kümelerden ilkini yani düğümleri cümleler temsil etmekte iken, diğer sonlu küme olan ayrıtları ise, ortak kelime sayıları temsil etmektedir. Böylece metinlerin yönlendirilmemiş ve ağırlıklı çizgeler ile temsili

sağlanabilmektedir. İfade şekilleri bakımından ise dinamik olarak ifade ve matrislerle ifade temsil tipleri mevcuttur. Bu bölümde gerçekleştirilen işlemler Şekil 1'de şematize edilen aşamalardan Metinsel Çizge aşamasına karşılık gelmektedir.

Şekil 2, basit bir metin örneğini ve bu metin için oluşturulan çizgeyi içermektedir. Oluşturulan çizge ağırlıklı ve yönsüz bir çizgedir. Deneysel çalışma sürecinde düğümlerin cümleler ile ayrıtların ise ortak kelime sayıları ile temsil edildiği bir dönüşüm gerçekleştirilmektedir. Cümlelerin her birinin, diğer bütün cümlelerle mevcut kelime kesişim sayıları hesaplanarak cümle ikililerinin ilişki düzeyleri ortaya konmaktadır. Bu sayede cümleleri ve cümleler arasındaki ilişkileri temsil gücü yüksek bir çizge ile ifade edebilmekteyiz. Elde edilen çizgeler tüm cümle çiftleri arasındaki bağlantı ve ilişkilerin düzeyini ortaya koyan bir çizgedir.

The mass emigration of thousands of disaffected East Germans has rekindled reunification talk in West Germany, where some legislators plan to begin exploring the possiblity of reuniting the two Germanys. No one is predicting a date for a possible reunification. Proponents acknowledge the risks of assuming too much of East Bloc reform trends, and most hedge their thoughts with careful references to Washington and Moscow. Yet growing signs of ferment in East Germany have given Chancellor Helmut Kohl's government reason to consider the possibilities. Eduard Lintner, parliamentary speaker on German affairs for the joint Christian Democrat and Christrian Socialist caucus in the Bundestag, said his committee has placed reunification on its October agenda. He said the sessions, which will include government officials and outside experts, hopefully can devise reunification models to discuss with West Germany's Western allies, particularly the US "It has become clear to the world that the Communist system in East Germany is in trouble," Mr Lintner said. "The time has come for us to show our allies, but also reform-minded people in the East, that German reunification is no longer just a theory but a subject requiring a serious appraisal of how it could come about. " "There are still a number of imponderables," Mr Linter conceded. "For instance, we want the federal government to approach our allies and get their thoughts on the reunification topic. " The US ambassador to Bonn, Vernon A Walters, earlier this month reaffirmed Washington's policy of supporting a German reunification based on free, democratic elections. But Mr Walters caused a stir by predicting that reunion could occur in the relatively near future. Emerging opposition to uncompromising Communist rule in East Germany has raised questions on how long East Berlin's leaders can resist the democratic reforms sweeping through neighboring East Bloc countries. Reforms in Poland, Hungary and even the Soviet Union have magnified East Germany's Stalinist controls.



Şekil 2. DUC-2002 Veri Seti'nde yer alan d069f dosyasına ait örnek bir metin ve bu metni temsil eden örnek çizge

Metinler içerisindeki her cümle için temsili çizgeye bir düğüm eklenmiştir. Düğümler arasındaki ayrıtlar için ise metinde bulunan cümlelere dair kesişen kelime sayıları dikkate alınarak ayrıt ağırlığı eklenmiştir. Şayet cümleler arasında kesişen kelime bulunmamakta ise ilgili düğümler arasına ayrıt eklenmemektedir. Bütün cümleler, tüm kombinasyonları kapsayacak şekilde birbirleri ile olan ortak içerikleri bakımından ilişkilendirilmiş ve cümleler arasındaki ilişkiler çizgeye doğru bir şekilde aktarılabilmiştir. Şekil 2'de Veri Önişleme ve Hazırlık aşamalarının sonrasında oluşturulan örnek bir paragraf ve bu paragrafa karşılık gelen çizge görülmektedir. Örnek paragraf DUC-2002 Veri Seti'nde yer alan örnek metinlerden seçilmiştir.

2.3. Düğüm Merkezlik Ölçütleri

Merkeziyet kavramı, Bavelas tarafından, bir bireyin sosyal ağ içindeki yapısal konumunun, bu kişinin grup çapında süreçlerdeki etkisini nasıl belirlediğini anlamak için ortaya atılmıştır [17]. Sonrasında yapılan çalışmalarda çok sayıda merkezlik ölçümü önerilmiştir. Bu önlemlerin her biri bir ağda "merkezi" olmanın ne demek olduğuna dair farklı fikirleri savunmaktadır. Merkeziyet ölçümleri birçok farklı alanda merkez noktaları veya merkez düğümleri tespit etmek amacı ile kullanılmıştır [18, 19]. Bu çalışmada metinlerden elde edilen çizgelerde, cümleleri temsil eden düğümler içerisinden en değerli olanları belirlemek amacı ile Arasındalık Merkezliği, Yakınlık Merkezliği, Derece Merkezliği ve Öz değer vektör Merkezliği ölçümleri kullanılmaktadır.

Bir çizgedeki merkezliği ölçmek adına kullanılan yöntemlerden biri arasındalık merkezliğidir. Bu ölçütte temel fikir, x ve y düğümlerinin çoğunda V düğümü yer alıyorsa bu düğümünün önemli bir düğüm olduğudur. Bu yöntem önemli düğümlerin diğer düğümlere bağlandığı varsayımına dayanır [2].

$$C_b(v) = \sum_{x,y \in N} \frac{G_{x,y}(v)}{G_{x,y}} \tag{1}$$

N çizgede bulunan düğümler kümesidir. V düğümüne ait arasındalık merkezliği değeri Denklem 1' deki gibi hesaplanmaktadır.

Yakınlık merkeziyeti, uzaklığın tersi, yani bir düğüm ile tüm diğer düğümler arasındaki en kısa mesafelerin toplamı olarak tanımlanmaktadır. Yakınlık merkezliği Denklem 2' deki gibi hesaplanmaktadır [20].

$$C_c(v_i) = \frac{N-1}{\sum_{v_j \in v} uzaklık(v_i, v_j)} \tag{2}$$

Denklemde kullanılan $uzaklık(v_i, v_j)$ fonksiyonu iki düğüm arasındaki mesafe değeridir. Düğümler arasındaki ağırlık ağırlık değeri toplamı olarak tanımlanmaktadır. Bir düğüm üzerinde meydana gelen bağlantıların sayısı, derece merkezliği olarak tanımlanmaktadır [21, 22]. Verilen bir $G := (V, E)$ çizgesi N adet düğümden oluşmuş olsun ve V düğümünün derece merkezliği değeri C_D (v) Denklem 3'deki gibi hesaplanmaktadır;

$$C_D(v) = \frac{derece(v_i)}{N-1} \tag{3}$$

Bu yöntemde yüksek önem değerine sahip düğümlere olan bağlantıların düşük önem değerine sahip düğümlere olan bağlantılarla aynı puanlamayı önem değerini vermekten ziyade yüksek değere sahip düğümlerin bağlantılı olduğu düğümlere daha fazla önem verdiği prensibine dayanmaktadır [23]. V_j ve V_i düğümler arasındaki kenarlarının ağırlığı ve λ bir sabit olmak üzere, V_i düğümünün öz vektör merkezliliği Denklem 4' deki gibi hesaplanmaktadır [20].

$$C_e(V_i) = \frac{1}{\lambda} \sum_{V_j \in N(V_i)} V_{ji}\ x\ C_e(V_j) \tag{4}$$

2.3. Önerilen Metot

Çalışma kapsamında Metinsel Çizge aşamasında belirtildiği şekilde çizgelerle temsilinin sağlanmasının akabinde bahsedilen düğüm ağırlıklandırma metotlarının bir arada kullanılarak farklı kabuller çalışmaya dahil edilmiştir.

Çalışma kapsamında en güçlü sonuç için (metinlerin genelinde hangi cümlelerin belirleyici olduğuna dair) deneysel süreçler, var olan düğüm merkezlik yöntemlerinin tümü etrafında yürütülerek genel ve bütüncül bir sonuç elde edilmiştir. Çalışma kapsamında yürütülen süreç düğümlere ait dört farklı merkezlik hesabı için ayrı ayrı hesaplanmaktadır. Önerilen yaklaşım gereği, düğümler kullanılan düğüm merkezlik yöntemlerine göre aldıkları puan sırası ile toplanmaktadır.

V_i düğümüne dair önerilen yaklaşım düğümlerin toplam ağırlık değerlerini Denklem 5'teki gibi hesaplanmaktadır.

$$Düğüm Merkezilik Değeri\ (Vi) = \sum_{i=0}^{m} \frac{1}{d_s} \tag{5}$$

Denklemde kullanılan düğüm merkezliği metot sayısı *m* ile kullanılan düğüm merkezliklerine göre düğümlere ait puan sırası ise d_s ile gösterilir. Önerilen denklem ile düğümlere dair farklı merkezlik hesaplama varsayım ve kabullerinin dahil olduğu sayısal düğüm önem değerleri elde edilmiştir. Bu sayede özeti aranan metni oluşturan cümlelerin her biri için metnin bütünü üzerindeki temsil ve önemi elde edilmiştir dahası bu yaklaşım ile cümleler arasındaki ilişkiler yerel bir ilişki düzeyi ile sınırlı kalmaması sağlanmaktadır.

3. Deneysel Çalışma

Çalışmanın bu bölümünde önerilen özetleme yöntemini test etmek amacı ile deneysel süreçler boyunca kullanılan veri seti anlatılmaktadır. Ayrıca, özetleme sistemlerinin doğruluğunu değerlendirmek için kullanılan popüler değerlendirme ölçütü türlerine yer verilmektedir. Son olarak sunulan özetleme yönteminin performansını diğer özetleme sistemlerine kıyasla değerlendirmek için yapılan bir dizi deney sonuçları sunulmaktadır.

3.1. Veri Seti

Bu çalışmada, önerilen özetleme yönteminin doğruluğunu test etmek amacı ile Document Understanding Conference (DUC-2002) veri setinden faydalanılmıştır. Bu veri seti özetleyici ve çıkarımsal özetlemeye yönelik dokümanlar bulundurmaktadır. Yapılan çalışmada çıkarımsal bir özetleme yöntemi önerildiği için çıkarımsal özetleme dosyalarından faydalanılmıştır. DUC-2002 veri setinde farklı değerlendiriciler tarafından çıkarımsal özeti oluşturulan 59 adet dosya bulunmaktadır. Ayrıca bu dosyalara ait 200 ile 400 kelime sayısı sınır alınarak oluşturulan ve her birinden iki farklı değerlendirici tarafından oluşturulan özetler bulunmaktadır.

3.2. Performans Değerlendirme Metriği

Özetleme sistemlerinde kullanılan en popüler değerlendirme ölçütleri Recall-Oriented Understudy for Gisting Evaluation (ROUGE) performans ölçütleridir. ROUGE, özetlerin otomatik olarak değerlendirildiği bir performans metriğidir. Bu ölçütler, n-gram, kelime dizileri bazında değerlendirme yapmaktadır. Özetleme sistemleri tarafından oluşturulan özetler ile insanlar tarafından oluşturulan ideal özetler arasında meydana gelen kesişimlerin baz alındığı bir ölçüttür [24]. ROUGE ölçütleri tarafından üretilen puanlar 0 ile 1 arasındadır. Üretilen puanlar ne kadar yüksek olursa, model özeti ile o kadar fazla paylaşılan içerik mevcut demektir ve bu durumda otomatik özet daha iyi ve daha bilgilendirici olarak kabul edilmektedir. Bu çalışmada, önerilen yaklaşımının performansını değerlendirmek için ROUGE-N (N-1, N-2), ROUGE L, ROUGE-W-1.2 ve ROUGE-SU ölçülerini kullanıyoruz. ROUGE-N, önerilen özet ve model özet arasında paylaşılan n-gram sayısını değerlendirir.

$$ROUGE - N = \frac{\sum_{C \in \{Referans\ Özetler\}} \sum_{gram_n \in S} Eşleşme_{Sayıları}(gram_n)}{\sum_{C \in \{Referans\ Özetler\}} \sum_{gram_n \in S} Eşleşme(gram_n)} \quad (6)$$

Formülde yer alan n, $gram_n$ uzunluğuna eşittir. $Count_{match}(gram_n)$, aday özet ve refrans özette kesişen maksimum n-gram'ların sayısıdır. Denklem 6'te açık bir şekilde görüldüğü üzere Rouge-N, duyarlılık ile ilişkili bir ölçümdür. Çünkü eşitliğin paydası referans özette oluşan n-gramların sayılarının toplamıdır [24]. Benzer şekilde ROUGE -L değerinde verilen iki alt kelime dizgesinde en uzun ortak kelime dizisini hesaplamaktadır. ROUGE-W-1.2 önerilen özet ve model özet arasında art arda meydana gelen en uzun eşleşmeleri hesaplar. ROUGE-SU4 ise iki özet arasında ortak olan skip-bigrams sayısını ölçmektedir.

3.3. Bulgular

Bu çalışmanın temel amacı, çıkarıcı metin özetlemek amacı ile genel, denetimsiz ve çizge tabanlı bir süreç sunmaktır. Önerilen özetleme sisteminin başarımını analiz etmek için DUC-2002 veri setinde yer alan metinlere ait özetler sunan deneysel çalışmalar gerçekleştirilmiştir. Deneysel süreçler boyunca metinlerin üzerinde çalışılabilecek biçimlere aktarılması, ayırt ediciliği olmayan gereksiz verilerden soyutlanması gibi normalizasyon adımları ile mümkün olmuştur. Bu normalizasyon adımlarına ilaveten yazım şekilleri bakımından birbirlerinden farklı olan ancak anlamsal olarak ortak bir kelime kökünden türeyen yakın anlamlı kelimelerin çizgeler oluşturulurken farklı kelimeler gibi değerlendirilmesi, cümleler arasındaki bağlantı ve ilişkilerin tespitini zorlaştırmaması için özetleme öncesinde kelimeler ve ifadeler üzerinde bir dönüşüm sağlanmıştır. Sözü edilen dönüşüm Net framework 4.7'de C# etiketleri kullanılarak gerçekleştirilmiştir.

Elde edilen özetlerin başarısı, literatürde en yaygın olarak kullanılan belirli performans metrikleri hesaplanarak ölçülmüştür ve yaklaşımın performansı geleneksel çizge tabanlı özetleme sistemlerinden olan LexRank ve TextRank özetleme modelleri ile karşılaştırılmıştır.

Cümlelerde yer alan ortak kelimelerden yola çıkılarak metinsel çizge oluşturma adımında düğümlerin cümleleri, ayrıtların ortak kelime sayılarını temsil ettiği çizgeler ve bu çizgelere ait ilişki matrisleri Python etiketleri ile oluşturulmuştur.

Son aşamada elde edilen çizgelerin farklı düğüm merkezlik hesaplamaları (Derece merkezliği, Yakınlık merkezliği, Arasındalık merkezliği ve Öz değer vektör merkezliği) ile çizge içerisindeki puan sıralamaları oluşturulmaktadır. Her bir yaklaşım için 200 kelimeden ibaret olan özetler elde edilmiştir. Gerçekleştirdiğimiz çalışmaya ait özetleme performansını değerlendirmek için DUC-2002 veri setinde yer alan metinlere ait model özet (200 kelimelik) ile bu çalışmada sunulan yöntem tarafından elde edilen sistem özetleri karşılaştırılmaktadır. Bu amaç ile özetleme sonuçlarının performansını ölçmek için ROUGE özetleme performans metrikleri kullanılmıştır. Deneysel çalışmada önerilen sistem özetinin özetleme başarısını ölçmek amacı ile ROUGE-N (N=1, 2, 3, 4) ROUGE-L, ROUGE-W-1-2 ve ROUGE-SU performans metrikleri hesaplanmıştır. Her bir ROUGE metriği için de Duyarlılık, Kesinlik ve F-Skor tiplerinde değerler elde edilmektedir. Çalışmada sunulan modelinin özetleme performansını geleneksel çizge tabanlı özetleme sistemlerinden olan LexRank ve TextRank özetleme modelleri ile karşılaştırılmıştır. Yöntemimizin ve LexRank, TextRank modellerinin 200 kelimelik özetleme performanslarına Tablo 2'de yer verilmiştir.

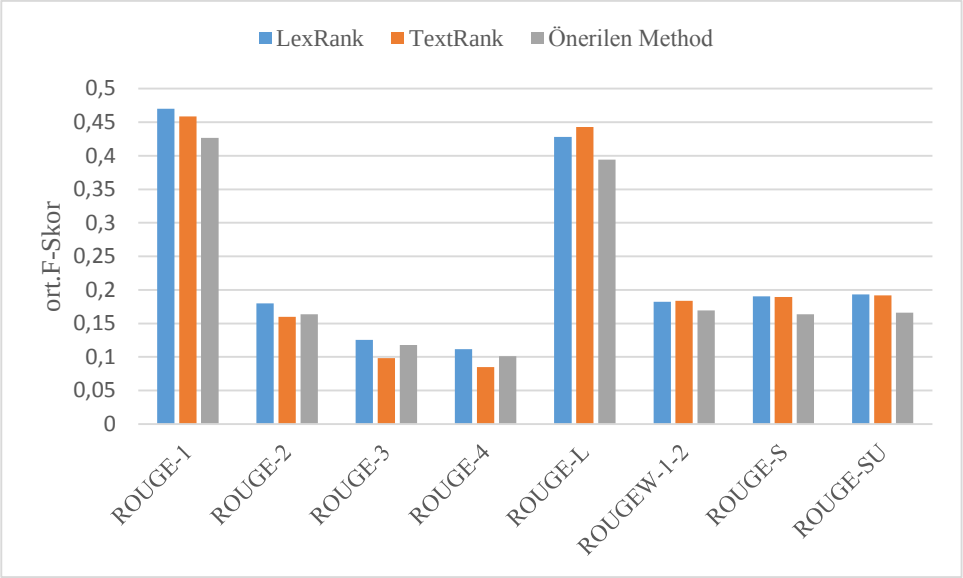
Metrik ölçüm tipleri oldukça farklı değerler almışlardır. Net bir değerlendirme elde edilebilmesi amacı ile ilgili tabloda yer alan tüm metotlar Duyarlılık, Kesinlik ve F-Skor ölçümleri ayrı ayrı baz alınarak karşılaştırılmaktadır.

Tablo 2. 200 kelimeden oluşan özetler için önerilen yöntem ile, geleneksel TextRank ve LexRank modellerinin karşılaştırılması: %95 güven aralığında ROUGE metrikleri kullanılarak ortalama duyarlılık, ortalama kesinlik, ortalama F-skor değerleri

Özetleme Metotları	ROUGE METRİKLERİ (% 95 Güven Aralığı)					
	ROUGE-1			ROUGE-2		
	Duyarlılık	Kesinlik	F-Skor	Duyarlılık	Kesinlik	F-Skor
TextRank	0.48417	0.43685	0.45868	0.16921	0.15195	0.16000
LexRank	0.48124	0.45984	0.46997	0.18507	0.17566	0.18010
Önerilen Metot	**0.44291**	**0.41222**	**0.42668**	**0.17068**	**0.15772**	**0.16383**
	ROUGE-3			ROUGE-4		
TextRank	0.10382	0.09316	0.09816	0.08994	0.08066	0.08501
LexRank	0.12907	0.12227	0.12546	0.11500	0.10910	0.11187
Önerilen Metot	**0.12287**	**0.11292**	**0.11762**	**0.10588**	**0.09694**	**0.10116**
	ROUGE-L			ROUGE-W-1-2		
TextRank	0.46748	0.42183	0.44289	0.15655	0.22316	0.18382
LexRank	0.43813	0.41867	0.42787	0.15201	0.22825	0.18237
Önerilen Metot	**0.40910**	**0.38057**	**0.39402**	**0.14277**	**0.20899**	**0.16954**
	ROUGE-S			ROUGE-SU		
TextRank	0.21283	0.17241	0.18958	0.21541	0.17468	0.19200
LexRank	0.20054	0.18242	0.19049	0.20321	0.18494	0.19309
Önerilen Metot	**0.17762**	**0.15263**	**0.16370**	**0.18015**	**0.15494**	**0.16611**

DUC-2002 veri setinden yer alan metin özetleri ile çalışma kapsamında elde edilen metin özetleri kullanılarak elde edilen ROUGE değerlerine Tablo 2'de yer verilmektedir. Önerilen metot ile elde edilen sonuçlar ilgili tabloda kalın yazı tipi ile gösterilmektedir.

Tablo 2'de görüldüğü üzere tüm metotlar için Duyarlılık ve Kesinlik değerleri arasında büyük dalgalanmalar ve farklılıklar rapor edilmektedir. Ancak F-Skor tipi değerlendirme Duyarlılık ve Kesinlik değerlerine göre daha dengeli değerler almaktadır. Şekil 3 farklı metotlar için tüm ROUGE metrikleri bakımından ortalama F-Skor değerlerini göstermektedir.



Şekil 3. 200 kelimeden oluşan özetler için önerilen yöntem ile, geleneksel TextRank ve LexRank modellerinin karşılaştırılması:%95 güven aralığında ortalama F-Skor değerleri

4. Sonuç ve Öneriler

Bu çalışmada, çıkarıcı metin özetlemek amacı ile genel, denetimsiz ve çizge tabanlı bir süreç sunulmaktadır. Yöntemimiz bir dizi sıralı adımı takip ederek cümleler arasında bir önem sırası oluşturma esasına dayanmaktadır. Önerilen özetleme yaklaşımında, konuşma dilinin beraberinde getirdiği kuralsızlıklar geliştirilen bir yazılım aracı ile giderilmekte ve bu yazılım aracının kendine özgü algoritması ile metinler özetleme için hazırlanmaktadır. Özetin oluşturulması için düğümlerle temsil edilen cümlelerin önemini, sayısallaştırabilmek amacı ile belirlenen düğüm merkezlik ölçüm değerleri hesaplanmıştır. Her bir düğüm merkezlik puanı ayrı ayrı hesaplanmış ve önerilen yaklaşım ile bir hesaplama gerçekleştirilmiştir. Kullanılan merkezlik ölçümleri bakımından en yüksek değerleri alan cümleler seçilerek boyutları 200 kelime ile sınırlandırılmış özetler oluşturulmaktadır. Önerilen yaklaşımın metin özetleme probleminde kullanılabileceğini gösteren deneysel sonuçlara ilgili tablo ve grafikte yer verilmiştir. Ayrıca bu çalışma kapsamında sistemimizin performansı geleneksel LexRank ve TextRank özetleme yöntemleri ile karşılaştırılmıştır.

Genel olarak ilgili tablo ve grafikte görüldüğü üzere geleneksel LexRank özetleme metodu yine geleneksel bir başka özetleme metodu olan TextRank ve çalışma kapsamında sunduğumuz yöntemden daha iyi ROUGE performans değerleri rapor etmektedir. Ayrıca çalışma kapsamında sunulan yöntem ile elde edilen ROUGE değerleri bakımından geleneksel TexRank özetleme metodu birçok metrik türü bakımından (ROUGE-2, ROUGE-3, ROUGE-4) geride bırakılmıştır. Önerilen yöntem ile LexRank özetleme yönteminin karşılaştırılması durumunda ise LexRank yönteminin daha yüksek ROUGE değerleri rapor ettiği açıkça gözlemlenmektedir. Deneysel süreçler boyunca elde edilen bir başka bulgu metin özetleme sistemlerinde sıklıkla kullanılan LexRank yaklaşımının, yine sıklıkla kullanılan TextRank yaklaşımından daha yüksek ROUGE değerleri rapor ettiğidir.

Açıkça ilgili tablo ve şekillerde ifade edildiği gibi kullanılan performans metrikleri açısından elde edilen ümit verici sonuçlar neticesinde, önerilen yöntemin metin özetlemede kullanışlı ve etkili bir araç olabileceği açıkça görülmektedir. Çalışma kapsamında yer verilen bütün düğüm merkezlik ölçüm yöntemleri farklı kabul ve varsayımlar üzerinden merkezi düğümleri tayin etmeye çalışılmıştır. Bu çalışmada ilgili merkezlik yöntemlerinin herhangi birine bağlı kalınmamış bir çok yöntemin dahil edildiği bütüncül bir yenilik sunulmuştur. Ayrıca deneysel çalışmaların sonuçlarına göre KUSH yazılımının araştırmacılar tarafından sıklıkla kullanılan sınıflandırma ve kümeleme yöntemleri öncesinde de kullanılabilecek bir araç olabileceğine inanıyoruz Bu çalışmada farklı kabullere göre çalışan yöntemlerin tamamının gücünü kullanarak basit, uygulanabilir ve rekabetçi bir yöntem sunulmuştur.

Kaynaklar

[1] O. Durmaz. 2011. Metin Sınıflandırmada Boyut Azaltmanın Etkisi ve Özellik Seçimi. Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

[2] Hark C., Uçkan T., Seyyarer, Abubekir Karci A. 2018. Metin Özetleme İçin Çizge Tabanlı Bir Öneri. IDAP 2018 - International Artificial Intelligence and Data Processing Symposium, pp: 1-6.

[3] Canberk G., Sağıroğlu Ş. 2006. Bilgi ve Bilgisayar Güvenliği : Casus Yazılımlar ve Korunma Yöntemleri. Grafiker Yayıncılık, Ankara.

[4] Aydemir E. 2018. Weka ile Yapay Zeka. Seçkin Yayınevi, Ankara, 216s.

[5] Radev D.R., Hovy E., McKeown K. 2002. Introduction to the special issue on summarization. Comput Linguist, 28 (4): 399-408.

[6] Erkan G., Radev D.R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. J Artif Intell Res., 22: 457-479.

[7] Joshi A., Fidalgo E., Alegre E., Fernández-Robles L. 2019. SummCoder: An Unsupervised Framework for Extractive Text Summarization Based on Deep Auto-encoders. Expert Syst Appl., doi:10.1016/j.eswa.2019.03.045.

[8] Kaynar O., Görmez Y., Işık Y.E., Demirkoparan F. 2017. Comparison of graph based document summarization method. 2017 International Conference on Computer Science and Engineering (UBMK), pp: 598-603.

[9] Cigir C., Kutlu M., Cicekli I. 2009. Generic text summarization for Turkish. 2009 24th International Symposium on Computer and Information Sciences (IEEE), pp: 224-229.

[10] Sarkar K., Saraf K., Ghosh A. 2015. Improving graph based multidocument text summarization using an enhanced sentence similarity measure. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS 2015 - Proceedings, pp: 359-365.

[11] Mihalcea R., Tarau P. 20041. TextRank: Bringing Order into Texts. Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions - (Association for Computational Linguistics, Morristown, NJ, USA), pp: 20-es.

[12] Parveen D., Ramsl H.-M., Strube M. 2015. Topical coherence for graph-based extractive summarization. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp: 1949-1954.

[13] Hark C., Seyyarer A., Uçkan T., Karci A. 2017. Doğal dil işleme yaklaşimlari ile yapisal olmayan dökümanlarin benzerliği. IDAP 2017 - International Artificial Intelligence and Data Processing Symposium, pp: 1-6.

[14] Mihalcea R., Tarau P. 2005. A Language Independent Algorithm for Single and Multiple Document Summarization. Proceedings of IJCNLP 2005, 2nd International Join Conference on Natural Language Processing, pp: 19-24.

[15] Karci A. 1998. Çizge Algoritmaları ve Çizge Bölmeleme. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ.

[16] Von Luxburg U. 2007. A Tutorial on Spectral Clustering Available at: www.springer.com. (Acces date: 24.12. 2018).

[17] Bavelas A. 1948. A Mathematical Model for Group Structures. Hum Organ, 7 (3): 16-30.

[18] Radev D.R. 2004. LexRank : Graph-based Lexical Centrality as Salience in. Artif Intell, 22: 457-479.

[19] Kutlu M., Cigir C., Cicekli I. 2010. Generic text summarization for Turkish. Comput J., 53 (8):1315-1323.

[20] Boudin F. 2013. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction. International Joint Conference on Natural Language Processing (IJCNLP) (Nagoya, Japan), pp: 834-838.

[21] McPherson M., Smith-Lovin L., Cook J.M. 2001. Birds of a Feather: Homophily in Social Networks. Annu Rev Sociol., 27 (1):415-444.

[22] Analysis B.N. 2016. Centrality and Hubs. (1979). doi:10.1016/B978-0-12-407908-3.00005-4.

[23] Kosorukoff A. 2011. Social Network Analysis Theory and Applications (Passmore, D. L, 2011).

[24] Lin C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. Text Summ Branches Out.