DOKUZ EYLÜL ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ
FEN VE MÜHENDİSLİK DERGİSİ
Cilt/*Vol.*:17■ No/*Number*:2■Sayı/*Issue*:50■Sayfa/*Page*:95-108■ Mayıs 2015 / *May 2015*

Makale Gönderim Tarihi (Paper Received Date): 11 Mart 2015 Makale Kabul Tarihi (Paper Accepted Date): 15 Mayıs 2015

TÜRKÇE METİN ÖZETLEMEDE MELEZ MODELLEME

(A HYBRID MODELLING FOR TURKISH TEXT SUMMARIZATION)

Akif HATİPOĞLU[1] Sevinç İlhan OMURCA[2]

ÖZ

Orijinal belgelere ait en önemli cümlelerin belirlenmesi için gerçekleştirilen bilgisayar programı ile otomatik metin özetleme işlemi bir doğal dil işleme problemidir. Doğal dil işlemede temel olarak iki çeşit metin özetleme yaklaşımı bulunmaktadır. Bu yaklaşımlar cümle seçerek özetleme ve yorumlayarak özetleme olarak ikiye ayrılmaktadır. Cümle seçerek özetleme iki farklı alt yönteme ayrılmaktadır. Birincisi özetlenecek metinde cümleleri istatistiksel olarak puanlandırma yöntemidir. İkinci yöntem ise sezgisel olarak gizli anlam çıkarımı yöntemidir. Özetleme çalışmalarında bu iki yöntemi birleştirip melez bir yapı kurularak özetleme gerçekleştirilmektedir. Bu makale kapsamında cümle seçerek özetleme hedeflenmiştir. Türkçenin yapısal özelliklerine göre istatistiksel olarak puanlandırılması ve gizli anlam çıkarım yöntemlerini sezgisel olarak birleştirerek cümle seçimi yapan melez bir model sunulmuştur.

Anahtar Kelimeler: Metin özetleme, Gizli Anlam Analizi, Doğal dil işleme

ABSTRACT

Automatic text summarization with a computer program in order to retain the most important sentinces of the original documant is a natural language processing problem. There are basically two types of text summarization approaches in natural language processing. These approaches are defined as summarization by selecting sentences and summarization by interpreting sentences. The summarization by selecting sentences method is also seperated into two sub-methods. The first one is the medhod of scoring the sentences statistically. The second one is the method of latent semantic analysis of the sentences. In this study, summarization is realized by selecting sentiences. A hybrid model which selects the sentences by combining two methods: statistically scoring sentences due to structural featuers of Turkish and latent semantic analysis method.

Keywords: *Text summarizition, Latent semantic analysis, Natural language processing*

[1] Kocaeli Üniversitesi, Mühendislik Fakültesi, Bilgisayar Bölümü, Kocaeli, akifhatipoglu.tech@gmail.com (sorumlu yazar)
[2] Kocaeli Üniversitesi, Mühendislik Fakültesi, Bilgisayar Bölümü, Kocaeli, silhan@kocaeli.edu.tr

1. GİRİŞ

İnternet kullanımının artması ile birlikte bilgi kaynakları da çoğalmıştır. Günümüzde bir konu hakkında araştırma yapıldığında bu konu ile ilgili farklı birçok kaynağa kolayca ulaşılmaktadır. Doğal olarak bu kaynakların hepsini inceleme fırsatı yakalanamamaktadır. Bilgi kaynakları arttıkça incelenmek üzere depolanan metinlerin sayısı ve uzunlukları da artmaktadır. Bu açıdan, metin özetleme ile incelenmesi zaman alan uzun metinleri daha anlamlı kısa metinler haline getirmek büyük önem taşımaktadır. Tüm bunların sonucunda metin özetleme çalışmaları günümüzde büyük önem kazanmıştır.

Literatürde bilgisayar yardımı ile metin özetleme işlemine ait yaygın olarak uygulanan iki farklı yaklaşım bulunmaktadır. Birincisi cümleyi yorumlayarak özetleme, ikincisi ise seçerek özetleme yaklaşımıdır. Yorumlayarak özetleme, metni kelime bazında anlamlandırıp yeni cümleler türetmeye dayalı bir yaklaşımdır. Bu yaklaşımda cümleler anlamlı olarak birleştirilerek yeni bir metin oluşturulmakta veya kelimeler anlamlı olarak birleştirilerek yeni cümleler oluşturulmaktadır. Seçerek özetleme yaklaşımı ise iki farklı alt yöntemden oluşmaktadır. Birincisi özetlenecek metinde cümleleri istatiksel olarak puanlandırma yoluyla cümle seçimidir. İstatistiksel puanlandırma, cümlelerin doğal dil açısından önemine göre cümleye puan atamaya dayanmaktadır. İstatistiksel puanlandırma çeşitleri şu şekilde açıklanmaktadır. Bir cümlenin metinde bulunduğu pozisyon o cümle için bir puan oluşturmaktadır. Aynı şekilde bir cümlede geçen kelime sayısı içerdiği bilgi açısından bir puan oluşturmaktadır. Bir cümlenin içerisinde o metinin başlığında geçen kelimeler bulunuyor ise o cümle özet cümle olmak için daha güçlü aday kabul edilip belirli bir puan almaktadır. Son olarak cümlede geçen kelimelerin frekansı hesaplanmaktadır. Bu frekans, kelimenin hem cümle içerisinde geçme hem de metin içerisinde kullanım sıklığı ile belirli bir değeri oluşturmaktadır. Bu frekans değerleri ile belge terim matrisi kurulup her cümle için belirli bir puan oluşturulmaktadır [1].

Cümle seçimi ile metin özetlemede ikinci yöntem ise sezgisel olarak gizli anlam çıkarımıdır ya da gizli anlam analizi olarak da adlandırılmıştır. Bu yöntem birliktelik kuralları ile fark edilemeyen bağlantıları çözümleyip gözlemlememizi sağlamaktadır. Bu yöntemde öncelikle kelime cümle matrisi kurulur. Bu matris, kelimelerin cümlelere göre frekanslarından oluşmaktadır. Bu kurulan matris üzerinde tekil değer ayrıştırılması yöntemi uygulanarak daha küçük yönetilebilir bir matris elde edilir. Bu matris bize metindeki kavramları vermektedir. Bu kavramlara göre cümlelere puan verilir.

Otomatik metin özetleme literatürde sık çalışılmış bir alan olmakla birlikte Türkçe metinler üzerinde gerçekleştirilen çok fazla çalışma bulunmamaktadır [2, 3, 4].

Bu çalışmada, Türkçe metinler üzerinde, cümlenin yapısal ve anlamsal özelliklerini melez şekilde birleştirerek otomatik metin özetlemeyi gerçekleştiren bir model geliştirilmiştir [2]. Cümlenin uzunluğu, belge içindeki pozisyonu, dokumana ait başlık kelimelerini içerme sıklığı ve içerdiği terimlerin ağırlıkları cümlenin yapısal özelikleri olarak belirlenmiş; cümlenin anlamsal özellikleri ise LSA temelli özellikleri ile belirlenmiştir. Özetlenecek metine ait her cümle için yapısal ve anlamsal özellikler skorlanmış ve sezgisel olarak belirlenen özellik katsayıları ile çarpılarak lineer bir model ile birleştirilmiştir. Uygulanan bu melez yöntemin sonucunda en yüksek puan alan cümleden en düşük puan alan cümleye doğru sıralanmış bir liste elde edilmektedir. Sıralanmış cümle listesi içerisinden kullanıcı tarafından belirlenen sayıda cümle seçilip kullanıcıya özet cümleler olarak sunulmaktadır. Deneysel çalışmalar

kısmında geliştirilen sistemi test etmek üzere Türkçe Vikipedi metinleri kullanılmış; seçilen Vikipedia belgesinin özet metni gösterilmiştir.

Liteartürde otomatik metin özetleme sistemlerinin geçerlilik testinin güç bir işlem olduğu yönünde değerlendirmeler yapılmaktadır [5]. Bu çalışmada modelin performansı, özet metin için seçilen cümleler ile 10 ayrı kişinin seçtiği cümleler karşılaştırılarak değerlendirilmiştir.

Makalenin geri kalanı şu şekilde organize edilmiştir: Bölüm 2'de uygulanan metin özetleme yöntemi açıklanmaktadır, bölüm 3'de deneysel sonuçlar raporlanmakta ve bölüm 4'de çalışmadan elde edilen sonuçlara yer verilmektedir.

2. METİN ÖZETLEME

İlk olarak özetlenilecek metin üzerinde veri ön işleme adımları gerçekleştirilir.Bu adımlar gereksiz kelimelerin elenmesi, metin içerisindeki cümlelerin ayrıştırılması ve son olarak kelimelerin köklerinin bulunmasıdır. Veri ön işleme adımları gerçekleştirildikten sonra Yapısal olarak istatistiksel puanlandırma yöntemleri ve sonrasında gizli anlamsal analiz (GAA) aşamaları gerçekleştirilmiştir.

2.1. Metin Önişleme

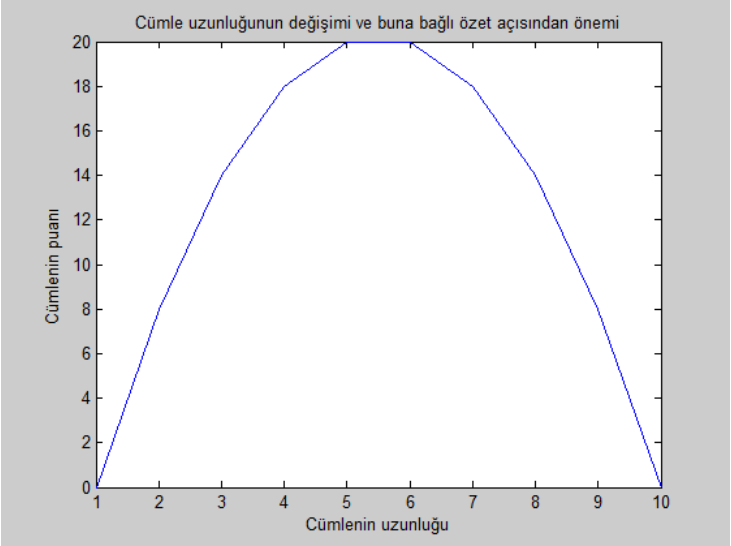
Ön işlemenin ilk adımı olarak metin içerisindeki gereksiz kelimeler elenir. Bunlar genel olarak o dilde sık kullanılan kelimeler, içerikten bağımsız kelimeler, bağlaçlar, imleçler, sayılar, kalıplaşmış kısaltmalardır. Ayrıca noktalama işaretleri ve tek başına duran alfabedeki harfler de gereksiz kelimeler listesine eklenebilir [6, 7].

Sonraki adımda, metin içerisindeki cümleler ayrıştırılır. Bu işlem Zemberek kütüphanesi yardımı ile gerçekleştirilmiştir [8]. Zemberek kütüphanesi açık kaynak kodlu Türkçe doğal dil işleme kütüphanesidir. Platform bağımsız olması ve Türkçe ile ilgili problemlere çözüm getirmesi amacı taşıyan bir projedir. Zemberek dil hatalarına uyarı vermek, ön izleme, düzletme, heceleme, köklerine ayırma, köklerini inceleme, deyimleri belirleme, ascii dönüşümleri gibi birçok işlevi olan Türkçe ile ilgili en detaylı projedir. Zemberek kütüphanesinin "Sentence boundary detector" fonksiyonu sayesinde bir metin içerisindeki tüm cümleler ayrıştırılabilmektedir. Ancak ayrıştırılan cümlelerde noktalama işaretlerinden kaynaklanan hatalar olmaktadır. Örneğin ":" karakteri cümle sonu gibi algılanarak Zemberek tarafından yanlış cümle ayrıştırmalarına denen olmaktadır. Zemberekten kaynaklanan bu ve benzeri eksiklikler geliştirilen kod ile düzeltilmiştir.

Cümleler ayrıştırıldıktan sonra cümlelerde geçen her bir kelime köklerine ayrıştırılır.Türkçe Ural–Altay dil ailesinden gelmektedir ve sondan eklemeli bir dildir. Türkçede bir kelimenin kökünden sonra farklı ekler gelerek kökten bağımsız farklı anlamlar elde edilebilir. Bu nedenle metin özetleme işlemi gerçekleştirilirken bütün kelimelerin kelime köklerinin bulunması gerekmektedir. Doğal dil işleme işlemlerinin gerçekleştirilmesi için gereken kelime-cümle matrisi kurulurken ve bu matrise bağlı işlemler yapılırken bütün kelimelerin köklerinin bulunmuş olması gerekmektedir. Tüm kelimelerin köklerine ayrıştırılması aşaması Zemberek kütüphanesi yardımı ile yapılmıştır. Her cümleye ait kelime kökleri elde edildikten sonra cümleler üzerinde puanlandırma aşaması gerçeklenmiştir.

2.2. Cümle Uzunluğu

Cümlenin uzunluğunu temsil eden özellik vektörü bir cümlenin içerisinde barındırdığı kelime sayısını ifade etmektedir. Veri ön işleme aşamasından sonra cümle içerisinde cümleye ait kelimelerin kökleri yer almaktadır. Kelime köklerinin sayısı o cümleye ait puanlandırmada kullanılmaktadır [1, 2]. Metin içerisindeki tüm cümleler için ortalama cümle uzunluğu hesaplanır. Bu ortalama cümle uzunluğuna yakın uzunlukta olan cümleler daha önemli cümleler olarak kabul edilir. Ortalamadan daha uzun cümlelerin ayrıntı içeren cümleler olması, daha kısa olan cümlelerin ise yeterli bilgi taşımıyor olması olasılıkları bu cümleleri özet metin için daha zayıf adaylar haline getirmektedir.



Şekil 1. Cümle uzunluğunun özet açısından önemi

Cümlenin uzunluğunun cümle puanına etkisi Şekil 1'de gösterilmiştir. Bu kabulü şu şekilde formüle edebiliriz. Cümle uzunlukları bulunduktan sonra cümle uzunluklarının ortalaması bulunur. Bu ortalamaya sahip ya da bu ortalamaya yakın cümleler metin özetleme için önemlidir. Çok kısa cümleler veya çok uzun cümleler metin özetleme açısından daha az öneme sahiptir.
D={$S^{(1)}, S^{(2)}, \dots, S^{(m)}$}belgesinde geçen bir $S^{(i)}$ cümlesi için "cümle uzunluğu" özelliği Eşitlik 1'deki gibi formüle edilmiştir.

$$S^{(i)}\_puan_{f_{i1}} = \left(uzunluk_{S^{(i)}} - {}^{1}/_{2m}\left(\sum_{j=1}^{m} luzunluk_{S^{(j)}}\right) - 1/2\right)^2 \quad (1)$$

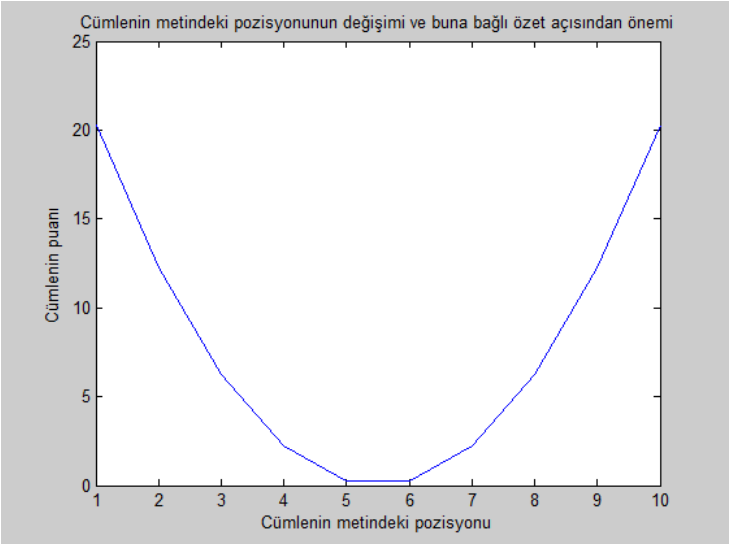
$$S^{(i)}\_puan_{f_{i1}} = -1 \times S^{(i)}\_puan_{f_{i1}} + max_{S^{(i)}\_max_{f_{i1}}}\left\{S^{(1)}\_puan_{f_{i1}}, S^{(2)}\_puan_{f_{i1}}, \dots, S^{(m)}\_puan_{f_{i1}}\right\} \quad (2)$$

Burada $S^{(i)}\_puan_{f_{i1}}$ $S^{(i)}$ cümlesine ait uzunluk özelliğinin hesaplanan değeridir. m belgede geçen cümle sayısını ifade etmektedir.

2.3. Cümlenin Metindeki Pozisyonu

Cümlenin pozisyonunu temsil eden özellik vektörü bir cümlenin metin içerisindeki pozisyonunu ifade etmektedir. Türkçe metinler giriş, gelişme ve sonuç bölümlerinden oluşmaktadırlar. Verilen bir konuyu açıklayabilmek için önce açıklanması gereken düşünce bulunur ve bu düşünce giriş bölümünde belirtilir. Giriş bölümünde sadece açıklanması gereken düşünce kısa ve öz bilgiler ile belirtilir; olay örneklenmez veya açıklama yapılmaz. Gelişme bölümünde, giriş bölümünde belirtilen düşünce daha genişletilmiş ve gerekirse örneklerle zengişleştirilmiş şekilde açıklanır. Gelişme bölümü giriş bölümünden daha uzun yazılır, konu ayrıntıları ile anlatılır. Sonuç bölümünde, giriş ve gelişmenin ortak düşüncesi yani ana düşünce yazılır. Bu bölüm yazılı anlatımın diğer bir kısa bölümüdür, sadece çıkarılan ana düşünceyi anlatır.

Bu çalışmada cümleye ait pozisyon özelliği, her cümlenin özetlenmesi gereken metnin başlangıcına olan uzaklığını temsil etmektedir. Cümlenin başlangıca olan uzaklığı değiştikçe cümlenin ağırlığı da değişmektedir. Türkçede metinlerin giriş-gelişme-sonuç şeklinde yazıldığı düşünüldüğünde; çalışma kapsamında cümlenin pozisyonu metnin başlangıcına yaklaştıkça ya da metnin sonuna yaklaştıkça özet açısından önemi artmaktadır şeklinde bir varsayım ile hareket edilmiştir.



Şekil 2.Cümle pozisyonunun özet açısından önemi

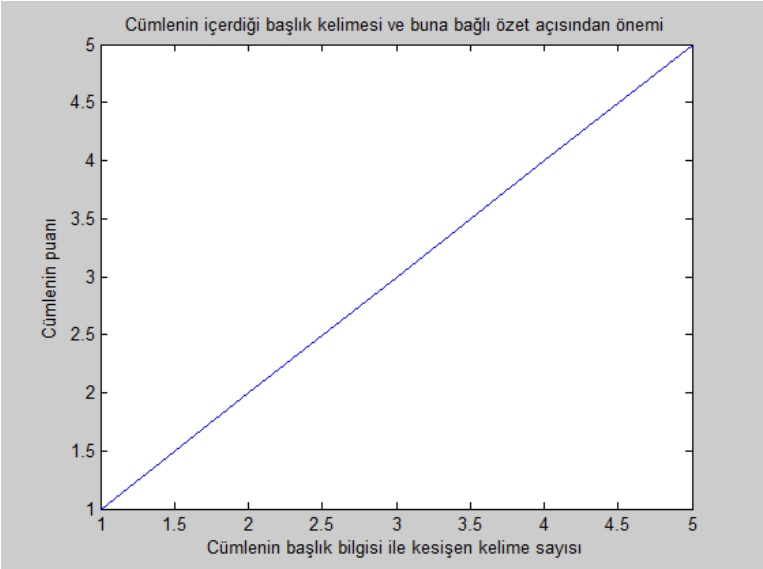
Şekil 2'de gösterildiği gibi cümlenin pozisyonu değiştikçe özet açısından önemi de değişmektedir. Bütün cümlelerin metnin başlangıcına olan uzaklığı o cümlenin pozisyonunu vermektedir. Bu pozisyon değişimi cümlenin giriş, gelişme, sonuç, bölümlerinde olabileceğini göstermektedir. Cümlenin pozisyonu metinin ortasına göre arttıkça veya azaldıkça özet açısından önemi değişmektedir. Eğer cümlenin metin içerisindeki pozisyonu metinin ortalarına karşılık geliyor ise cümlenin gelişme bölümünde bulunma ihtimali artmaktadır. Cümlenin pozisyonu metnin başlangıcına yaklaştıkça o cümlenin giriş bölümünde bulunma ihtimali artmaktadır. Cümlenin pozisyonu metnin sonuna doğru yaklaştıkça cümlenin sonuç bölümünde bulunma ihitimali artmaktadır[1, 2]. Bu durum Eşitlik 3'te formüle edilmiştir.

$$S^{(i)}\_puan_{f_{i2}} = \left(pozisyon_{S^{(i)}} - {}^{1}/_{2}\left(pozisyon_{S^{(m)}}\right) - 1/2\right)^{2} \quad (3)$$

Burada $S^{(i)}_puan_{f_{i2}}$S$^{(i)}$ cümlesine ait pozisyon özelliğinin hesaplanan değeridir. $pozisyon_{S^{(m)}}$ belgenin son cümlesinin posizyon değeridir.

2.4. Başlık Kelimeleri

Başlık kelimelerini temsil eden özellik vektörü her cümlenin sayfanın başlık bilgisi ile kesişen kelimelerin sayısını ifade etmektedir. Eğer cümle başlıkta geçen kelimeyi içeriyor ise bu cümle metin özetleme açısından önemlidir ve cümle bu varsayıma göre puanlandırılır [1, 2]. Bir cümle içerisinde birden fazla başlık ile kesişen kelime bulunabilir. Her cümle için, başlık bilgisi ile kesişen kelime sayısı ile cümlenin puanı doğru orantılıdır.



Şekil 3. Cümlenin içerdiği başlık kelimesi ve özet açısından önemi

Şekil 3'de cümlenin başlık bilgisi ile kesişen kelime sayısının, cümlenin özeti açısından önemini gösterilmektedir.

2.5.Cümlelerin Ağırlıklandırılması

Cümlenin ağırlığını temsil eden özellik vektörü her terimin her cümle için ağırlık bilgisini temsil etmektedir. Terimlerin cümledeki ağırlık bilgisi Tf-tbf (Terim frekansı – ters belge frekansı) yöntemi ile hesaplanmakta ve belge – terim matrisi kurulmaktadır [1, 2, 9, 10]. Bütün cümlelerin içerdikleri terimlerin belirli bir frekansı bulunmaktadır. Tf-tbf ağırlıklandırma yöntemi metinde geçen terimlerin çıkarılması ve bu terimlerin frekanslarına göre çeşitli hesaplamaların yapılması üzerine kurulmuştur. Tek bir belgenin özetlenmesi yapıldığından dolayı belge yerine cümleler kullanılmıştır. Bütün metinde geçen kelimeler üzerinde Tf-tbf ağırlıklandırılması yapılmakta ve terim – cümle matrisi oluşturulmaktadır. Metinde bir kelime kökü birden çok cümlede geçebilmektedir. Bu yüzden ilk önce metin içerisinde birden çok yerde geçse dahi tekil kabul edilen kelimeler bulunmakta ve bir listede tutulmaktadır. Bu kelimelerin listesi terim–belge matrisindeki terimleri oluşturmaktadır. Terim–belge matrisinde geçen belge bilgisi ise cümleler olarak belirlenmektedir. Terim–belge matrisinin hesaplanmasında her terimin cümle ile ilgili olan ağırlık bilgisi hesaplanmaktadır. Ağırlık bilgisi iki değişkene bağlıdır. Ağırlık bilgisi hesaplanırken ilk değişken bir kelimenin

cümlede geçme sıklığının bulunmasıdır; ikinci değişken ise bir kelimenin metinde geçme sıklığının bulunmasıdır. Bu işlem tüm kelimeler için gerçekleştirilir. Bir cümlede sıkça geçen ancak diğer cümlelerde pek bulunmayan terimin ağırlığı daha yüksek olur. Yüksek ağırlıklı kelimeler içeren cümleler özet için daha önemlidir. Bu varsayım Eşitlik 4'de formüle edilmiştir.

$$\begin{aligned} tf &= terimin\ cümlede\ geçme \div cümkedeki\ toplam\ terim\ sayısı \\ tbf &= \log_{10}(toplam\ cümle\ sayısı \div terimi\ bulunduran\ cümle\ sayısı) \\ \mathrm{Tf-tbf} &= tf \times tbf \end{aligned} \tag{4}$$

Her tekil kelime ve bağlı olduğu cümle için $\mathrm{tf-tbf}$ değeri hesaplanmaktadır [2, 9]. Terim–cümle matrisinin sütunundaki değerlerin toplamı her cümle için hesaplanır. Yani cümleye ait kelimelerin ağırlıklarının toplamı o cümlenin toplam ağırlığını vermektedir. Bu toplam ağırlık puanlandırma ile doğru orantılıdır. Ağırlık yöntemi metininin içerdiği cümle sayısı ya da cümlelerin içerdiği kelime sayısı arttıkça daha iyi sonuçlar vermektedir.

2.6. Gizli Anlamsal Analiz

GAA birliktelik kuralları ile fark edilemeyen bağlantıları çözümleyip gözlemlememizi sağlamaktadır [2, 9, 11]. Bu yöntem belgelerde yer alan her cümle içerisindeki terimlerin benzerliklerinden kavramlar çıkartmaya dayanmaktadır. Kavramlar, belgenin konusu hakkında bilgi veren doğal dildeki kelime (terim) ya da kelime öbekleridir. GAA ise terim benzerliklerine dayalı matris ayrıştırma tekniğidir. Cümle-terim matrisi GAA aşamasında da kullanılmaktadır. Cümle-terim matrisi tekil değer ayrışımı (TDA) yöntemi ile daha küçük yönetilebilir bir matrise dönüştürülmektedir. İndirgenmiş matris, metin ile ilişkili terim gruplarını kapsamaktadır. Bu matriste bulunan terimler gruplanarak birbirleri ile ilişkili terim grupları saptanabilmektedir. GAA girdi matrisinin kurulması, TDA ve cümle seçimi olmak üzere üç aşamadan oluşmaktadır.

Girdi matrisi üzerinde yapılacak olan işlemler her kelimenin cümleye göre frekansına bağlı olduğundan, çalışma kapsamında bu aşamada tf-tbf matrisi kullanılmıştır. Her tekil kelimenin, geçtiği cümlelere göre tf-tbf değerleri hesaplanarak terim-cümle matrisi kurulmuştur.

Bir önceki adımda kurulmuş olan girdi matrisi m × n boyutlu bir cümle-terim matrisidir. Matris boyutları m > n şeklindedir yani matrisin kelime sayısı cümle sayısından fazladır. TDA bu girdi matrisinin öz değer matrisi ve öz vektörleri matrislerini elde etmemizi sağlamaktadır. Bu aşamadan sonra girdi matrisi, A matrisi olarak adlandırılmıştır. A matrisinin rankı r ≤ min (m , n) olarak gösterilmektedir. A matrisi,

$$A = USV^T \tag{5}$$

olarak TDA algoritması tarafından çözümlenir. Burada U matrisi m × r sütun ortogonal matrisi olup sol tekil vektörü bize verir. V^T matrisi r × n ortonormal matrisi verir ve bu matris sağ tekil vektörüdür. Bu sağ ve sol vektörel matrisler A matrisimizin öz vektör matrislerini oluşturmaktadır. S matrisi ise r × r tekil değerlerin diagonal matrisidir [12].

Diğer bir açıdan bakıldığında TDA A matrisinin indirgenmiş matrislerini vermektedir. A matrisi r tane kavram içermektedir. S matrisi seçilmiş konuların önemini belirtmekte, bir diğer

deyişle kavramları vermektedir. V^T matrisinin her sütunu, metindeki cümleleri; satırları ise kavramları vermektedir. U matrisi ise r tane konu için m adet terim arasındaki eşleşmeyi vurgulamaktadır [2, 9].

TDA algoritmasının uygulanmasında "Apache Commons Math" açık kaynak kodlu kütüphanesi kullanılmıştır. Bu kütüphanenin seçilme nedeni kıyaslama testlerinde diğer kütüphanelere göre daha performanslı çıkmasıdır. "Apache Commons Math" kütüphanesi büyük matrisler üzerinde TDA uygulamasının daha performanslı yürütüldüğü gözlendiğinden seçilmiştir.

GAA'nın ilk aşamasında girdi matrisi oluşturulmuştur. Girdi matrisi kullanılarak TDA aşaması gerçekleşmiştir. Bu aşama sonucunda üç farklı matris elde edilerek girdi matrisi bu üç matrise indirgenmiştir. Bu aşamalar sonrasında indirgenmiş matrisleri kullanarak cümle seçimi gerçekleştirilmiştir. Cümle seçimi aşaması cümlelere puan vermeye dayalı bir aşamadır. En yüksek puana sahip cümleler metin özetleme açısından daha önemlidir. Cümle seçimi aşamasında çapraz yöntem ve benzerlik yöntemi melez olarak kullanılmıştır. Bu iki yöntem cümleleri puanlandırmada özellik vektörünün devamı niteliğindedir.

2.6.1. Çapraz Yöntem

Çapraz yöntem TDA aşamasından sonra elde edilen V^T matrisini kullanarak hesaplama yapmaktadır. V^T satırları kavramları ve sütunları cümleleri temsil eden kavram-cümle matrisidir. Bu yöntemde öncelikle her satırın ortalaması hesaplanmaktadır. Hesaplanan ortalama değer kavram için eşik değeri kabul edilmektedir. Bu ortalamanın altında kalan kavramların değerleri sıfır yapılmaktadır. Sonrasında yeni oluşturulan V^T matrisinin her sütunu için yani her cümle için o cümlenin uzunluğu hesaplanmaktadır. Bu işlemler Eşitlik 6'da formüle edilmiştir.

$$S^{(i)}\_puan_{f_{i5}} = \sqrt{V^T_{(i,j)} \times S_{(j,j)}} \tag{6}$$

Her cümle için çapraz yöntemden elde edilen ağırlık bilgisi, V^T matrisin o sütunundaki değerleri ile S matrisinin ilgili sütununa ait öz değeri kullanılarak hesaplanmaktadır [2, 8, 9, 10].

2.6.2. Benzerlik Yöntemi

Benzerlik yöntemi TDA aşamasından sonra elde edilen U, S, V^T matrislerini kullanarak hesaplama yapmaktadır. TDA aşamasından sonra elde edilen indirgenmiş matrisler, k maksimum tekil değerler kadar yeniden indirgenir. Eski U, S, V^T matrisleri indirgenerek yeni bir $A_k = U_k S_k {V_k}^T$ matrisi elde edilir.

İndirgeme işlemi tamamlandıktan sonra her cümle için cümle ilişki bağlantısı kurularak her cümle arası bağlantı oluşturulur. A_k matrisinin sütunları cümle vektörleridir; cümlelerin bağlantıları eşit benzerlikte olurlarsa geçersiz sayılırlar. A_k matrisine ait bütün cümle vektörleri arasındaki benzerlikler Eşitlik 7'de gösterildiği gibi kosinüs benzerliğine göre hesaplanmaktadır.

$$S^{(i)}_{puan_{f_{i6}}} = Benzerlik\left(S_i, S_j\right) = \frac{\overrightarrow{S_i} \times \overrightarrow{S_j}}{|\overrightarrow{S_i}| \times |\overrightarrow{S_j}|} \quad (7)$$

Bütün i ve j cümleleri için benzerlik puanları hesaplandıktan sonra bu benzerlik değerleri cümleler arasındaki bağlantıya ait olan değerlerdir [2, 9, 13, 14].

3. Deneysel Sonuçlar

Uygulamamızda özetlenecek metni oluşturan cümlelere yapısal analiz ve GAA açısından puan ataması gerçekleştirilmiştir. Cümlenin genel özellik vektörü kurulurken melez bir yapı esas alınmıştır. Kurulan özellik vektörü içerisinde altı farklı özellik bulunmaktadır. Özellik f_{ij} i. cümleye ait j. özelliği temsil etmektedir. Çalışmada kullanılan özellikler ve açıklamaları Çizelge 1'de verilmektedir.

Çizelge 1. i. Cümle özellikleri

Özellik	Açıklaması
f_{i1}	Uzunluk
f_{i2}	Pozisyon
f_{i3}	Başlık
f_{i4}	Ağırlık
f_{i5}	Çapraz
f_{i6}	Benzerlik

Sonraki aşamalarda özellikler analiz edilerek metindeki her cümlenin özet cümle olmaya ne kadar uygun olduğu hesaplanmaktadır. Cümlelerin özet cümle için puanları, Türkçenin dil özellikleri göz önüne alınarak aşağıdaki Eşitlik 8'de olduğu gibi hesaplanmaktadır.

$$F(x) = 0.150 \times f_{i1} + 0.135 \times f_{i2} + 0.140 \times f_{i3} + 0.160 \times f_{i4} + 0.090 \times f_{i5} + 0.070 \times f_{i6} \quad (8)$$

Burada, i cümlesinin toplam puanı, cümlenin tüm özelliklerinin belirli bir ağırlık değeri ile çarpılıp toplama eklenmesi sonucunda elde edilmiştir [2, 9]. Eşitlikteki katsayılar literatürde daha önceden benzer şekilde gerçekleştirilen bir çalışmasında '2' belirlenmiş kaysayılar olarak seçilmiştir. Katsayıların optimizasyonu için farklı yöntemler denenebilir yada uzman görüşüne bağlı olarak farklı değerler seçilebilir.

Çalışmamızda, tüm cümleler Eşitlik 7 ile puanlandırıldıktan sonra, cümle seçimi aşaması gerçekleştirilmektedir. Cümleler yüksek puan alandan az puan alana doğru sıralanmaktadır. Bu sıralı listenin % 40'ı oranında cümle seçilip özet metine eklenmiştir. Son olarak özet metin çıktısı sorguyu oluşturan kullanıcıya gösterilmektedir. Deneylerde kullanılan metinlere ait özet bilgiler Çizelge 2'de verilmektedir.

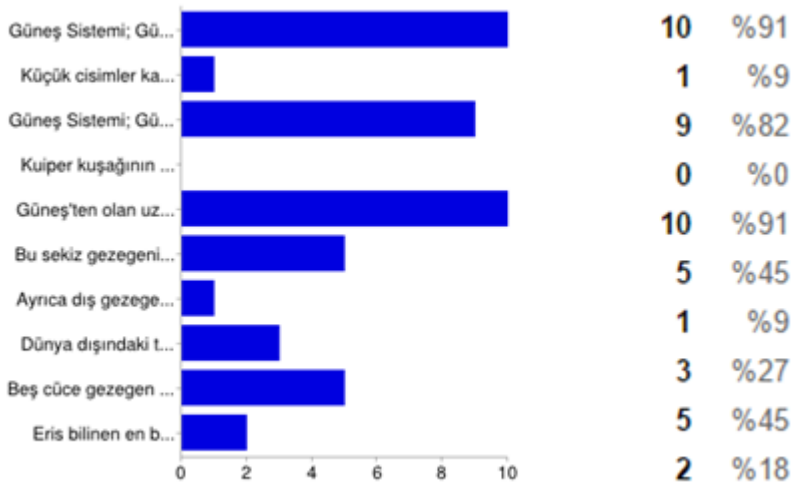
Çizelge 3'de "Güneş Sistemi" belgesine ait özetleme örneği gösterilmektedir.

Cümlelerin puanlarına göre sıralanmış özet cümleleri, cümlelerin orijinal metindeki sıralamasına göre otomatik olarak düzenlendiğinde elde edilen sonuç yukarıdaki çizelgede "Metindeki cümle sırasına göre sıralanmış özet cümleleri" başlığında belirtilmiştir. Metinin anlam bütünlüğü açısından daha doğru bir sıralama bu şekilde elde edilmiştir.

Çizelge 3'de verilen Güneş Sistemi metni proje hakkında bilgisi olmayan 10 farklı kullanıcıya verilmiş ve bu metinden 4 farklı cümle seçerek özetlemeleri istenmiştir. Bu çalışmanın önermiş olduğu cümleler ile kullanıcıların seçtikleri cümleler % 77.5 eşleşmiştir. Şekil 4'de kullanıcıların özet için seçtikleri cümleler ve bu cümlelerin kaç kişi tarafından seçildiğine dair yüzdelik dilim gösterilmiştir.

Çizelge 2. Deneylerde kullanılan metinler

Belge No	Başlık	Cümle Sayısı	Kaynakça
1	Güneş Sistemi	10	http://tr.wikipedia.org/wiki/Güneş_Sistemi
2	Charles Bukowski	6	http://tr.wikipedia.org/wiki/Charles_Bukowski
3	Güneş Ocağı	6	http://tr.wikipedia.org/wiki/Güneş_ocağı
4	Shine On You Crazy Diamond	5	http://tr.wikipedia.org/wiki/Shine_On_You_Crazy_Diamond
5	Fil	32	http://tr.wikipedia.org/wiki/Fil
6	Gandalf	22	http://tr.wikipedia.org/wiki/Gandalf
7	Yıldız Savaşları	13	http://tr.wikipedia.org/wiki/Yıldız_Savaşları
8	Java Programlama Dili	13	http://tr.wikipedia.org/wiki/Java_(programlama_dili)
9	Bilgisayar	13	http://tr.wikipedia.org/wiki/Bilgisayar
10	Afrika	10	http://tr.wikipedia.org/wiki/Afrika



Şekil 4. Metindeki her cümlenin kullanıcılar tarafından seçilme yüzdesi

Çizelge 4'de verilen Charles Bukowski metnini aynı şekilde proje hakkında bilgisi olmayan 10 farklı kullanıcıya verilmiş ve bu metinden 2 farklı cümle seçerek özetlemeleri istenmiştir. Bu çalışmanın önermiş olduğu sonuçlar ile kullanıcıların seçtikleri cümleler % 82

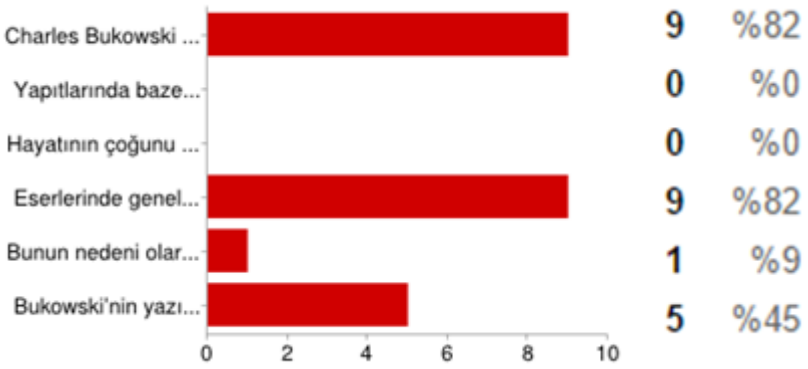
eşleşmiştir. Şekil 5'de kullanıcıların özet için seçtikleri cümleler ve bu cümlelerin kaç kişi tarafından seçildiğine dair yüzdelik dilim gösterilmiştir.

Çizelge 3. Örnek metin özetlemesi

Metin başlığı: Güneş Sistemi
Ana Metin
Güneş Sistemi; Güneş ve onun çekim etkisi altında kalan sekiz gezegen ile onların bilinen 166 uydusu, altı cüce gezegen (Ceres, Plüton, Eris, Haumea, Makemake,Sedna) ile onların bilinen altı uydusu ve milyarlarca küçük gök cisminden oluşur. Küçük cisimler kategorisine asteroitler, Kuiper kuşağı nesneleri, kuyrukluyıldızlar, gök taşları ve gezegenlerarası toz girer. Güneş Sistemi; Güneş, dört Yer benzeri iç gezegen, küçük, kaya ve metal içerikli asteroitlerden oluşan bir asteroit kuşağı, dört gaz devi dış gezegen, ve Kuiper kuşağı denen buzsu cisimlerden oluşan ikinci bir kuşaktan ibarettir. Kuiper kuşağının ötesinde ise seyrek disk, gündurgun (heliopause) ve en son olarak da varsayımsal Oort bulutu bulunur. Güneş'ten olan uzaklıklarına göre gezegenler sırasıyla Merkür, Venüs, Dünya, Mars, Jüpiter, Satürn, Uranüs ve Neptün'dür. Bu sekiz gezegenin altısının çevresinde doğal uydular döner. Ayrıca dış gezegenlerin her birinin toz ve diğer parçacıklardan oluşan halkaları vardır. Dünya dışındaki tüm gezegenler adlarını Yunan ve Roma mitolojisi'nin tanrılarından alır. Beş cüce gezegen ise; Kuiper kuşağında yer alan Plüton, Haumea ve Makemake; asteroit kuşağındaki en büyük cisim olan Ceres ve seyrek diskte yer alan Eris'tir. Eris bilinen en büyük cüce gezegendir.
Cümlelerin puanlarına göre sıralanmış özet cümleleri
Güneş'ten olan uzaklıklarına göre gezegenler sırasıyla Merkür, Venüs, Dünya, Mars, Jüpiter, Satürn, Uranüs ve Neptün'dür. Güneş Sistemi; Güneş ve onun çekim etkisi altında kalan sekiz gezegen ile onların bilinen 166 uydusu, beş cüce gezegen (Ceres, Plüton, Eris, Haumea, Makemake) ile onların bilinen altı uydusu ve milyarlarca küçük gök cisminden oluşur. Beş cüce gezegen ise; Kuiper kuşağında yer alan Plüton, Haumea ve Makemake; asteroit kuşağındaki en büyük cisim olan Ceres ve seyrek diskte yer alan Eris'tir. Güneş Sistemi; Güneş, dört Yer benzeri iç gezegen, küçük, kaya ve metal içerikli asteroitlerden oluşan bir asteroit kuşağı, dört gaz devi dış gezegen, ve Kuiper kuşağı denen buzsu cisimlerden oluşan ikinci bir kuşaktan ibarettir.
Metindeki cümle sırasına göre sıralanmış özet cümleleri
Güneş Sistemi; Güneş ve onun çekim etkisi altında kalan sekiz gezegen ile onların bilinen 166 uydusu, beş cüce gezegen (Ceres, Plüton, Eris, Haumea, Makemake) ile onların bilinen altı uydusu ve milyarlarca küçük gök cisminden oluşur. Güneş Sistemi; Güneş, dört Yer benzeri iç gezegen, küçük, kaya ve metal içerikli asteroitlerden oluşan bir asteroit kuşağı, dört gaz devi dış gezegen, ve Kuiper kuşağı denen buzsu cisimlerden oluşan ikinci bir kuşaktan ibarettir. Güneş'ten olan uzaklıklarına göre gezegenler sırasıyla Merkür, Venüs, Dünya, Mars, Jüpiter, Satürn, Uranüs ve Neptün'dür. Beş cüce gezegen ise; Kuiper kuşağında yer alan Plüton, Haumea ve Makemake; asteroit kuşağındaki en büyük cisim olan Ceres ve seyrek diskte yer alan Eris'tir.

Çizelge 4. Örnek metin özetlemesi

Metin başlığı: Charles Bukowski
Ana Metin
Charles Bukowski (16 Ağustos 1920 – 9 Mart 1994), asıl adı Heinrich Karl Bukowski olan Amerikalı yazar ve şair. Yapıtlarında bazen Henry Chinaski ismini de kullanmıştır. Hayatının çoğunu ABD'nin Los Angeles şehrinde geçirmiştir. Eserlerinde genellikle toplum dışı insanları ve depresyonu konu alması ve alkolizme yakın bir hayat tarzını anlatmasıyla ünlüdür. Bunun nedeni olarak kendisinin bu hayatı yaşaması gösterilebilir. Bukowski'nin yazılarında kendi hayatını yazıp yazmadığı tartışma konusu olmuştur; hayranlarının bir kısmı bunları kurguladığını, çoğunluğu ise yaşamadan bu tip kurguları yapmasının mümkün olmayacağını ve o karakterde bir insanın bu hayatı sürmesinin zaten doğal olduğu görüşünü savunmaktadır.
Cümlelerin puanlarına göre sıralanmış özet cümleleri
Eserlerinde genellikle toplum dışı insanları ve depresyonu konu alması ve alkolizme yakın bir hayat tarzını anlatmasıyla ünlüdür. Charles Bukowski (16 Ağustos 1920 – 9 Mart 1994), asıl adı Heinrich Karl Bukowski olan Amerikalı yazar ve şair.
Metindeki cümle sırasına göre sıralanmış özet cümleleri
Charles Bukowski (16 Ağustos 1920 – 9 Mart 1994), asıl adı Heinrich Karl Bukowski olan Amerikalı yazar ve şair. Eserlerinde genellikle toplum dışı insanları ve depresyonu konu alması ve alkolizme yakın bir hayat tarzını anlatmasıyla ünlüdür.



Şekil 5. Metindeki her cümlenin kullanıcılar tarafından seçilme yüzdesi

4. SONUÇLAR

Bu çalışmada bir doğal dil işleme uygulaması olan metin özetleme problemi üzerinde çalışılmıştır. Özet cümlelerin seçimi, özetlenecek metinlerin Türkçenin dil özelliklerine dayalı istatistiksel puanlanlandırılması ve anlamsal puanlandırılması yöntemlerinin melez şekilde değerlendirilmesi ile gerçekleştirilmiştir. Cümlelerin değerlendirilmesinde ele alınacak yapısal ve anlamsal özellikler, literatürdeki çalışmalara benzer şekilde gerçekleştirilmiş ve önerilmiştir. Özetlenecek metinlerde yer alan cümlelerin özet cümle adaylığı için aldıkları puanlar, yapısal ve anlamsal özelliklerin sezgisel bir ağırlıklandırma yöntemi ile birleştirilmesi ile belirlenmiştir. Çalışma kapsamında veri ön işleme, yapısal olarak istatistiksel puanlandırma, GAA analiz ve melez cümle seçimi aşamaları Türkçe yazılmış metinler üzerinde başarıyla gerçekleştirilmiştir.

Elde edilen sonuçların değerlendirilmesi için geliştirilen özetleme sistemine ve farklı kullanıcılara aynı metinler verilmiştir. Özetleme sisteminin önerdiği cümleler ile kullanıcıların önermiş olduğu cümleler karşılaştırılmıştır. Bu karşılaştırma sonucunda "Güneş Sistemi" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler % 77.5, "Charles Bukowski" metninin özeti ile kullanıcıların bu metinden seçtiği cümleler % 82 oranında eşleşmiştir. Burada gösterilemeyen benzer sonuçlar değerlendirildiğinde kısa metinlerde sistemin başarımı daha yüksek iken metinlerin uzunluğu arttıkça sistemin başarımı düşmektedir. Uzun metinlerde başarımı artırmak için metin özellikleri için önerilmiş ağırlıkların optimizasyonu gerçekleştirilebilir.

Çalışmanın bundan sonraki aşamalarında değerlendirilen belge sayısı artırılarak, özetleme için hesaba katılan metin özelikleri zenginleştirilmeye çalışılacaktır.

KAYNAKLAR

[1] Uzundere E, Dedja E, Diri B, Amasyalı MF. Türkçe Haber Metinleri İçin Otomatik Özetleme, *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*, 2008, s.1-3.

[2] Güran A, Güler BN, Gürbüz ZM. Efficient Feature Integration with Wikipedia-Based Semantic Feature Extraction for Turkish Text Summarization, *Turkish Journal of Electrical Engineering & Computer Sciences*, 2013, s.3-11.

[3] Çığır C, Kutlu M, Çicekli I. Generic Text Summarization for Turkish, *The Computer Journal*, 2010, s.1315-1323.

[4] Güran A, Bayazıt NG, Bekar B. Automatic Summarization of Turkish Documents Using Non-negative Matrix Factorization, *Innovations in Intelligent Systems and Applications (INISTA)*, 2011, s.480-484.

[5] Das D, Martins AFT, A Survey on Automatic Text Summarization, *Literature Survey for the Language and Statistics II course at CMU*, 2007.

[6] [http://en.wikipedia.org/wiki/Stop_words], Erişim tarihi: 20.01.2015.

[7] [http://www.konumuzseo.com/stop-words-ve-kullanim-amaclari/], Erişim tarihi: 20.01.2015.

[8] [http://tr.wikipedia.org/wiki/Zemberek_(yazılım)], Erişim tarihi: 20.01.2015.

[9] Özsoy GM. Text Summarization Using Latent Semantic Analysis, *Proceeding of the 23rd International Conference on Computational Linguistics*, 2011, s.22-41.

[10] Yohei S. Sentence Extraction by tfidf and Position Weighting from Newspaper Articles, *Third NTCIR Workshop*, 2003, s.2-6.

[11] Ünaldı İ, Kırkgöz Y. Latent Semantic Analysis: An Analytical Tool for Second Language Writing Assessment, *Mustafa Kemal University Journal of Social Sciences Institue*, 2011, s.2-9.

[12] Golub HG, Reinsch C. Singular Value Decomposition and Least Squares Solutions, *Numerische Matematik*, 1970, s.1-14.

[13] Steinberger J, Jezek K. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation, *ISIM2004*, s.2-21.

[14] Özsoy GM, Çiçekli İ, Alpaslan F. Text Summarization of Turkish Texts Using Latent Semantic Analysis, *Proceeding of the 23rd International Conference on Computational Linguistics*, 2010, s.3-6.

ÖZGEÇMİŞ / CV

Akif Hatipoğlu

Akif Hatipoğlu, 27 Mayıs 1993 yılında Balıkesir'in Bandırma ilçesinde doğdu. 18 yaşına kadar yaşamını Bandırma'da sürdürdü. Şu an ki yaşamını Kocaeli'nde sürdürmektedir. İlköğretim ve ortaöğretimini Bandırma Evyapan İlk Öğretim Okulu'nda okudu. Liseyi Bandırma Anadolu Lisesi'nde okudu. 2011 LYS sınavı ile şu an okumanda olduğu Kocaeli Üniversitesi Bilgisayar Mühendisliği bölümünü kazandı. Bu bölümde 4.sınıf öğrencisi olarak halen öğrenim hayatına devam etmektedir.

Akif Hatipoglu, was born on May 27,1993 in Bandırma/Balıkesir. He continued living up to 18 years in Bandırma. He continues his current life in Kocaeli. He studied at the elementary and secondary education Bandırma Evyapan Elementary School. He studied high school in Bandırma Anatolian High School. He won the Kocaeli University Computer Engineering Department in 2011. He still continues his education career as a 4th year student of Computer Engineering at the Kocaeli University.