Koşullu Rastgele Alanlar ile Türkçe Haber Metinlerinin Etiketlenmesi

(Labelling Turkish News Stories with Conditional Random Fields)

Seda KAZKILINÇ
İTÜ Bilgisayar Mühendisliği Bölümü
kazkilinc@itu.edu.tr

Eşref ADALI
İTÜ Bilgisayar Mühendisliği Bölümü
adali@itu.edu.tr

Özetçe

Her geçen gün belge sayısı artan Web'in tam potansiyeliyle kullanılması için anlamsal ağ alanındaki çalışmaların Web'in geleceğini oluşturacağı düşünülmektedir. Belge sayısındaki bu artışa bağlı olarak istenilen metne erişebilmek için bu metni en iyi temsil eden söz öbeklerinin bulunması doğru bir yaklaşım olmaktadır. Tüm metni okumadan o metni en iyi ifade edecek söz öbeklerine erişmek hem kullanıcı açısından hem de tarayıcı açısından büyük önem taşımaktadır. Bu çalışmanın amacı haber metinlerinde, haber metninin öznesi, yüklemi, yer ve zamanını belirtecek söz öbeklerinin metinde bulunup, metnin etiketlenmesidir. Haber metnin öznesi, metindeki en baskın kişi, şey veya süjeyi ifade eder. Metnin yüklemi ise metindeki oluşu ifade eder. Metnin yeri ve zamanı ise metindeki olayın geçtiği zaman ve yeri ifade eder. Bu amaçla, metinde geçen cümleler içerisinden seçilen en baskın özne, yüklem, yer ve zaman bilgilerinin çıkarılması hedeflenmektedir. Kapsam olarak Türkçe haber metinleri seçilmiştir. Elle etiketleme işlemi yapılan metinler otomatik etiketleme işlemi esnasında bir kısmı eğitim ve diğer kısmı ise sınama verisi olarak kullanılmıştır.

Anahtar Sözcükler *:* Doğal Dil İşleme, Bilgi Çıkarımı, Koşullu Rastgele Alanlar, Varlık İsmi Tanıma

Abstract

Drastical document increase in Web requires semantic web applications in order to lead the Web to its full potential. Extracting important phrases in a document facilitates finding expected information. In this paper, a new approach that is labelling the main subject, main predicate, main location and main date of an electronic document is introduced. The main subject label tells whom or what the document about. The main predicate label tells what the subject is or does. The main location label tells where the activities passed and the main date label tells when the document passed. With the help of this new methodology, extraction of not only high level description of the content, but also the attribute of a phrase in a document is provided. As an experimental set Turkish news stories are selected. To use as a training and test set, manual labeling is made by human annotators. Then, different models for each label are implemented to extract the labels automatically and they are compared to manually labelled results to evaluation process of this study.

Key words: Natural Language Processing, Information Extraction, Conditional Random Fields, Named Entity Recognition

1. Giriş

İnternetin yaygınlaşması ile birlikte artan belge sayısı, istenilen belgeye erişmedeki zorluğuda beraberinde getirmiştir. İstenilen belgeye daha kolay ulaşabilmek için belgeyi en iyi temsil eden söz öbeklerinin bulunması doğru bir yaklaşım olmaktadır. Metni okumadan metni en iyi tasvir eden söz öbeklerine ulaşmak kullanıcı ve tarayıcı açısından büyük önem taşımaktadır.

Çalışmanın amacı metni en iyi niteleyen özne, yüklem, yer ve zaman etiketlerinin metinden çıkarılmasıdır. Elde edilen bu etiketler sayesinde metnin konusu çıkarılabilir. Türkçe dili için geliştirilen bu çalışmanın çerçevesi 50-300 sözcük içeren haber metinleridir. Veri seti, Türkiye, Dünya, Siyasi, Ekonomi, Bilim ve Teknoloji ile Spor konularından 75'şer adet haberin internet üzerindeki haber kaynaklarından çekilmesiyle oluşturulmuştur [21][22][23][24].

Kavram olarak metnin öznesi,yüklemi, yer ve zamanından bahsedilemez. Çalışmada özne olarak kastedilen kavram metindeki ana karakter, yüklemin bildirdiği durumu üzerine alan şey veya kimseyi ifade eder. Metnin yüklemi ise oluş, iş ve hareket bildiren sözcük veya sözcük kümesini ifade eder. Metnin yeri ve zamanı ise metnin geçtiği yer ve zamanı ifade eder.

Ele aldığımız metinler Türkçe olduğu için anlamlı verilere erişebilmek için tüm sözcüklerin biçimbilimsel analizi yapılmıştır. Biçimbilimsel çözümleme için Oflazer'in çalışması kullanılmıştır [18]. Biçimbilimsel analizden çıkan sonuçlar birden fazla olduğun için en doğru biçimbirimi bulmak için biçimbilimsel belirsizlikleri giderilmiştir. Belirsizlik giderici olarak Sak'ın çalışması kullanılmamıştır [19].

Daha sonra sözcüğün cümle içindeki niteliğini belirlemek için sözdizimsel çözümleme yapılmıştır. Sözdizimsel çözümleyici olarak Eryiğit ve arkadaşlarının çalışması kullanılmıştır [20].

Çalışmamızda bir metin ilk olarak yukarıda sıralanan üç aşamalı çözümleme işleminden geçirilmiştir. Çalışmanın ilk kısmında biçimbilimsel ve sözdizimsel çözümü çıkarılmış olan metinlerden kurallar çıkarılarak etiketleme işlemi yapılmaya çalışılmışsa da yeterli başarımı elde edilememiştir. Bu nedenle, çıkaramadığımız bazı kuralları çıkarabileceğini düşünerek makine öğrenmesi yöntemleri üzerinde çalışılmıştır. Makine öğrenmesi yöntemi olarak bir dizilim sınıflandırıcısı olan Koşullu Rastgele Alanlar (KRA) üzerinde çalışılmıştır. Kural tabanlı yaklaşımda elde ettiğimiz bazı kuralları kullanarak ve çözümleyici çıktılarını kullanarak metindeki her bir sözcüğe ait nitelikler belirlenmiştir. Önceden elle işaretlediğimiz metinleri ve belirlenen nitelikleri kullanarak, KRA modelimiz oluşturulmuştur. Daha sonra önceden etiketlenmemiş metinleri, bu model sayesinde etiketleme işlemini geliştirilmiştir.

Bu çalışmanın bilimsel ve teknik katkısını ortaya çıkarabilmek için, sınama kümesindeki elle etiketlediğimiz metinlerin etiketlerini KRA'in ürettiği etiketler ile karşılaştırıp başarımızı tutturma ve bulma olasılıkları ve bunlardan türeyen F-ölçüm oranı cinsinden ölçülmüştür.

2. Benzer Çalışmalar

Bugüne kadar birebir metnin öznesi, yüklemi, yeri ve zamanı alanında bir çalışma olmasa da, çalışmaya referans olacak çalışmalar incelenmiştir.

Bunlardan ilk grup Varlık İsmi Tanıma çalışmalarıdır. Varlık İsmi Tanıma (VİT) bilgi çıkarımının bir alt dalı olup, metinlerde daha önceden çıkarılmış veya elde var olan bilgileri kullanarak kişi, kurum, kuruluş, yer isimleri, zaman ifadeleri, para birimleri gibi varlıkları tanıma işlemidir [1].

Kural tabanlı yaklaşımlara örnek olarak İngilizce dili için yapılmış olan Crystal[2] çalışması verilebilir. Bu çalışma dilden örüntüler çıkarılarak oluşturulmuş bir sözlük benzer sözcüklerin çıkarılması için kullanılabilir. Bu yöntem, bunun için kavramlar sözlüğünün otomatik olarak oluşturulmasını sağlamaya çalışır. Makine öğrenmesi yöntemleriyle eğitim kümesinin sistemi eğitmesiyle oluşturulur.

Diğer bir örnek olarak Nymble [3] ise varlık isimlerini metinlerden çıkarmak için Saklı Markov Modeli'ni kullarak eğitilmiş bir modeldir. Eğitim kümesinin istatistiksel yöntemlerde başarı oranını doğrudan etkilemesinden dolayı başarısı yüksek bir yöntemdir. İngilizce ve İspayolca için uygulanmıştır.

Diğer bir önemli çalışma ise NetOwl'dur [4]. İleri dil işleme yöntemlerini kullanarak anahtar kavramları çıkarıp sınıflandırmayı hedefler.

Küçük tarafından yapılan çalışma [5] kural tabanlı bir yaklaşımdır. Kişi isimleri, tanınmış kişiler, tanınmış organizasyon isimleri gibi sözlükleri bulmaktadır. Ayrıca Türkçe için belirli örüntüler çıkarılır. Bunlara bağlı olarak haber metinlerinde varlık isimlerini çıkarmaktadır ve %78 oranında bir başarı elde edilmiştir. Ancak masalları ve tarih konulu yazılarda başarı oranı yeterli düzeyde değildir.

Bayraktar ve arkadaşları tarafından yapılan "Finansal Haber Metinlerinde Kişi İsmi Etiketleme" isimli çalışma [6] ise yerel dilbilgisi yaklaşımı üzerine yoğunlaşmıştır. Yerel dilbilgisi yaklaşımı varlık tanıma esnasında diğer varlık tanıma sistemlerinin aksine hiç bir genel sözlük, isim, organizasyon ya da yer sözlüğüne ihtiyaç duymamaktadır. Sonuç olarak yerel dilbilgisi yaklaşımı daha önce görülmemiş metinlerde varlıkları tanımakta ve sınıflandırmaktadır. Diğer varlık tanıma sistemleri yerel dilbilgisi yaklaşımının aksine örüntü oluşturmadan önce bazı anlamsal ve yapısal analizlere ihtiyaç duymaktadır. Kişi isimlerini çıkarmada kullanılan bu yöntem ile yerel dilbilgisi yaklaşımının sıklık analizi, uygunluk analizi ve eşdizimlilik analizi yapılarak Türkçe'ye uygulanabilirliğini araştırılmıştır.

Varlık İsmi Tanıma çalışmalardan biri olan Cucerzan ve arkadaşlarının çalışması [1] kişi, yer, kuruluş ve diğer önemli isimleri metinden çıkarmayı hedefler. Dilden bağımsız geliştirilen bu çalışma tekrarlı eğitime dayanan ve biçimbilimsel örüntüleri kullanarak ve bağlama bağlı olarak hiyerarşik bir model oluşturur. Sadece dilden bağımsız olarak elle etiketlenmiş bir veri kümesini model oluşturmak için kullanır. Bu veriler sayesinde o dile bağlı örüntüler çıkarır. Bu yöntem önyükleme algoritması izlenerek oluşturulmuş bir yöntemdir. Bir çok dil için uygulanan bu yöntem Türkçe için de uygulanmıştır [1].

Sak ve arkadaşları tarafından yapılan "Türkçe için İstatistiksel Bilgi Çıkarım Sistemleri" isimli çalışamada [7] Saklı Markov Modeli içinde gömülü n-gram dil modelini kullanılmıştır. Sözlük modeli ve biçimbilimsel modelin birlikte uygulanması sonucu ortaya çıkan bu yeni model ile % 91.56 oranında başarı elde edilmiştir.

Diri ve Özkaya, "Türkçe Metinlerde Şartlı Rastgele Alanlarla Varlık İsmi Tanıma" [25] isimli çalışmada, bir sıralı makine öğrenmesi yöntemi olan Koşullu Rastgele Alanlar'ı kullanarak Türkçe metinlerde Varlık İsmi Tanıma üzerinde çalışmışlardır.

Nallapati ve arkadaşlarının yaptığı [8] haber metinlerinden anahtar sözcük çıkarımı çalışması anahtar kişiler, anahtar yerler, anahtar isimler ve anahtar eylemleri haber metinlerinden çıkarmayı hedefler. Buna bağlı olarak bu sorunu sınıflandırma problemi olarak görür. Öncelikli olarak anahtar sözcükleri çıkarılır ve anahtar sözcükleri Naive Bayes, Saklı Markov modeli ve Maksimum Entropi Model'i ile anahtar sözcükleri sınıflandırır. Arama motorlarınca dikkate alınmayan ve çok tekrarlanan ve sıralama hesaplarına dahil edilmeyen sözcüklerin ayıklanmasıyla elde edilen anahtar sözcüklerin Maksimum Entropi Model'i ile sınıflandırılması sonucu en iyi sonuçlar elde edilmiştir. Bizim çalışmamızın İngilizce dili için yapılmış bu çalışmadan farkı, bu işlemi Türkçe gibi eklemeli bir dil ile yapmasının yanında çıkartılan etiketlerin cümlelerin öğeleri gibi metnin öğelerini çıkaran bir yaklaşım izlemesi ve bu amaca yönelik bilgi çıkarma yöntemine gitmesidir.

Bu çalışmaya kaynak olabilecek diğer tür çalışmalar ise anahtar sözcük öbeği çıkarımını inceleyen çalışmalardır. Çoğunlukla en sık geçen sözcükleri bulma, TF*IDF ve sözcüğün ilk gözlemlendiği yeri dikkate alarak geliştirilirler. Cohen'in [9] ve Matsuo ile Ishizuka'nın çalışması [10] örnek verilebilir.

Plas ve arkadaşları [11] WORDNET kullanarak konuşma dilinde bulunan anahtar sözcük çıkarımını kullanmıştır. Hulth [12] ise çalışmasında sözdizimsel kurallara ek olarak isim öbekleri yığınları ve n-gram yöntemlerini uygulamış ve başarılı sonuçlar elde etmiştir.

Makine öğrenmesine dayalı anahtar sözcük öbeği çıkarımlarında en önemli çalışmalardan biri Anahtar Sözcük Öbeği Çıkarım Algoritması (Keyword Extraction Algorithm; KEA)'dır [13].

Anahtar sözcük öbeği çıkarım algoritmalarına bir örnek de yapay sinir ağları kullanılarak oluşturulmuş bir modeldir. Wang ve arkadaşları [14] bu yöntemde

TF*IDF özelliği bir ağırlık değeri olarak kullanmıştır. Bunun yanında ise başlık ve alt başlıklar, sözcük öbeğinin bulunduğu paragraf sayısı ağırlık değeri olarak kullanılmıştır. Bu ağırlık değerleri yardımıyla oluşturulan Yapay Sinir Ağı algoritması eğitim ve sınama süreçlerine sahiptir. Bu uygulamanın bulma ve tutturma yönteminin başarısı %30'dur. Kullanıcı bazlı değerlendirmede ise başarı %65'dir.

Bir başka sözcük öbeği çıkarım algoritması ise C4.5 ve GenEx algoritmalarıdır [15]. Her ikisi de güdümlü öğrenme algoritmalarıdır. Öncelikle tüm olası sözcük öbekleri metinden çıkarılır. Sözcük öbeğinin geçme sıklığı, metinde ilk kullanıldığı yer, özel isim olup olmadığı gibi nitelikler yardımıyla bir model oluşturulur.

Kalaycılar ve Çiçekli tarafından önerilen TurkeyX [16] ise anahtar sözcük öbeği çıkarımında kullanılan bir güdümsüz öğrenme modelidir. Bir metinde istatistiksel olarak isim öbeklerinin bulunma sıklığına bakar. KEA ve GenEx'in bazı özelliklerini kullanan bu yöntem ilk olarak tüm aday sözcük öbeği listesini çıkarır. Bu kısımda biçimbilimsel analizi yapılmış sözcükler kullanılır. Daha sonra başka bir sözcük öbeğinin içinde geçen öbeklerden az sözcüklü olanı elenir. Daha sonra en çok geçen sözcük öbekleri anahtar sözcük öbeği olarak adlandırılır.

2.1. Koşullu Rastgele Alanlar

KRA, Lafferty ve arkadaşları [17] tarafından önerilen istatistiksel dizilim sınıflandırmasına dayanan bir makine öğrenmesi yöntemidir. Dizilim sınıflandırıcıları bir dizilim içerisindeki her birime bir etiket atamaya çalışırlar. Olası etiketler üzerinde bir olasılık dağılımı hesaplar ve en olası etiket dizilimini seçerler. Buna göre KRA modeli $p(y^*|x^*)$ olasılığını hesaplamak üzere geliştirilmiş bir olasılık modeli olarak tanımlanabilir. Burada $y^* = y_1,...,y_n$ olası çıktı etiketlerini belirtirken, $x^* = x_1,...,x_n$ giriş verilerini belirtir. Buna göre KRA modeli bağıntı 1 ile gösterilebilir.

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, x_t)\right\} \quad (1)$$

Burada Z_φ tüm olası etiket dizileri için normalleştirme faktörüdür ve bağıntı 2'deki gibi tanımlanır:

$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y}\in Y^T} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, x_t)\right\}. \quad (2)$$

Burada, bağıntı 2'de de görüleceği üzere nitelik fonksiyonu parametreleri t. etiket y_t ve t-1. etiket y_{t-1} ve sözcük dizilimi x olan bir fonksiyonudur. Nitelik fonksiyonları makine öğrenmesinde kullanmak istenilen nitelikleri belirleyen fonksiyonlardır.

3. Metinlerin Önişlenmesi

Çalışmamızda amacımız haber metinlerinden özne, yüklem, yer ve zaman etiketlerinin çıkarılmasıdır. Kapsam olarak Türkçe dilinde yazılmış metinler incelendiği için Türkçe diline özgü bir çözüm geliştirilmektedir. Türkçe dilinin eklemeli bir dil olması nedeniyle sözcük kökleri çok sayıda ek alabilmektedir. Bu nedenle ilk olarak sözcüklerin gövdelerine erişmek gerekir. Biçimbilimsel çözümleyici işlemi sonucu sözcüklerin ek ve kök bilgileri yanında sözcüklerin türlerine ait bilgilere de erişilmektedir. Bu işlem için Oflazer'in biçimbilimsel çözümleyicisi kullanılmıştır. [18] Biçimbilimsel çözümleyiciden elde edilen çıktı Tablo 1'de verilmiştir.

Çizelge-1: Biçimbilimsel çözümleyiciye bir örnek

Çözümleyici Girdisi	Çözümleyici Çıktısı
Her	Her Her Noun+Prop+A3sg+Pnon+Nom
şey	şey şey+Noun+A3sg+Pnon+Nom
çok	çok çok+Adverb
	çok çok+Det
	çok çok+Adj
	çok çok+Postp+PCAbl
güzel	güzel güzel+Adj
olacak	olacak ol +Verb+Pos+Fut+A3sg
	olacak ol +Verb+PosDB+Adj+FutPart+Pnon
.	. . +Punc

Türkçe'de sözcüklerin ortalama iki adet biçimbilimsel çözümü bulunmaktadır. Bu nedenle en doğru çözümü bulabilmek için belirsizlik giderme işlemi yapılmıştır. Belirsizlik giderici olarak Sak tarafından hazırlanan belirsizlik giderici çıkarılmıştır [19]. Belirsizlik giderici'Nin girdi ve çıktılarına ait bir örnek Tablo 2'de verilmiştir.

Çizelge-2: Belirsizlik gidericiye bir örnek

Belirsizlik Giderici Girdisi	Belirsizlik Giderici Çıktısı
Her	Her Her Noun+Prop+A3sg+Pnon+Nom
şey	şey şey+Noun+A3sg+Pnon+Nom
çok	çok çok+Adverb
güzel	güzel güzel+Adj
olacak	olacak ol +Verb+Pos+Fut+A3sg
.	. . +Punc

Son olarak sözdizimsel çözümleme yapılarak sözcüklerin cümle içerisindeki görevlerine erişilebilmiştir. Eryiğit'in çalışmasından faydalanılmıştır [20]. Sözdizimsel çözümleyiciye bir örnek Tablo 3'de verilmiştir.

Çizelge-3: Sözdizimsel çözümleyiciye bir örnek

Etiketleyici Girdisi
Her Noun Prop Prop A3sg|Pnon|Nom _ _ _ _ şey şey Noun Noun A3sg|Pnon|Nom _ _ _ _ çok çok Adv Adv _ _ _ _ _ güzel güzel Adj Adj _ _ _ _ _ olacak ol Verb Verb Pos|Fut|A3sg _ _ _ _ . . Punc Punc _ _ _ _ _

Etiketleyici Çıktısı
Her Noun Prop Prop A3sg|Pnon|Nom 5 SUBJECT _ _ şey şey Noun Noun A3sg|Pnon|Nom 5 SUBJECT _ _ çok çok Adv Adv _ 4 MODIFIER _ _ güzel güzel Adj Adj _ 5 MODIFIER _ _ olacak ol Verb Verb Pos|Fut|A3sg 6 SENTENCE _ _ . . Punc Punc _ 0 ROOT _ _

4. Metinlerin İşlenmesi

Biçimbilimsel çözümleyici, belirsizlik giderici ve sözdizimsel çözümleyici çıktılarından faydalanılarak ilk olarak kural tabanlı yaklaşımlar ile hedefimize ulaşmak istenmiştir. Ancak başarı oranı düşük olasından dolayı makine öğrenmesi yöntemlerinden faydalanılmıştır. Türkçe gibi kuralların bol olduğu bir dilde hedefe ulaşmak için bir çok kuralın tanımlanması gerekmektedir. Bu nedenle probleme makine öğrenmesi yöntemi ile yaklaşmak doğru bir yaklaşım olmaktadır.

Çalışmamızda amaç bir dökümana ait özne, yüklem, yer ve zaman etiketlerini bulmaktır. Bu bakış açısı ile problem dizilimlerden oluşan, bir sınıflandırma problemidir. Her bir sözcük ya belgenin öznesidir, ya yüklemidir, ya yeridir, ya zamanıdır yada bunlardan hiçbiridir. Tablo 4'de etiketler ve anlamları gösterilmiştir.

Çizelge-4: Etiketler ve Anlamları

Etiket	Anlamı	Örnek
SUBJ	Metnin Öznesi	Mustafa Kemal
PRED	Metnin Yüklemi	Çıktı
LOC	Metnin Yeri	Samsun
DATE	Metnin Zamanı	19 Mayıs 1919
O	Yukarıdakilerden hiç biri	-

Sorunumuzu dizilim sınıflandırma problemi olarak ele aldığımızdan ve sistemi eğitmedeki veriminden dolayı Koşullu Rastgele Alanlar yöntemi ile sistemi modelleme tercih edilmiştir. Bu çalışmada, doğrusal zincir Koşullu Rastgele Alanlar kullanılmıştır.

4.1. Niteliklerin Belirlenmesi

Koşullu Rastgele Alanlarda sistemi eğitmek için nitelikleri belirlemek gerekir. Kural tabanlı çalışmalarımızda elde ettiğimiz deneyimlerimizin sonucu olarak, KRA için gerekli olan nitelikler belirlenmiştir.

4.1.1. Kural Tabanlı Nitelikler

Bu kısımda bahsedilen nitelikler, sorunu kural tabanlı olarak çözmek için yapılan araştırmalardan sonra çıkarılmıştır. Sadece kural tabanlı olarak sorunu çözmeye çalışarak, istenilen başarım oranına erişilememiştir. Fakat bu kurallardan bazılarının başarım oranını olumlu etkilelediği gözlemlenmiş ve KRA için nitelik olarak kullanılmıştır.

4.1.1.2. Özne ve Yer Etiketleri için Kural Tabanlı Nitelikler

Tanım olarak bir metnin öznesinden bahsedilemez. Bizim çalışmamızda bir metnin bütününün öznesi olarak kabul edilebilecek sözcüğe metnin öznesi adını vermekteyiz. Çalışmamızda kullanılan metin kümesi, haber nitelikli olduğu için, özne etiketleri ile yer etiketleri genelde özel isimlerden oluşmaktadır. Daha önce değinildiği gibi özne tek bir sözcük olabileceği gibi sözcük öbeği de olabilir.

Aşağıda ayrıntıları anlatılan kurallar ile sözcüğün bir özel isim kümesine dahil olup olmadığı olasılığı bir nitelik olarak alınmıştır. Buna göre eğer bir sözcük özel isim kümesinin bir üyesi ise ve bu sözcük

cümlede özne olarak kullanılmış ise 'POSSUB' isimli bir nitelikle tanımlanır.

Eğer sözcük özel isim kümesinin bir üyesi ise ve bu sözcük cümlede dolaylı tümleç olarak kullanılmış ise bu sözcük 'POSLOC' olacak şekilde nitelik olarak tanımlanır.

Özel isim öbekleri

Türkçe'de yazım kuralları gereği özel isimler büyük harfle başlar. Büyük harfle başlayan her sözcük özel isim değildir. Buna en iyi örnek cümle başındaki sözcüktür.

Buna göre aşağıdaki kurallar çıkarılabilir.
Kural 1: Eğer bir sözcüğün ilk harfi büyük harf ise ve cümle başında değil ise özel isimdir.

Kural 2: Eğer bir sözcük cümlenin ilk sözcüğü ise ve özel isim türü ile etiketlenmiş ise (Prop etiketli) özel isimdir.

Kural 3: Eğer iki özel isim arasında bağlaç var ise bu bağlaç özel isim öbeğine aittir. Örnek olarak, "Çalışma ve Sosyal Güvenlik Bakanlığı", "Gençlik ve Spor Bayramı" verilebilir.

Özel isim öbekleri için sınır kuralları

Özel isim grupları çıkarılırken kullanılan sınır kuralları aşağıda detaylı anlatılmıştır.

Sınır Kuralı 1: Eğer özel bir isim –i halini almış bir sözcük ise bu sözcük özel isim öbeğinin son sözcüğüdür. Sadece son sözcüğü –i hali almış özel isim öbekleri içerisinde bağlaç barındırabilir. Örnek olarak, "Savunma ve Sosyal Güvenlik Bakanlığı", "Öğrenci Seçme ve Yerleştirme Sınavı" verilebilir.

Sınır Kuralı 2: Herhangi bir noktalama işareti o sözcük öbeğinin bittiğine işaret eder. Bu sayede virgül "," ve kesme "'" işareti sadece sözcük öbeğindeki son sözcükte bulunabilir.

4.1.3. Biçimbilimsel nitelikler

Biçimbilimsel nitelikler biçimbilimsel çözümleyici ve sonrasında belirsizliği giderilmiş çözümler içerisinden seçilir. Temel olarak sözcük sınıfları ve ek sınıflarından oluşan bu nitelikler her bir sözcüğün biçimbilimsel belirsizlik gidericiden çıkmış hali işlenerek çıkartılır.

4.1.4. Sözdizimsel nitelikler

Sözdizimsel çözümleyiciden çıkan çözümlerden o sözcüğe ait sözdizimsel niteliği çıkarılır.

4.1.5. Yapısal nitelikler

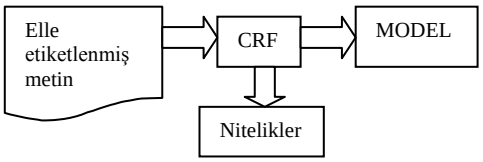
Yapısal niteliklere veri seti içerisinde metnin sırası, metin içerisinde cümle sırası, sıklık ve ilk gözlemlendiği yer olarak ve büyük harf ile başlama verilebilir.

Nitelik Seçimi ve Performans İlişkisi

Nitelikleri seçerken Koşullu Rastgele Alanlar yönteminin bağıntısından da anlaşılacağı üzere nitelik fonksiyonlarının ağırlıkları normalize edilmiş şekildedir. Uygulamanın gerçeklenmesi kısmında da bahsedileceği üzere fazla sayıda nitelik seçmenin sistemin performansına etkisi ihmal edilebilir düzeydedir. Bu nedenle Maksimum İlinti Minimum Fazlalık (MrMr) gibi nitelik seçme işlemine veya en çok etki eden nitelikleri bulma gibi bir yönteme başvurulmamıştır.

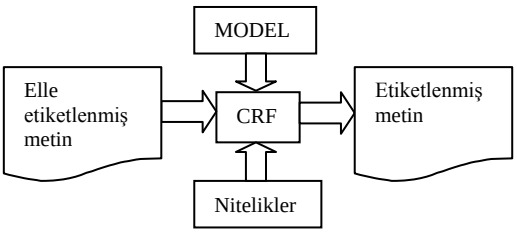
4.2. Koşullu Rastgele Alanlar'ın Kullanımı

Elle etiketlenmiş sözcükler nitelikleri ile birlikte kaydedilerek, KRA modellenmesinde eğitim kümesi olarak kullanılmıştır. Böylece eğittiğimiz KRA yapısına hiç etiketlenmemiş belgeler girdi olarak verilmiş ve etiketlenmeleri sağlanmıştır. Şekil 1 ve Şekil 2'de tarafımızca tasarlanan eğitim ve sınama aşamaları gösterilmiştir.



Şekil-1: KRA ile eğitim işlemi

Son olarak öğrenme kümesi kullanılarak eğittiğimiz KRA yapısına elle işaretlenmiş belgeler, giriş olarak uygulanmış geliştirdiğimiz yöntemin bulma ve tuturma olasılıklarıyla birlikte başarımı ölçülmüştür.



Şekil-2: KRA ile sınama işlemi

4.3. Başarımın Ölçülmesi

Başarımı ölçerken bulma(precision) ve tuturma(recall) oranları ile tutturma ve bulmanın harmonik ortalamasından hesaplanan f-ölçümü hrsaplanmıştır.

Bulma oranı sınama kümesi içende var olan etiketlerden bulabildiklerimizdir ve bağıntı 3'de bulunabilir.

$$Bulma = \frac{elle\ atanan\ etiketler\ içerisinde\ programın\ bulduğu\ etiketler}{sınama\ kümesindeki\ etiketler} \quad (3)$$

Tutturma oranı ise bağıntı 4'de görüleceği üzere bulabildiğimiz etiketlerin elle etiketlenmiş etiket listesinde olmasının oranıdır.

$$Tutturma = \frac{elle\ atanan\ etiketler\ içerisinde\ programın\ bulduğu\ etiketler}{programın\ bulduğu\ etiketler} \quad (4)$$

Sonuçların karşılaştırılması için kullanılan F-değeri bağıntı 5'de gösterilmiştir.

$$F = \frac{2 * tutturma * bulma}{tutturma + bulma} \quad (5)$$

Bu bilgiler ışığında çıkarılan sonuçlar Tablo 5'de gösterilmiştir.

Çizelge-5: Her bir etiketin başarı oranı

Etiket Tipi	Elle Etiketlenen	Programın Bulduğu	Çakışan
Özne Etiketi	50	62	40
Yüklem Etiketi	50	64	39
Yer Etiketi	12	11	8
Zaman Etiketi	11	7	5
Etiket Tipi	**Bulma**	**Tutturma**	**F-Ölçümü**
Özne Etiketi	0,8	0,645	0,72
Yüklem Etiketi	0,78	0,61	0,684
Yer Etiketi	0,667	0,727	0,695
Zaman Etiketi	0,45	0,71	0,55

5. Sonuç ve Öneriler

Bu çalışmada metni en iyi şekilde temsil eden özne, yüklem, yer ve zaman etiketlerinin metinden çıkarımı hedeflenmiştir. Bu hedef doğrultusunda öncelikle kural tabanlı yöntemler üzerinde çalışılsa da, başarım oranındaki düşüklükten dolayı, daha sonra makine öğrenmesi yöntemlerinden KRA üzerinde çalışılmıştır. Kural tabanlı çalışma sırasında gözlemlediğimiz bazı kuralları nitelik olarak kullanarak KRA yönteminin başarımı arttırılmıştır.

Bu çalışmada sadece haber metinleri üzerinde çalışıldığı için, haber metinlerine özel bazı nitelikler kullanılmıştır. Bu niteliklere haber metnin öznesi ve yer'i ifadesinin sıklıkla özel isim grubu olması örnek olarak verilebilir. İlerdeki çalışmalarda başka tür metinleri de incelemek hedeflenmiştir. Ancak bu çalışmada incelenen haber kümesi bir çok avantajın yanında dezavantaj da barındırmaktadır. Bir dezavantaj olarak, haber metinleri imla ve noktalama kurallarına uymayan bir çok hatayı da içinde barındırır. İnternet ortamında ne yazık ki haber metinleri editörler tarafından yeterince incelenmeden yayımlanmaktadır. İmla ve noktalama kuralları özellikle sözdizimsel çözümleme olmak üzere çalışmanın bütününde olumsuz etki yaratmaktadır. Başarı oranlarındaki düşüklüğün en büyük nedeni olarak bu hataları varsayabiliriz. Bu nedenle ilerideki çalışmalarımızda yazım ve noktalama hatalarını düzelten ek bir modülün geliştirilmesi hedeflenmektedir. Bu modülün

sistemin başarı oranını büyük düzeyde etkileyeceği düşünülmektedir. Bununla birlikte biçimbilimsel çözümleme, belirsizlik giderici ve sözdizimsel çözümleme işlemlerini kapsayan modülün hata oranları da sistemin başarı oranını doğrudan etkiler. Bu modüllere yönelik geliştirmelerde sistemin başarısını olumlu yönde etkilemektedir.

Başarım oranını etkileyen önemli bir nedende eğitim kümemizin niceliğinin sınırlı olmasıdır. Eğitim kümesi arttırılarak başarım oranında artış olacağı öngörülmektedir.

İlerdeki çalışmalarımızda başarım oranımızı arttırmayı ve elde edilen çözümü, anlamsal web uygulaması olarak kullanmayı hedeflemekteyiz.

Kaynakça

[1] **Silviu Cucerzan and David Yarowsky**, Language Independent Named Entity Recognition, Combining Morphological and Contextual Evidence. s. 90-99, 1999

[2] **Soderland, S., Fisher, D., Aseltine, J. ve Lehnert,W.**, 1995, CRYSTAL: Inducing a Conceptual Dictionary,

[3] **Bikel, D.M., Miller, S., Schwartz, R. ve Weischedel R.**, 1997. Nymble: a high-performance learning name-finder, Proceedings of the fifth conference on Applied natural language processing, ANLC '97, Association for Computational Linguistics, Stroudsburg, PA, USA, s.194–201

[4] NetOwl Server, Proceedings of the fifth conference on Applied natural language processing, ANLC '97, Association for Computational Linguistics, Stroudsburg, PA, USA, s.15–16

[5] **Kucuk, D. ve Yazici, A.**, 2009. Named Entity Recognition Experiments on Turkish Texts, Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS '09, Springer-Verlag, Berlin, Heidelberg,

[6] **Özkan Bayraktar**, 1991. Local Grammar, Person Name Recognition in Turkish Financial Texts by Using Local Grammar Approach, METU, s.19–27.

[7] **Tür, G., Hakkani-tür, D. ve Oflazer, K.**, 2003. A statistical information extraction system for Turkish, Nat. Lang. Eng., 9(2), 181–210,

[8] **Nallapati, R., Allan, J. ve Mahadevan, S.,** Extraction of Key Words from News Stories.

[9] **Cohen, J.D.,** 1995. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting, Journal of the American Society for Information Science, 46(3), 162–174.

[10] **Matsuo, Y. ve Ishizuka, M.**, 2004. Keyword extraction from a single document using word co-occurrence statistical information, International Journal on Artificial Intelligence Tools, 13(1), 157–169

[11] **van der Plas, L., Pallotta, V., Rajman, M. ve Ghorbel, H.,** 2004. Automatic Keyword Extraction from Spoken Text. A Comparison of two Lexical Resources: the EDR and WordNet, CoRR, cs.CL/0410062

[12] **Hulth, A.**, 2003. Improved automatic keyword extraction given more linguistic knowledge, Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA, s.216–223

[13] **Pala, N. ve Çiçekli, I.,** 2007. Turkish Keyphrase Extraction Using KEA, in, Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007.

[14] **Wang, J., Peng, H. ve Hu, J.s.**, 2006. Automatic keyphrases extraction from document using neural network, Proceedings of the 4th international conference on Advances in Machine Learning and Cybernetics, ICMLC'05, Springer-Verlag, Berlin, Heidelberg, s.633–641

[15] Quinlan, J.R., 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.

[16] Cicekli, I. ve Kalaycilar, F., 2008. TurKeyX: Turkish Keyphrase Extractor, Proceedings of the 23rd International Symposium on Computer and Information Sciences, TeX Users Group, s.84–89.

[17] Lafferty, J., McCallum, A. ve Pereira, F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, s.282–289.

[18] Oflazer, K., 1993. Two-level description of Turkish morphology, Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, EACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA, s.472–472,

[19] Sak, H., Güngör, T. ve Saraçlar, M., 2008. Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus, Proceedings of the 6th international conference on Advances in Natural Language Processing, GoTAL '08, Springer-Verlag, Berlin, Heidelberg, s.417–427,

[20] Eryiğit, G., 2007. ITU Treebank Annotation Tool, Proceedings of the ACL workshop on Linguistic Annotation (LAW 2007), Prague.

[21] http://www.ntvmsnbc.com/

[22] http://www.hurriyet.com.tr/anasayfa/

[23] http://www.milliyet.com.tr/Haber/

[24] http://www.zaman.com.tr/

[25] Ozkaya, S.; Diri, B., "Named Entity Recognition by Conditional Random Fields from Turkish informal texts," Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference on , vol., no., pp.662-665, 20-22 April 2011