

**T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**TÜRKÇE METİNLERDE ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ
TANIMA**

SERAP ÖZKAYA

**YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI**

**DANIŞMAN
DOÇ. DR. BANU DİRİ**

İSTANBUL, 2013

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERDE ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ
TANIMA

Serap ÖZKAYA tarafından hazırlanan tez çalışması 18.02.2013 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Banu DİRİ
Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Banu DİRİ
Yıldız Teknik Üniversitesi

Yrd. Doç. Dr. M. Fatih AMASYALI
Yıldız Teknik Üniversitesi

Yrd. Doç. Dr. A. Cüneyd TANTUĞ
İstanbul Teknik Üniversitesi

ÖNSÖZ

Varlık İsmi Tanıma, Doğal Dil İşleme'nin çalışma alanlarından biri olup kişi, yer, kurum, zaman, tarih, formül ve para gibi farklı varlık isimlerinin bulunup, anlamsal açıdan sınıflandırılmasıdır. Varlık İsmi Tanıma'nın iki temel amacı vardır. Birincisi varlıkların bulunup tanınması, ikincisi ise bu varlıkların sınıflandırılmasıdır.

Varlık İsmi Tanıma'da kural tabanlı veya makine öğrenmesi veya bunların birlikte kullanıldığı sistemler ile tanıma yapılabilmektedir.

Bu çalışmada, konudan bağımsız resmi olmayan dokümanlardan Şartlı Rastgele Alanlar kullanılarak kişi, kurum ve yer isimlerinin etiketlenmesi gerçekleştirilmiştir.

Bu çalışmanın başından sonuna kadar yapmış olduğu öneriler, karşılıklı tartışmalar ve yönlendirmeleri ile tezde yapmış olduğu detaylı incelemeler ve katkıları için hocam Doç. Dr. Banu Diri'ye çok teşekkür ederim.

Hayatım boyunca desteklerini hiçbir zaman eksik etmeyen, beni yüreklendiren ve bana güvenen sevgili aileme çok teşekkür ederim.

Ocak, 2013

Serap ÖZKAYA

İÇİNDEKİLER

	Sayfa
SİMGE LİSTESİ.....	vii
KISALTMA LİSTESİ	vii
ŞEKİL LİSTESİ.....	ix
ÇİZELGE LİSTESİ	x
ÖZET.....	xi
ABSTRACT.....	xiii
BÖLÜM 1	
GİRİŞ	1
1.1 Literatür Özeti	1
1.2 Tezin Amacı.....	3
1.3 Hipotez.....	4
BÖLÜM 2	
ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ TANIMA.....	6
2.1 Sıralı Verinin Etiketlenmesi.....	6
2.2 Yönsüz Grafik Modelleri	8
2.2.1 Potansiyel Fonksiyonlar.....	9
2.3 Şartlı Rastgele Alanlar	10
2.4 Maksimum Entropi	11
BÖLÜM 3	
SİSTEM TASARIMI	12
3.1 Kullanılan Yöntem	13
3.2 Türkçe E-Posta için Yapılan Çalışma	13
3.2.1 E-Posta Veri Seti	14
3.2.2 Temel Özellikler.....	16

3.2.2.1	Başlık Bilgisi Özellikleri.....	16
3.2.2.2	Etiketlenebilen Özellikler.....	16
3.2.2.3	Kural Tabanlı Özellikler	17
3.3	Doküman Türünden Bağımsız Şartlı Rastgele Alanlar Kullanımı	17
3.3.1	Eğitim Veri Seti.....	18
3.3.2	Olasılık Hesabı	19
3.3.3	Özellik Çıkarımı.....	20
3.3.4	Temel Özellikler.....	21
3.3.4.1	Büyük Harf Kullanımı	21
3.3.4.2	İlk Kelime Olması.....	21
3.3.4.3	Noktalama İşaretleri.....	21
3.3.4.4	Nümerik Değer İçerme	21
3.3.4.5	Özel Kelimeler Kullanma.....	22
3.4	Kural Tabanlı Yaklaşımın Eklenmesi.....	22
3.4.1	Kişi İsimleri	23
3.4.2	Kurum İsimleri	24
3.3.3	Yer İsimleri	25
BÖLÜM 4		
GELİŞTİRİLEN SİSTEM		26
4.1	Veri Seti Kullanımı	26
4.2	Türkçe E-Posta için Geliştirilen Sistem	28
4.2.1	Kelime Nesnesi Oluşturma	28
4.2.2	N-Gram Çıkarımı	29
4.2.3	E-Posta Özellikler Çıkartılması	29
4.2.4	Etiketlenebilen Özelliklerin Çıkartılması.....	29
4.2.5	Sınıflandırma	30
4.2.6	Uygulamanın Çalıştırılması	31
4.3	Yapılandırılmamış Dokümanlar için Geliştirilen Sistem.....	32
4.3.1	Sistemin Eğitilmesi	33
4.3.2	Özel Kelime Kullanılması	33
4.3.3	Özellik Çıkarımı.....	33
4.3.4	Şartlı Rastgele Alanların Kullanımı	34
4.3.5	Kural Tabanlı Yaklaşım.....	34
4.3.6	Uygulamanın Çalıştırılması.....	35
BÖLÜM 5		
DENEYSEL SONUÇLARIN DEĞERLENDİRİLMESİ		38
5.1	Test Veri Seti	39
5.2	Yarı-yapılandırılmış Dokümanlar için DeneySEL Sonuçlar	40
5.3	Yapılandırılmamış Dokümanlar için DeneySEL Sonuçlar	42
5.3.1	Eğitim Veri Setinin Etkisi	44
5.3.2	Test Veri Setinin Etkisi	46
5.3.3	DeneySEL Sonuçların Karşılaştırılması	48

BÖLÜM 6

SONUÇ VE ÖNERİLER49

KAYNAKLAR51

EK-A

ÖZEL KELİMELER.....54

A-1 Özel Kelime Listesi54

ÖZGEÇMİŞ60

SİMGE LİSTESİ

p	Dağılımın beklentisi
s_k	Durum özellik işlevi
t_j	Geçiş özellik işlevi
x	Gözlem dizisi
y	Etiket dizisi
$Z(x)$	Normalleştirme faktörü
λ_i	Çalışma verilerinden tahmin edilebilecek parametre

KISALTMA LİSTESİ

Bcc	Blind Carbon Copy
Cc	Carbon Copy
DT	Doğru Tespiti
GMM	Gizli Markov Model
GT	Gerçek Tespit
MEMM	Maksimum Entropi Markov Model
MUC	Message Understanding Conference
MUC-6	Sixth Message Understanding Conference
MUC-7	Seventh Message Understanding Conference
NER	Named Entity Recognition
P	Precision-Tutturma
R	Recall-Bulma
TDK	Türk Dil Kurumu
TOG	Toplum Gönüllüleri Vakfı
UT	Uygulama Tespiti

ŞEKİL LİSTESİ

	Sayfa
Şekil 2. 1 Diziler için zincir yapılı Şartlı Rastgele Alanların grafik yapısı	9
Şekil 3. 1 Örnek Kural Listesi.....	14
Şekil 3. 2 Örnek E-Posta	15
Şekil 3. 3 MUC standartları	18
Şekil 3. 4 Örnek MUC standardı.....	19
Şekil 4. 1 Etiketlenmiş eğitim veri set	28
Şekil 4. 2 Özellik çıkarımı	31
Şekil 4. 3 Uygulama akışı	32
Şekil 4. 4 Uygulamanın çalıştırılması.....	33
Şekil 4. 5 Yapılandırılmamış dokümanlar için Varlık İsmi Tanıma	37
Şekil 4. 6 Uygulamanın çalıştırılması.....	38
Şekil 5. 1 Varlık isimlerinin farklı alanlardaki F-Ölçüm değerleri	43
Şekil 5. 2 Varlık isimlerinin F-Ölçüm değerleri.....	45
Şekil 5. 3 Eğitim veri seti kullanımı ve F-ölçüm değerleri.....	46

ÇİZELGE LİSTESİ

	Sayfa
Çizelge 4. 1 E-posta veri seti özellikleri	28
Çizelge 4. 2 Eğitim veri seti özellikleri	29
Çizelge 5. 1 Test veri seti özellikleri	40
Çizelge 5. 2 Tespit edilen varlık sayıları	41
Çizelge 5. 3 Akademik E-Posta F-Ölçüm Değerleri	42
Çizelge 5. 4 Kurumsal E-Posta F-Ölçüm Değerleri	42
Çizelge 5. 5 Kişisel E-Posta F-Ölçüm Değerleri	42
Çizelge 5. 6 Yapılandırılmamış dokümanlarda tespit edilen varlık sayıları	44
Çizelge 5. 7 Şartlı Rastgele Alanlar kullanıldığında F-ölçüm değerleri	44
Çizelge 5. 8 Kural tabanlı yaklaşım eklenildiğinde F-ölçüm değerleri	44
Çizelge 5. 9 Eğitim veri setinin dörtte biri ile sistem eğitildiğinde	45
Çizelge 5. 10 Eğitim veri setinin yarısı ile sistem eğitildiğinde	45
Çizelge 5. 11 Eğitim veri setinin dörtte üçü ile sistem eğitildiğinde	45
Çizelge 5. 12 Test veri seti miktarı artırıldığında tespit edilen varlık sayıları	47
Çizelge 5. 13 Test veri seti miktarı artırıldığında F-Ölçüm değerleri	47
Çizelge 5. 14 İki yöntem birlikte kullanıldığında F-Ölçüm değerleri	47

TÜRKÇE METİNLERDE ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ TANIMA

Serap ÖZKAYA

Bilgisayar Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Tez Danışmanı: Doç.Dr. Banu DİRİ

Varlık İsmi Tanıma, Doğal Dil İşleme’de bilgi çıkarımının bir alt dalı olup, kişi ve kurum ismi, yer, zaman, saat, kısaltma ve para gibi önceden belirlenmiş olan varlık sınıflarını dile bağımlı veya dilden bağımsız olarak bir doküman içerisinde arayıp bulan sistemlerdir.

Varlık İsmi Tanıma için kullanılan ilk sistemler kural tabanlı iken, son zamanlarda makine öğrenmesi teknikleri kullanılarak modern sistemler geliştirilmiştir. Hem kural tabanlı hem de makine öğrenmesini birlikte kullanan hibrit sistemler de mevcuttur.

Bu tez çalışmasında resmi olmayan bir dilde yazılmış Türkçe dokümanlar için tasarlanmış, Şartlı Rastgele Alanları kullanan Varlık İsmi Tanıma çalışmasından bahsedilmektedir. Sınıflama ve etiketleme işlemi kişi, kurum ve yer isimleri temel olmak üzere tarih ve para varlıkları için de gerçekleştirilmiştir. Çalışmanın gerçekleştirilmesi sırasında makine öğrenmesi teknikleri kullanılarak daha verimli sonuçlar elde etme işlemine odaklanılmıştır.

Çalışma üç adımdan oluşmaktadır. İlk olarak, resmi olmayan bir dilde yazılmış Türkçe e-postalar içerisinde Şartlı Rastgele Alanlar kullanılarak özel isimlerin etiketlenerek çıkartılması gerçekleştirilmiştir. Çalışmanın ikinci adımında, belli bir çalışma alanına bağlı olmayan, konudan bağımsız dokümanlar için Şartlı Rastgele Alanlar kullanılarak dokümanlardaki varlık isimlerinin belirlenmesi gerçekleştirilmiştir. Çalışmanın son

adımında ise, yarı eğitici teknik ile uygulanan çalışma kural tabanlı yöntem ile desteklenerek dokümanlardan varlık isimlerinin çıkartılması gerçekleştirilmiştir.

Varlık İsmi Tanıma için gerçekleştirilen çalışmada, eğitim verisi ne kadar çok etiketlenmiş varlık içerirse başarı oranı da o ölçüde etkilenmektedir. Çalışmada Şartlı Rastgele Alanlar yöntemi ile kişi isimlerinde elde edilen başarı kurum ve yer isimlerinde elde edilen başarıdan daha yüksek çıkmıştır.

Anahtar Kelimeler: Varlık İsmi Tanıma, Doğal Dil İşleme, Şartlı Rastgele Alanlar, Resmi Olmayan Türkçe, Türkçe E-Posta

ABSTRACT

NAMED ENTITY RECOGNITION BY CONDITIONAL RANDOM FIELDS FROM TURKISH INFORMAL TEXTS

Serap ÖZKAYA

Department of Computer Engineering

MSc. Thesis

Advisor: Assoc.Prof. Banu DİRİ

Named Entity Recognition is a subtask of information extraction that seeks to locate and classify predefined entities, such as names of persons, locations, organizations, etc. in unstructured texts.

Named Entity Recognition rule-based approach used for the first, while the recently developed modern systems using machine learning techniques. It is a hybrid system that uses both rule-based and machine learning are also available.

In this study, Conditional Random Fields has been used to extract name from informal texts. Classification and labeling is based on people, organization and location names including date and the money. The study of implementing more efficient by using machine learning techniques during to process focused on achieving results.

The study consists of three steps. First, Conditional Random Fields has been used to extract name entities which are person, location and organization names from informal Turkish e-posta. The second step of the study, Conditional Random Fields has been used to extract name entities from domain independent for formal and informal texts. In the last step of the study, Semi-supervised learning approach enrichment with the rule based approach has been used to extract name entities.

The training data contains so much labeled entity that the success rate can be influenced for the Named Entity Recognition system that includes a machine learning component.

Keywords: Named Entity Recognition, Natural Language Processing, Conditional Random Fields, Turkish informal texts, Turkish e-posta

GİRİŞ

Varlık İsmi Tanıma olarak Türkçe'ye çevirdiğimiz "Named Entity Recognition-NER", veri ve doküman madenciliği, doğal dil işleme, bilgi çıkarımı ve bilgiye erişim gibi birçok disiplinle ilişkilidir. Varlık İsmi Tanıma'nın amacı resmi olan veya olmayan bir dilde yazılmış, belli bir çalışma alanına bağlı veya bağımsız olan tüm dokümanlar içerisinde, dile bağımlı veya bağımsız olarak varlık isimlerini bulmak ve bunları kişi, kurum ve yer ismi, zaman, tarih ve parasal ifadeler olarak sınıflandırmaktır.

Doğal dil işleme alanında yaygın olarak kullanılan Varlık İsmi Tanıma ilk defa 1996 yılında 6. Mesaj Anlama Konferansı'nda (Sixth Message Understanding Conference-MUC-6) tanılmıştır [1]. Varlık İsmi Tanıma'ya yönelik yapılan mühendislik çözümlerinin hepsindeki standart yapı, girdi olarak bir dokümanı alan ve çıktı olarak bulunan varlıkların türleri ve sınıfları hakkında bilgi veren bir Varlık İsmi Tanıma sistemidir. Günümüzde Varlık İsmi Tanıma yöntemleriyle büyük doküman koleksiyonları çok daha çabuk ve verimli bir şekilde taranabilmektedir.

1.1 Literatür Özeti

Varlık İsmi Tanıma ile ilgili çalışmalar 1990'lı yıllarda başlamış olup, "Message Understanding Conference-MUC" platformunda farklı diller için yapılan çalışmalar ile devam etmiştir. Çalışmalarda özellikle yaygın olarak kullanılan dil İngilizce olup, Varlık İsmi Tanıma'da başarımlar %90'larda olmaktadır. İngilizce'nin dışında Arapça, Çince, Japonca, Hintçe gibi farklı dillerde de çok sayıda çalışma yapılmıştır. Bu alandaki ilk çalışmalardan biri Rau [2] tarafından yapılan doküman içerisinde şirket isimlerini çıkarmak ve tanımdır.

Varlık İsmi Tanıma için kullanılan ilk sistemler kural tabanlı iken, son zamanlarda makine öğrenmesi tekniklerini kullanan modern sistemlerde geliştirilmiştir. Hem kural tabanlı hem de makine öğrenmesini birlikte kullanan hibrit sistemler de mevcuttur. Yapılan çalışmalar, kullanılan dile (tek dil veya çok dilli), dokümanın türüne veya alanına ve varlık isminin çeşidine göre değerlendirilmektedir [1].

On beş yıldan fazla bir süredir, ileri temel bilgi ve mühendislik yöntemleri ile Varlık İsmi Tanıma için bir çok sistem geliştirilmiştir. Geliştirilen bu sistemlerin hemen hemen hepsinde veri girişi olarak bir doküman verilir ve çıktı olarak girilen dokümanda bulunması istenilen varlık isimleri işaretlenerek gösterilecek şekilde sonuç üretilir. İlk yapılan çalışmalarda çoğunlukla elle oluşturulan kural tabanlı yöntemler kullanılırken, son zamanlarda yapılan çalışmalarda ise eğitici makine öğrenme teknikleri kullanılmıştır [3].

Günümüzde Varlık İsmi Tanıma’da kullanılan öğrenme yöntemleri eğitici, yarı-eğitici ve eğitici olmamak üzere üç grupta ele alınmaktadır. Eğitici sistemlerde Saklı Markov Modelleri, Karar Ağaçları, Maksimum Entropy Modeli, Destek Vektör Makineleri ve Şartlı Rastgele Alanların kullanımı örnek olarak verilebilir. Bu sistemlerde işaretlenmiş geniş bir derlem, ezberlenecek varlıklar listesi ve belirsizliği giderecek kurallar kullanılır [1]. Yarı eğitici sistemlerde ise, başlangıçta öğrenmek için küçük bir küme kullanılarak eğitim yapılır, sonrasında sistem kendi kendine öğrenir. Eğitici yöntemlerde ise genel yaklaşım kümelemedir. Grup içi benzerlikler minimum, gruplar arası benzerlikler maksimum tutularak varlık isimleri tanınmaya çalışılır. Eğitici yöntemlerde bazen WordNet gibi sözlüksel kaynaklarda kullanılır [4].

Yarı denetimli öğrenme tekniğini temel alan Brin [5], 1998 yılında kitap başlıkları ile yazarlarının eşleştirildiği bir listeyi oluşturabilmek için belirli bir düzene göre işletilen sözcük özelliklerini kullanan bir sistem tasarlamıştır.

Türkçe için yapılan Varlık İsmi Tanıma çalışmaları ise sınırlıdır. Bunlardan ilki Cucerzan [6] tarafından gerçekleştirilen dilden bağımsız olarak Türkçe, İngilizce, Romence, Yunanca ve Hintçe dillerinde varlık tanıma çalışmasıdır. Türkçe finans haber dokümanları içerisinde kişi isimlerinin belirlenmesine yönelik yapılan çalışmada ise sadece Türkçe dilbilgisi yapısı kullanılarak varlık isimleri belirlenmiştir [7]. Türkçe için ilk

kural tabanlı varlık ismi tanıma çalışmasını yapan Küçük [8], haberler, makaleler, çocuk hikayeleri ve tarihsel metinler gibi farklı türlerdeki dokümanlarda sistemin başarısını ölçümlemiştir. Benzer bir çalışma [9] tarafından yapıp, Türkçe için konudan bağımsız kural tabanlı yer, kurum-kuruluş ve özel isimlerin doküman içerisinden etiketlenerek çıkartılmasına yönelik bir sistem geliştirmiştir.

Son zamanlarda makine öğrenmesi teknikleri kullanılarak Türkçe’de Varlık İsmi Tanıma’ya yönelik çalışmalar da yapılmıştır. Saklı Markov Model kullanılarak istatistiksel yaklaşım modeli ile İngilizce için geliştirilmiş Varlık İsmi Tanıma sisteminin bir benzeri başka bir çalışmada Türkçe için geliştirilmiş ve sonuçlar karşılaştırılmalı olarak verilmiştir [10]. Morfolojinin Varlık İsmi Tanıma için geliştirilen sistemlerdeki etkisinin belirlenmesine yönelik yapılan çalışmada ise, Şartlı Rastgele Alanlar kullanılmıştır. Morfolojik özelliklerin kullanılmasının yanında yeni bir sembolleştirme sunarak diğer çalışmalardan farklı bir yaklaşım [11] da sergilenmiştir.

Varlık İsmi Tanıma için kullanılan istatistiksel modellerden biri olan Şartlı Rastgele Alanlar’ın Türkçe için kullanıldığı başka bir çalışmada [26], Türkçe morfolojik özellikleri Şartlı Rastgele Alanlar ile birlikte uygulanarak genel haber metinlerinde varlık isimlerinin çıkartılmasına yönelik bir sistem geliştirilmiştir. Türkçe varlık isimlerinin belirlenmesinde morfolojik özelliklerin kullanılmasına yönelik başka bir çalışma yapan Çiçekli [27], diğer eklemeli dillerde de kullanılabilecek bir sistem geliştirmiştir. Varlık isimlerinin belirlenmesine yönelik hibrit bir sistem sunan çalışmada ise kural tabanlı tanıma işlemi istatistiksel yöntem (Rote learning-Freitag,1998) ile geliştirmiştir [28].

1.2 Tezin Amacı

Varlık İsmi Tanıma için yapılan çalışmaların çoğu İngilizce, Çince, Japonca ve Arapça için olurken, Türkçe için yapılan çalışmalar ise sınırlıdır. Özellikle İngilizce için büyük sözlük veritabanı olması ve çeşitli alanlarda yeteri kadar veri seti olmasından, Varlık İsmi Tanıma’ya yönelik yapılan çalışmalarda kullanılan yöntemler çeşitlilik göstermektedir. Türkçe bir dokümanda varlık isimlerinin belirlenmesine yönelik yapılan çalışmalarda ise genellikle iki yaklaşım görülmektedir. Birinci yaklaşım Türkçe dilbilgisi yapısı kullanılarak varlık isimlerinin belirlenmesi, ikinci yaklaşım ise kural tabanlı bir sistem geliştirilerek varlık isimlerinin belirlenmesidir.

Bu çalışmada, Türkçe için konudan bağımsız olarak belli bir çalışma alanına bağlı veya bağımsız dokümanlarda Şartlı Rastgele Alanlar (Conditional Random Fields) kullanılarak kişi, yer ve kurum-kuruluş isimlerinin etiketlenerek çıkartılması için bir çalışma yapılmıştır. Bu çalışma ile yarı eğitici makine öğrenme tekniği kullanılarak uygulanan yöntemde daha verimli sonuçlar çıkartılmaya odaklanılmıştır.

Yapılan çalışma üç adımdan oluşmaktadır. İlk olarak, resmi olmayan bir dilde yazılmış Türkçe e-postalar içerisinde Şartlı Rastgele Alanlar kullanılarak özel isimlerin etiketlenerek çıkarılmasına yönelik bir uygulama geliştirilmiştir. Yapılandırılmış doküman formatına sahip olan e-posta'lardan elde edilen deneysel sonuçlar 5.2'de değerlendirilmiştir.

Çalışmanın ikinci adımında, belli bir çalışma alanına bağlı olmayan konudan bağımsız, resmi olmayan dokümanlar için yarı eğitici makine öğrenmesi tekniği olan Şartlı Rastgele Alanlar kullanılarak dokümanlardaki varlık isimleri çıkarılmıştır. Elde edilen deneysel sonuçlar bölüm 5.3'de değerlendirilmiştir.

Çalışmanın son adımında ise, yarı eğitici teknik ile uygulanan çalışma kural tabanlı yöntem ile zenginleştirilip dokümanlardan varlık isimlerinin çıkartılması hedeflenmiştir. Elde edilen sonuçlar karşılaştırmalı olarak bölüm 5.3'de sunulmuştur.

1.3 Hipotez

Varlık İsmi Tanıma, varlıkların anlamsal açıdan sınıflandırılmasıdır. Varlık isminin iki temel amacı vardır. Birincisi varlıkların bulunup tanınması, ikincisi ise bu varlıkların sınıflandırılmasıdır. Türkçe için yapılan çalışmalarda çoğunlukla yerel dilbilgisi yaklaşımı veya kural tabanlı yaklaşım kullanılmıştır.

Yerel dilbilgisi yaklaşımında varlık isimlerinin belirlenebilmesi için net bir şekilde örüntü oluşturulması gerekmektedir. Ayrıca, yerel dilbilgisi yaklaşımının Türkçeye uygulanabilirliği için uygunluk analizi ve eşdizimlik analizine ihtiyaç bulunmaktadır. Yerel dilbilgisi yaklaşımının resmi olmayan dokümanlarda ki başarı oranı ise yükseltilememektedir.

Varlık İsmi Tanıma'da kural tabanlı yaklaşım çok maliyetli olup, başarıyı yükseltmek için çok kural tanımlamak gerekmektedir. Yerel dilbilgisi yaklaşımına benzer şekilde resmi

olan ve yapılandırılmış doküman formatlarında başarı oranı daha yüksek iken konudan bağımsız resmi olmayan dokümanlardaki başarı oranı ise daha düşüktür. Ayrıca, günümüzde kural tabanlı yaklaşımların çoğu belli bir alana göre hazırlanmış doküman türleri için kullanılmaktadır. Bu şekilde geliştirilen bir sistem başka bir alana ait dokümanlar için uygulanmak istenildiğinde, tanımlanan kurallar yetersiz kalabilmektedir.

Araştırmalar göstermiştir ki; belli bir alana göre oluşturulan Varlık İsmi Tanıma sistemleri başka bir alan için kullanıldığında yetersiz kalmaktadır. Bu çıkarım kural tabanlı, sembolik yaklaşım ve olasılık yaklaşımları için de geçerlidir [3].

Türkçe morfolojik olarak zengin bir dil olup, Varlık İsmi Tanıma'da kullanılan istatistiksel modellerden biri olan Gizli Markov Model tek başına verimli olmamaktadır. Karmaşık dil yapısına sahip olan dillerde Gizli Markov Model olasılık tanımlamada yetersiz kalmaktadır [15].

Varlık İsmi Tanıma'da kullanılabilen Şartlı Rastgele Alanlar kural tabanlı uygulamalara göre yapılandırılmamış dokümanlarda çok daha yüksek başarı oranı sunmaktadır. Ayrıca, Şartlı Rastgele Alanlarla geliştirilen Varlık İsmi Tanıma çalışmalarında doküman türünün bir önemi olmadığından farklı alanlar için kullanıldığında başarı oranında bir değişme görülmemektedir.

Şartlı Rastgele Alanlar istatistiksel modellerden biri olup, Maksimum Entropi Markov Model ve Gizli Markov Modeli ile birlikte kullanılarak bir olasılık sunduğundan tek başına Gizli Markov Model'e göre de daha yüksek başarı göstermektedir. Varlık İsmi Tanıma için geliştirilen çalışmada başarı oranını yükseltiği için Şartlı Rastgele Alanlar'ın kullanılması tercih edilmiştir.

ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ TANIMA

Doğal dil işlemede bilgi çıkarımı yöntemleri kullanılarak bir doküman içerisindeki belirli kriterlere uyan bilgilere erişim yapılabilmektedir. Bu işlem sırasında örneğin, bir kalıba uygun olan verilerin çıkartılması istenebilir. Bilgi çıkarım işleminde amaç çok miktardaki veriyi otomatik olarak işleyen bir yazılım üreterek, insan müdahalesini minimum seviyeye indirmektir. Bilgi çıkarımın bir alt dalı olan Varlık İsmi Tanıma'da da bir veri kaynağında geçen varlıkların ve isimlerinin algılanması hedeflenir. Bu varlıklar kişi, kurum, yer, soyut varlıklar veya sayısal ifadeler olabilir.

Varlık İsmi Tanıma'da kullanılan çeşitli yöntemler olmasına karşın günümüzde makine öğrenmesi teknikleri kullanılarak yeni sistemler geliştirilebilmektedir. Eğitici sistemlerden Şartlı Rastgele Alanlar kullanılarak dokümandaki varlık isimleri belirlenebilmektedir.

2.1 Sıralı Verinin Etiketlenmesi

Bir dizi gözlem dizilerine etiket sıralarının atanması biyobilişim, bilişimsel dilbilim ve konuşma tanıma [6, 10, 12] dâhil olmak üzere birçok alanda yapılmaktadır. Doğal dil işlemede sözcük türü analiz ile sıralı veri etiketlenebilmektedir. Örneğin;

<İsim>Ahmet</İsim><TamlayanEki>in</TamlayanEki><İsim>baba</İsim><TamlananEki>ı</TamlananEki>,<Sifat>beyaz</Sifat><İsim>koyun</İsim><NesneEki>u</NesneEki><İsim>araba</İsim><DiğerZarflar>ile</DiğerZarflar><İsim>köy</İsim><DolaylıTümleçEki>e</DolaylıTümleçEki><Yüklem>getir</Yüklem><ZamanEki>di</ZamanEki>

Cümlelerin bu şekilde etiketlenmesi, daha üst düzey doğal dil işleme görevleri için faydalı bir işleme adımıdır: Sözdizimsel analiz, açık olarak dilin özünde bulunan bazı yapılara sadece işaret etmek suretiyle sözcüklerin içerdiği bilgileri arttırır [19].

Böyle bir etiketleme ve parçalara ayırma görevlerinin gerçekleştirilmesi için kullanılan en yaygın yöntemlerden birisi, herhangi bir cümledeki sözcükler için en olası etiket dizisini tanımlamak amacıyla, Gizli Markov Modelleri (GMM) [13] ya da olasılıksal sonlu özdevininin kullanılmasıdır.

Gizli Markov Modeli doğal dil işleme, ses tanıma, video işleme ve bunun gibi zamana bağlı değişkenlerin olduğu alanlarda kullanılan bağlantı tabanlı bir modeldir. GMM yapısı itibariyle Markov Model teorisine dayanmaktadır. Markov Model yapısı ardışık olarak gelen düğümlerin (çizge modeli) olasılığının bulunmasında kullanılır. Markov Model yapısında düğümler sadece gözlemlenen varlıkları ifade eder. Ancak, GMM gözlemleyemediğimiz ve varsaydığımız bağlantılar için saklı düğümler oluşturarak bağlantıları bizim varsayımımıza göre yapar.

Markov Model ile bir denklemin matematiksel bir modeli çıkarılırken, GMM ile bu denkleme yakınsayan bir model çıkarılır [10]. GMM iki stokastik süreç içerir. İlk olan Markov süreci, zaman ile ilgili değişikliklerde kullanılır ve durumları içeren bir Markov zinciri üretir. Diğer süreç gözlemlenebilir olan özellik parametrelerini veya gözlemler denilen rastgele değişkenleri içerir [22].

GMM'ler, gözlem dizileri ve bunlarla ilişkili etiket sıraları boyunca sırasıyla yayılan X ve Y 'nin rastlantısal değişkenler olduğu bir $p(X,Y)$ ortak olasılık dağılımını tanımlayan üretken bir model biçimindedir. Bu yapıdaki bir ortak dağılımı tanımlamak için üretken modellerin tüm olası gözlem öğelerini sıralaması gerekir ve bu görev, bir gözlem dizisinde diğer elemanlardan bağımsız olarak gözlem elemanları izole edilmiş birimler şeklinde temsil edilmedikçe birçok alan için zorlu bir görevdir. Zaman içinde herhangi bir andaki gözlem elemanı sadece o zamandaki bir duruma ya da etikete doğrudan bağlı olabilir. Bu, birkaç basit veri dizileri için uygun bir varsayımdır [19].

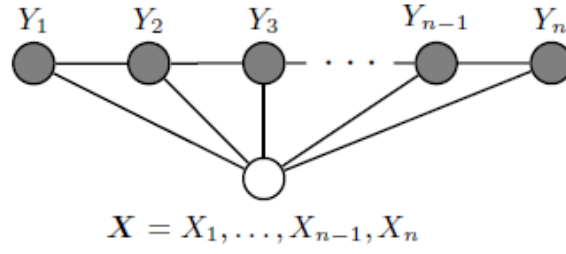
Bu temsil sorunu sıralı verilerin etiketlenmesinde en temel problemlerden birisidir. İzlenebilir çıkarımı destekleyen bir model gereklidir ama, nedensiz bağımsız varsayımlar yapmadan verileri temsil eden bir model de arzu edilmektedir. Bu her iki kriteri yerine

getiren bir yol, hem etiket hem de gözlem dizileri boyunca bir ortak dağılım yerine, belirli bir x gözlem dizisinde etiket dizileri boyunca $p(Y|x)$ şartlı olasılığını tanımlayan bir modelin kullanılmasıdır. Şartlı modeller, $p(y_*|x_*)$ şartlı olasılığını arttıran y_* etiket dizini seçmek suretiyle x_* yeni gözlem dizini etiketlemek için kullanılır. Bu tür modellerin koşullu doğası, gözlemleri modellemede herhangi bir çaba harcanmadığı ve bu diziler hakkında nedensiz bağımsız varsayımlar yapmak zorunda olmadığı anlamına gelir; modelleyicinin bu özelliklerin nasıl ilişkili olduğu hakkında endişelenmesine gerek kalmadan, gözlem verisinin rastgele özellikleri bu modelle elde edilebilir.

Şartlı Rastgele Alanlar [19] sıralı verilerin etiketlenmesi ve parçalara ayrılması için, koşullu yaklaşıma dayalı bir olasılıksal çerçevedir. Şartlı Rastgele Alanlar, belirli bir gözlem dizisindeki etiket dizileri boyunca bir tek log-lineer dağılımı tanımlayan yönsüz bir grafik modeli biçimidir. Şartlı Rastgele Alanların Gizli Markov Modele göre temel avantajı, GMM'lerin gerektirdiği bağımsız varsayımlardan rahatlamaya neden olacak şekilde şartlı yapıda, olmalarıdır. Buna ek olarak, Şartlı Rastgele Alanlar, Maksimum Entropi Markov Modelleri [10] (MEMM'ler) ve yönlendirilmiş grafik modellerine dayalı diğer şartlı Markov modelleri tarafından sergilenen bir zayıflık olan etiket önyargı problemini önlerler. Şartlı Rastgele Alanlar çok sayıda dizi etiketleme görevlerinde hem MEMM'ler hem de GMM'lerden üstündür [2, 14, 15].

2.2 Yönsüz Grafik Modelleri

Şartlı Rastgele Alan, gözlem dizilerini temsil eden rastgele değişken olan X üzerinde global olarak devam eden yönsüz bir grafik model ya da Markov Rastgele Alan [3] olarak görülebilir. Formal olarak, Y 'nin bir Y_v elemanını temsil eden her bir rastgele değişkene denk düşen bir $v \in V$ düğümü olacak şekilde bir yönsüz grafik olarak $G = (V, E)$ şeklinde tanımlanır. Şayet, her bir Y_v rastgele değişkeni, G yönünden Markov özelliğine uyarsa, bu durumda (Y, X) bir şartlı rastgele bir alandır. Teoride, modellenmekte olan etiket dizilerinde koşullu bağımsızlıkları temsil etmesi şartıyla, G grafiğinin yapısı rastgele olabilir. Ancak, diziler modellenirken karşılaşılan en basit ve en yaygın grafik yapısı, Şekil 2.1'de gösterildiği gibi, Y biçimi elemanlarına denk düşen düğümlerin basit birinci derece zincir oluşturmasıdır. Şekil 2.1'de görüldüğü gibi Y , basit birinci dereceden zincir oluşturur [19].



Şekil 2.1 Diziler için zincir yapılı Şartlı Rastgele Alanların grafik yapısı. Gölgesiz düğümlere denk düşen değişkenler model tarafından üretilmemiştir.

2.2.1 Potansiyel Fonksiyonlar

Şartlı Rastgele Alanın grafik yapısı, Y 'nin Y_v elemanları boyunca ortak dağılımını, koşullu bağımsızlık kavramından türetilmiş pozitif gerçek değerli potansiyel işlevlerin normalleştirilmiş bileşkesini çarpanlara ayırmak için kullanılabilir. Her bir potansiyel işlev, G 'deki köşe noktaları tarafından temsil edilen rastgele değişkenler alt dizisinde işlev görür. Yönsüz grafik modelleri için şartlı bağımsızlık tanımına göre, G 'deki iki köşe noktası arasında bir köşenin yokluğu, bu köşe noktaları tarafından temsil edilen rastgele değişkenlerin bu modeldeki tüm rastgele değişkenlerde şartlı olarak bağımsız olduğunu işaret eder. Bu nedenle potansiyel işlevlerin, şartlı bağımsız değişkenlerin aynı potansiyel işlevde görünmeyecek şekilde, ortak olasılığı çarpanlara ayırmanın mümkün olduğunu sağlaması gerekir. Bu gerekliliği yerine getirmenin en kolay yolu, her bir potansiyel işlevin ilişkili köşe noktalarının G içinde maksimal grup oluşturan rastgele değişkenler dizisinde işlev görmesini gerektirmektir. Bu, hiçbir potansiyel işlevin, köşe noktaları doğrudan bağlantılı olmayan herhangi bir rastgele değişkenler çiftine işaret etmemesini ve şayet iki köşe noktası bir grupta görünürse, bu ilişkinin açık olmasını sağlar. Şekil 2.1'de sergilendiği gibi zincir yapılı Şartlı Rastgele Alanlar durumunda ise her bir potansiyel işlev, Y_i ve Y_{i+1} komşu etiket değişkenlerinde işlev görmektedir.

İzole edilmiş potansiyel bir işlevin doğrudan olasılıksal yorumu olmayıp, bunun yerine tanımlanan işlevdeki rastgele değişkenler konfigürasyonunda sınırlamaları temsil etmektedir. Bu da sonuçta global konfigürasyonların olasılığını etkiler; yani yüksek

olasılığa sahip bir global konfigürasyon, düşük olasılığa sahip bir global konfigürasyona göre bu sınırlamalardan bir çoğunu yerine getirmesi olasıdır.

2.3 Şartlı Rastgele Alanlar

Lafferty ve diğerleri [25] x gözlem dizisindeki belirli bir y etiket dizisinin, her birinin Eşitlik 2.1’de oluşturulduğu gibi potansiyel işlevlerin normalleştirilmiş bileşkesi olacağını tanımlar:

$$\exp \left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i) \right) \quad (2.1)$$

burada $t_j(y_{i-1}, y_i, x, i)$ tüm gözlem dizisinin ve etiket dizisindeki i ve $i-1$ konumlarındaki etiketlerin geçiş özellik işlevidir; $s_k(y_i, x, i)$ ise i konumundaki etiket ve gözlem dizisinin durum özellik işlevidir; λ_i ve μ_k çalışma verilerinden tahmin edilecek parametrelerdir.

Özellik işlevlerini tanımlarken, model dağılımını tutması gereken eğitim verilerinin ampirik dağılımının bazı karakteristiklerini ifade etmek için, gözlemin gerçek değerli özellikler $b(x, i)$ dizisi oluşturulur. Böyle bir özelliğin örneği şu şekildedir:

$$b(x, i) = \begin{cases} 1 & i \text{ konumundaki gözlem "İstanbul" kelimesi ise} \\ 0 & \text{Diğer durum} \end{cases} \quad (2.2)$$

Şayet mevcut durum (durum işlevi durumunda) ya da önceki ve tüm durumlar (geçiş işlevi durumunda) belirli değerleri alırsa, her bir özellik işlevi, bu gerçek değerli gözlem özelliklerinden $b(x, i)$ ’den birisinin değerini alır. Bu nedenle tüm özellik işlevleri gerçek değerlidir. Örneğin, aşağıdaki geçiş işlevini düşünün:

$$t_i(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) & \text{Eğer } y_{i-1} \text{ sıfat ise, } y_i \text{ isimdir} \\ 0 & \text{Diğer durum} \end{cases} \quad (2.3)$$

Ayrıca, bu notasyon aşağıdaki gibi yazılarak basitleştirilmiştir:

$$s(y_i, x, i) = s(y_{i-1}, y_i, x, i) \quad (2.3)$$

ve

$$F_j(y, x) = \sum_{i=1}^n f_j(y_{i-1}, y_i, x, i) \quad (2.4)$$

Burada her bir $f_j(y_{i-1}, y_i, x, i)$ 'ye ya bir $s(y_{i-1}, y_i, x, i)$ durum işlevidir ya da bir $t(y_{i-1}, y_i, x, i)$ geçiş işlevidir [20]. Bu durum ise bir x gözlem dizisindeki y etiket dizisi olasılığının Eşitlik 2.5'deki gibi yazılmasına izin verir:

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(y, x)\right) \quad (2.5)$$

$Z(x)$ bir normalleştirme faktörüdür.

2.4 Maksimum Entropi

Şartlı Rastgele Alanlar örüntü tanımada kullanılan istatistiksel modellemelerden biridir. Şartlı Rastgele Alanlar sıralı veriyi işaretlemek ve bölümlere ayırmak için kullanılan Maksimum Entropi Markov Model ve Gizli Markov Modelin genel halini yansıtan bir olasılık sunar.

Şartlı Rastgele Alanın (2)'de verilen şeklinde ağırlıklı olarak Maksimum Entropi ilkeleri kullanılarak, eğitim veri setinden tahmini olasılık dağılımı hesaplanmaktadır. Olasılık dağılım entropisi bir belirsizlik ölçüsüdür ve söz konusu dağılım mümkün olduğu kadar homojen olduğunda maksimize edilir. Maksimum Entropi prensibi sonlu eğitim verileri gibi yetersiz bilgi üzerinden elde edilen olasılık dağılımında bile mevcut kısıtlamaları temsil eden bir dizi maksimum entropi olduğunu belirtmektedir. Başka bir dağılım yersiz varsayımlar içerecektir [7].

Eğitim veri seti içerisinde bulunan bilgiler bir dizi özellik fonksiyonu ile sunulmaktadır. Maksimum Entropi Model dağılımı eğitim verisinde her bir özellik fonksiyonunun sunduğu ampirik dağılımın mümkün olduğunca homojen olmasını beklemektedir [16,21].

BÖLÜM 3

SİSTEM TASARIMI

Bilgi çıkarımının bir alt dalı olan Varlık İsmi Tanıma'da bir veri kaynağında geçen varlıkların bulunması ve sınıflandırılması hedeflenir. Varlık İsmi Tanıma'ya yönelik geliştirilen sistemde ise resmi olmayan bir dilde yazılmış, belli bir çalışma alanından bağımsız olan tüm Türkçe dokümanlar içerisinde, varlık isimlerini bulmak ve bunları kişi, kurum ve yer ismi olarak sınıflandırmak amaçlanmıştır.

Türkçe'de Varlık İsmi Tanıma'ya yönelik yapılan çalışmalarda kural tabanlı yaklaşım çok kullanılmasına rağmen, son zamanlarda makine öğrenmesi teknikleri de kullanılmaktadır. Varlık İsmi Tanıma'ya yönelik geliştirilen bu sistemde, makine öğrenme tekniklerinden yarı eğitici bir yöntem olan Şartlı Rastgele Alanlar kullanılarak doküman türü ve içerikten bağımsız bir tasarım gerçekleştirilmiştir. Sistemin gerçekleşmesi sırasında varlık isimlerini içerisinde tutan herhangi bir sözlüksel kaynak kullanılmamıştır.

Çalışmada ilk önce resmi olmayan bir dilde yazılmış Türkçe e-postalar içerisinde Şartlı Rastgele Alanlar kullanılarak varlık isimlerinin bulunup, sınıflandırılması gerçekleştirilmiştir.

Türkçe e-postalarda elde edilen başarı ile uygulamanın belli bir formata ve çalışma alanına bağlı olmayan doküman türleri için geliştirilmesi hedeflenmiştir. Sadece Şartlı Rastgele Alanlar kullanılarak varlıkların belirlenebileceği görülmüştür.

Çalışmanın son aşamasında geliştirilen sistem, Türkçe'nin temel özelliklerine göre kurallar eklenerek kural tabanlı yaklaşım ile zenginleştirilmiştir. Böylece, hem makine

öğrenmesi hem de kural tabanlı yaklaşım yöntemi birlikte kullanılarak yüksek başarı elde edilmeye odaklanılmıştır.

3.1 Kullanılan Yöntem

Geliştirilen sistemde ön çalışma olarak doküman içerisindeki kelimeler ayırt edilmiştir. Kelimelerin birbirinden ayırt edilmesinde kelimeler ve noktalama işaretlerinden sonra gelen boşluklardan yararlanıldığı gibi, resmi olmayan yapılandırılmamış dokümanlarda noktalama işaretleri de dikkate alınmıştır.

Varlık İsmi Tanıma’da örüntü tanımada olduğu gibi olasılık değerlerini belirleyebilmek için kullanılan istatistiksel modellerden Şartlı Rastgele Alanlar yarı eğitici makine öğrenmesi tekniği ile uygulanmıştır. Sistemi eğitebilmek için ayrıca eğitim verisi oluşturulmuş, eğitim verisi artırıldıkça da olasılık değerlerinin daha verimli belirlenebildiği görülmüştür.

Varlık İsmi Tanıma sistemi için geliştirilen uygulama Java programlama dili kullanılarak gerçekleştirilmiştir.

3.2 Türkçe E-Posta için Yapılan Çalışma

E-postalar, yarı yapılandırılmış bir formata sahip olup, resmi olmayan bir dilde yazılmış Türkçe e-postalarda yer, kurum-kuruluş ve özel isimlerin etiketlenerek çıkartılması işlemi yapılmıştır. E-postaların belirli bir formata sahip olması özelliğinden yararlanılarak kural tabanlı yaklaşım ve Şartlı Rastgele Alanlar birlikte uygulanmıştır.

E-posta içerisindeki her kelime bir nesne olarak düşünülmüştür. Oluşturulan bu nesnelerde kelime ve kelimenin uzunluğu, adedi, kısaltma olup olmadığı, tipi (isim, sıfat, zamir, vd.), cümlelerin ilk kelimesi mi gibi özellik değişkenleri tutulmaktadır. Uygulama çalıştırıldığında, veri setindeki her e-posta ve içerisindeki her farklı kelime için bir nesne oluşturulup:

- Özellik çıkarımında kullanılmak üzere kelime sıklıkları ve karakter tabanlı 1-gram, 2-gram, 3-gram’lar çıkartılır.
- Çıkartılan her kelimenin daha önceden oluşturulan kaynaklar listesinde yer alıp almadığı kontrol edilir.

- Kelimenin büyük harfle başlaması, noktalama işaretlerini içermesi, başlık bilgisinde yer alması gibi özelliklere bakılarak, her özellik için kelime nesnesindeki özellik etiket değerine 1 atanır.
- Etiketleme işlemi bittikten sonra veri setinden çıkartılan özellikler kullanılarak sınıflandırma için optimizasyon parametreleri belirlenir.

Son aşamada Şartlı Rastgele Alanlar yöntemiyle etiket sınıfı belirlenir. Ancak, sınıflandırma işleminden önce uygulamaya belirlenen kural tanımları verilmelidir. Örneğin, bir kelimenin kişi ismi olabilmesi için çıkarılan kurallardan bir kaç aşağıda verilmektedir.

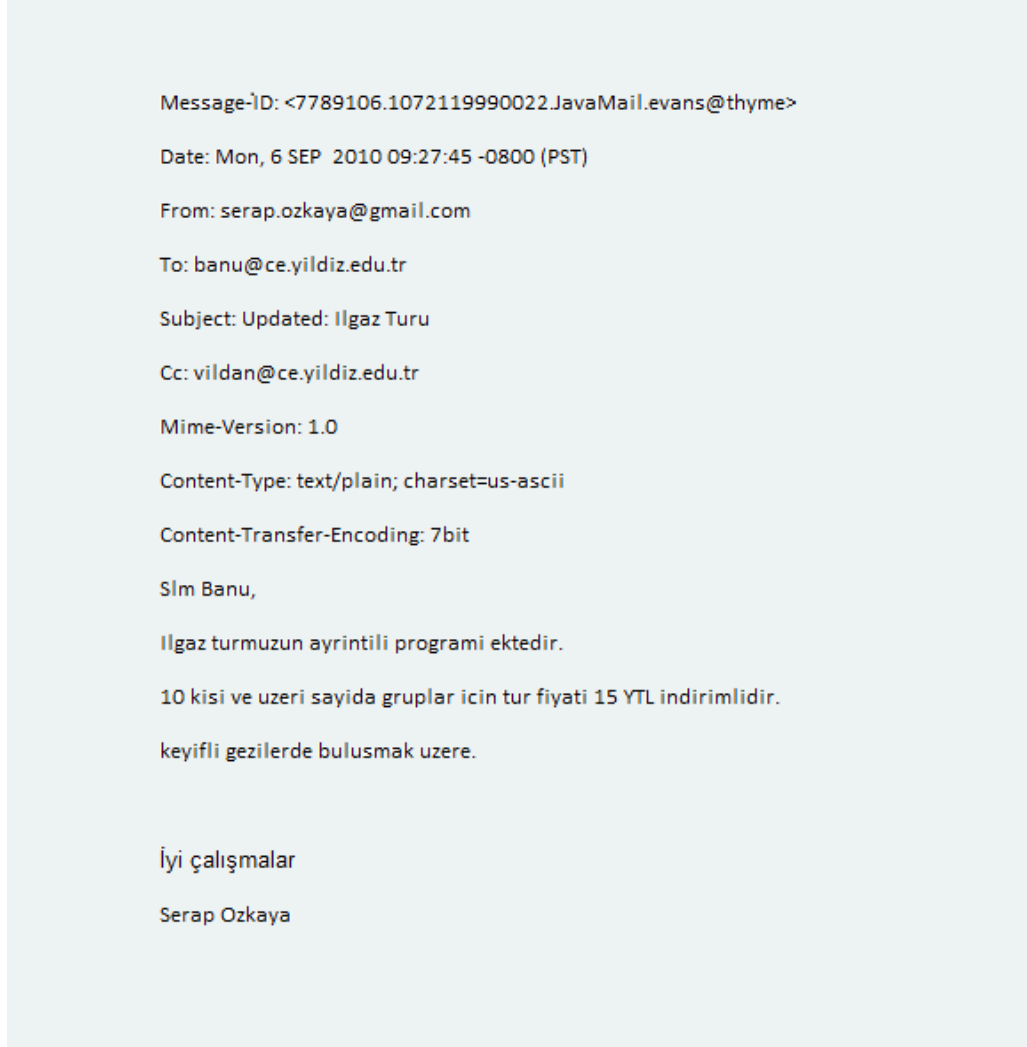
<kisi_ismi>From/BüyükHarfleBaşlama/<kisi_ismi>
Örneğin: <u>From:Serap Özkaya</u> serap.ozkaya@gmail.com
<kisi_ismi>ForwardedBy/BüyükHarfleBaşlama/BüyükHarfleBaşlama<kisi_ismi>
Örneğin: <u>Forwarded By Banu Diri</u>
<kisi_ismi>Ünvan/BüyükHarfleBaşlama/<kisi_ismi>
Örneğin: <u>Dr. Haluk Bey</u>
<kisi_ismi>BüyükHarfleBaşlama/Ünvan<kisi_ismi>
Örneğin: Merhaba <u>Ahmet Bey</u>
<kisi_ismi>DiğerÖzelKelime/BüyükHarfleBaşlama<kisi_ismi>
Örneğin: <u>Sayın Banu Hanım</u>
<kisi_ismi>DiğerÖzelKelime/Ünvan/BüyükHarfleBaşlama<kisi_ismi>
Örneğin: <u>Saygılarımla Av. Ayşen</u>

Şekil 3.1 Örnek Kural Listesi

3.2.1 E-Posta Veri Seti

E-postalar yazım kurallarına dikkat edilerek yazılmış dokümanlar olmasada kendilerine özgü bir formata sahiptirler. Bir e-posta iletilsinin özelliklerinden biri daha az söz

dizimine sahip iyi tanımlanmış bir başlık bilgisi ve iletinin sonunda yer alan ve genellikle kişi ismi olan imza bilgisidir. Ancak, e-postalarda metin bilgisi yapısal ve gramer olarak bir çok hatayı içerisinde bulunduran, alıcı ve gönderici tarafından anlaşılabilen özel kısaltmalar ve jargonları da içermektedir (Şekil 3.1).



Şekil 3.2 Örnek E-Posta

Resmi olmayan Türkçe dokümanlardan kişi, kurum ve yer varlık isimlerini çıkartmak için akademik, kurumsal ve kişisel e-postalardan veri seti oluşturulmuştur. E-postalar, maksimum 792 kelime, minimum da 98 kelime olmak üzere ortalama 158 kelime uzunluğundadır. Veri seti herbiri 50 adet olmak üzere akademik, kurumsal ve kişisel e-postalardan eşit sayıda toplanmış toplam 150 adet e-postadan oluşmaktadır.

E-postalarda Varlık İsmi Tanıma'ya yönelik oluşturulan sistemde temel özelliklerin çıkartılmasında yardımcı olan kısaltmalar, ünvanlar ve bazı özel kelimelerin tutulduğu sözlükler kullanılmıştır. Bu çalışmada 175 adet kısaltma listesi (İng., Fr., Tşk, Sok., Mah., TD, Sn, A.Ş., Ltd., vd.), 35 adet ünvan listesi (Prof., Dr., Av., Gen., vd.), ve 32 adet özel kelimeler listesi (Sayın, Hanım, Bey, Hocam, Üniversite, Bakanlık, Hastane, Dağ, Tepe,...) kullanılmıştır. Detayları Ek-1'de verilmiştir.

3.2.2 Temel Özellikler

Temel özellikler, herhangi bir kaynak kullanılmadan sadece veri setinden çıkartılmaktadır. E-posta için çıkartılan bu özellikler; e-posta başlık bilgisi özellikleri, etiketlenebilen özellikler ve kural tabanlı özellikler şeklinde 3 farklı grupta değerlendirilmektedir.

3.2.2.1 Başlık Bilgisi Özellikleri

Her e-posta bir başlık bilgisiyle başlamaktadır. İyi tanımlanmış bir başlık bilgisinde varlık isimleri yer alabileceğinden uygulamada aşağıdaki başlık bilgisi alanlarına bakılmaktadır.

- From, To, Cc ,Bcc, Send, Sender, In-Reply-to, Followup-to, Reply-to, Forwarded by
- Konu bilgisi

3.2.2.2 Etiketlenebilen Özellikler

Şartlı Rastgele Alan yönteminde, veri setindeki kelimelerin özellikleri etiket bilgisi ile gösterilebilmektedir. Örneğin, "ilk_kelime_mi=1" ataması kelimenin bir cümlede ilk kelime olduğunu göstermektedir. Eğer kelime cümledeki ilk kelime değilse bu değer 0 olacaktır [16].

E-postada geçen bir kelime aşağıdaki özelliklere sahip olduğu durumda etiket değerlerine 1 atanır. Böylece kelimenin sahip olduğu özellik bilgisi tutulur.

- Büyük harfle başlama : Kelime büyük harf ile başlıyor ise,

- Noktalama işaretleri : Kelime (-; ,” “) gibi noktalama işaretleriyle birlikte yazılmış ise,
- İlk kelime : Cümlelerin ilk kelimesi ise,
- Başlık bilgisi : Kelime başlık bilgisinde yer alıyorsa,
- Ünvanla başlama : Kelimeden önce gelen kelime ünvan listesinde var ise,
- Özel kelimelerle başlama : Kelimeden önce gelen kelime özel kelimeler listesinde var ise,

özellik etiket değerlerine 1 ataması yapılır.

3.2.2.3 Kural Tabanlı Özellikler

Kural tabanlı özellikler bir varlık isminin kişi, yer ve kurum isimlerinden biri olup olmadığının tahmininde kullanılmak için e-postalardan çıkartılan özelliklerdir.

- Ard arda gelen kelimeler : Veri setinde ele alınan her kelime için kelimeden önce gelen ve kelimeden sonra gelen kelimeler belirlenir. Bu özellik daha çok isim-soy isim eşleştirmesinde kullanılmaktadır.
- Karakter tabanlı 1-gram, 2-gram, 3-gram : Bazı son ekleri belirleyebilmek için bu özellik kullanılmaktadır.
- Kelime uzunluğu : Kelimenin kısaltma, ünvan, vd. olup olamayacağı tahmin edilebilmektedir.
- İsim, Soy İsim : Ard arda gelen kelimelerden kişi ismi olabilecek kelimelerin hangisinin isim, hangisinin soy isim olduğunun belirlenmesidir.

3.3 Doküman Türünden Bağımsız Şartlı Rastgele Alanlar Kullanımı

E-posta dokümanlarında elde edilen deneyim ve tecrübe ile birlikte resmi olmayan bir dilde yazılmış belli bir çalışma alanına bağlı olmayan Türkçe dokümanlar için Varlık İsmi Tanıma sistemi geliştirilmiştir. Varlık İsmi Tanıma’da örüntü tanımadaki olduğu gibi olasılık değerlerini belirleyebilmek için kullanılan istatistiksel modellerden Şartlı Rastgele Alanlar yarı eğitici makine öğrenmesi tekniği ile uygulanmıştır. Sistemi

eğitebilmek için ayrıca eğitim verisi oluşturulup, eğitim verisi artırıldıkça olasılık değerleri daha verimli belirlenebilmektedir. Bu çalışmada herhangi bir sözlüksel kaynak ya da başka bir dış kaynak kullanılmamıştır.

3.3.1 Eğitim Veri Seti

Varlık İsmi Tanıma’da yarı eğitici makine öğrenmesi tekniğini kullanabilmek için eğitim veri setinin oluşturulmasında iki ayrı çalışma yapılmıştır. İlk çalışmada, Varlık İsmi Tanıma’ya yönelik daha önceden yapılan çalışmalarda kullanılan eğitim veri setleri toplanıp sistemin beklediği formata göre hazırlanmıştır. Bu kapsamda gazete makalelerinden derlenerek oluşturulan ve [10] (Tur et al., 2003) çalışmasında kullanılan veri seti temin edilmiştir.

Eğitim veri setinin oluşturulmasında ikinci çalışmada ise farklı alanlarda derlenen dokümanlar elle özel anlamlı varlıkların belirlenmesiyle oluşturulmuştur.

Elde edilen eğitim veri setinde varlık isimlerinin etiketlenmesinde MUC-6 ve MUC-7 standartlarına göre yirmi araştırma görevlisi tarafından belirlenen etiketleme standardı kullanılmıştır[1].

<ENAMEX TYPE="PERSON">Kişi İsmi</ENAMEX>

<ENAMEX TYPE="ORGANIZATION">Kurum İsmi</ENAMEX>

<ENAMEX TYPE="LOCATION">Yer İsmi</ENAMEX>

<TIMEX TYPE="DATE">Zaman</TIMEX>

<NUMEX TYPE="MONEY">Para</NUMEX>

Şekil 3.3 MUC standartları

Kişi ismi ya da isim-soyisim “PERSON” etiketi ile, kurum, şirket ya da dernek ismi “ORGANIZATION” etiketi ile, siyasi ya da coğrafi olarak tanımlanan konum isimleri (şehir, ülke, uluslar arası bölge, dağ, nehir, vd.) “LOCATION” etiketi ile, gün, ay ve yıl değerleri “DATE” etiketi ile para miktarı ise “MONEY” etiketi ile belirtilmektedir. Örneğin;


```
<ENAMEX TYPE="ORGANIZATION">Davutpaşa Kışlası</ENAMEX> <TIMEX
TYPE="DATE">1999</TIMEX> yılında <ENAMEX TYPE="LOCATION">İstanbul</ENAMEX>'da
bulunan <ENAMEX TYPE="ORGANIZATION">Yıldız Teknik Üniversitesi</ENAMEX>
mülkiyetine verilmiştir.
```

Şekil 3.4 Örnek MUC standardı

3.3.2 Olasılık Hesabı

Varlık İsmi Tanıma'da örüntü tanıma için olasılık değerlerinin belirlenmesinde Şartlı Rastgele Alanlar kullanılmaktadır. Şartlı Rastgele Alanda ağırlıklı olarak Maksimum Entropi ilkesi kullanılarak, eğitim veri setinden tahmini olasılık dağılımı hesaplanmaktadır.

Eğitim verilerinin $\{(x^{(k)}, y^{(k)})\}$ bağımsız ve eşit dağıtılmış olduğu varsayıldığında, Eşitlik 3.1'deki fonksiyon parametresi olan λ 'nın tüm eğitim serisinin olabilirliği (likelihood) $p(\{y^{(k)}\}|\{x^{(k)}\}, \lambda)$ 'de ifade edildiği şekildedir. Maksimum olabilirlik eğitimi olabilirlik logaritmasında (log-likelihood), olasılığı maksimize edecek şekilde parametre değerleri seçer [23]. Şartlı Rastgele Alanlarda olabilirlik logaritması (3.1)'deki gibidir.

$$\mathcal{L}(\lambda) = \sum_k \left[\log \frac{1}{Z(x^{(k)})} + \sum_j \lambda_j F_j(y^{(k)}, x^{(k)}) \right] \quad (3.1)$$

Burda verilen fonsiyon, global maksimum için yakınsama garanti, iç bükey fonsiyonudur [16].

Olabilirlik logaritmasının farklılaştırılması λ_j parametresi ile belirtilir;

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda_j} = E_{\tilde{p}(Y, X)}[F_j(Y, X)] - \sum_k E_{p(Y|x^{(k)}, \lambda)}[F_j(Y, x^{(k)})] \quad (3.2)$$

Eşitlik 3.2'de verilen $\tilde{p}(Y, X)$ eğitim veri setinin ampirik dağılımını, $E_p[.]$ 'de p dağılımının beklentisini gösterir. Burdaki türevin sıfır verimi için Maksimum Entropi

Model'in kısıtı; her özelliğin model dağılımı eğitim veri setinin ampirik dağılımının altındaki değere eşit olmasıdır [16].

Şartlı Rastgele Alanlarda en yüksek olabilirlik parametre değerlerini belirleyebilmek için eğitim veri seti içerisindeki her gözlem dizisinin her özellik işlevinin dağılımını hesaplamak mümkün olmalıdır. $x^{(k)}$ gözlem dizisi olmak üzere;

$$E_{p(Y|x^{(k)},\lambda)}[F_j(Y, x^{(k)})] = \sum_y p(Y = y|x^{(k)}, \lambda) F_j(y, x^{(k)}) \quad (3.3)$$

Gözlem dizisi $x^{(k)}$ 'nin sahip olduğu n eleman için $n^{|Y|}$ kadar karşılık gelen etiket dizileri vardır.

3.3.3 Özellik Çıkarımı

Varlık İsmi Tanıma'da özellik çıkarımı kelimelerin karakteristik yapısına göre yapılmaktadır. Şartlı Rastgele Alanlar'da veri setindeki kelimelerin özellikleri etiket bilgisi ile ifade edilmektedir. Örneğin, bir kelime eğer büyük harf ile başlıyor ise bu kelimenin "*büyük_harf_ile_başlama*" özelliğine 1 atanacaktır, eğer büyük harf ile başlamıyorsa bu özelliğe 0 atanacaktır. Diyelim ki doküman içerisindeki her kelime için;

- Büyük harf ile başlama. Eğer büyük harf ile başlıyorsa 1 başlamıyorsa 0 atanacak.
- Kelime içerisindeki harf sayısı
- Kelimenin kendisi

şeklinde üç özelliği "Davutpaşa Kışlası 1999 yılında İstanbul'da bulunan Yıldız Teknik Üniversitesi mülkiyetine verilmiştir." cümlesine uygulayıp, özellik vektörünü çıkartırsak aşağıdaki gibi sonuç alınacaktır.

<1, 9, Davutpaşa>, <1, 7, Kışlası>, <0, 4, 1999>, <0, 7, yılında>, <1, 8, İstanbul>, <0, 7, bulunan>, <1, 6, Yıldız>, <1, 6, Teknik>, <1, 12, Üniversitesi>, <0, 11, mülkiyetine>, <0, 11, verilmiştir>

3.3.4 Temel Özellikler

Şartlı Rastgele Alanlar yönteminde kullanılmak üzere belirlenen temel özellikler, herhangi bir doküman türüne bağlı olmaksızın çıkartılmıştır. Temel özellikler Şartlı Rastgele Alanlar'da eğitim veri setinden olasılık hesaplama için kullanılmaktadır. Dokümandaki her kelime için özelliklere bakılır. Şartlı Rastgele Alanlarda her kelime için özellik bilgileri etiket bilgisi ile belirtilmektedir.

3.3.4.1 Büyük Harf Kullanımı

Türkçe bir dokümanda kelimenin büyük harf ile başlayıp başlamamasına dikkat edilmektedir. Doküman içerisinde bir kelime büyük harf ile başlıyorsa bu özellik etiket ile belirtilmektedir.

Türkçe dili için büyük harf kullanım özelliği Türk Dil Kurumu (TDK) tarafından belirtilmiştir. Büyük harf kullanımı ile ilgili TDK kaynakları baz alınarak tanımlamalar yapılmıştır.

3.3.4.2 İlk Kelime Olması

Doküman içerisindeki bir kelime cümlenin ilk kelimesi ise, bu özellik etiketi ile belirtilmektedir. Türkçe'de cümleler büyük harf ile başladığından büyük harf ile başlama özelliğinin yanında cümlenin ilk kelimesi olması durumu da dikkate alınmıştır.

3.3.4.3 Noktalama İşaretleri

Kelimeler kesme ('), çift tırnak (" "), nokta (.), virgül (,), noktalı virgül (;), iki nokta (:) şeklindeki noktalama işaretleri ile birlikte yazılmış ise noktalama işaret kullanımı özelliği etiketlenir.

3.3.4.4 Numerik Değer İçerme

Doküman içerisindeki herhangi bir kelime sayısal bir değer içerdiği zaman bu özellik etiketi kullanılmaktadır. Bu özellik tarih, para ve bazen de kurum isimlerini belirlemede kullanılmaktadır.

3.3.4.5 Özel Kelimeler Kullanma

Türkçe’de kişi isimlerinde özel ad ile birlikte ünvan veya makam sözcükleri kullanılabilir. Kişi adlarından önce ve sonra gelen ünvanlar, saygı sözleri, rütbe adları kelime için bir nitelik olup, kimi zaman özel adın yerine de kullanılmaktadır. Bir kelimedenden önce belirtilen bir ünvan veya makam sözcüğünün olması durumunda özel kelime kullanımı özelliği etiketlenmektedir.

Kişi isimleri ile birlikte kullanılan bu özel kelimelerin bir benzerleride kurum isimleri içinde kullanılmaktadır. Ünvan ve makam kullanımına örnek olarak;

Cumhurbaşkanı Mustafa Kemal Atatürk, Sayın Doç. Dr. Banu Diri, Mimar Sinan, Ayşe Hanım, Sinan Bey verilebilir.

3.4 Kural Tabanlı Yaklaşımın Eklenmesi

Türkçe bir dokümanda Varlık İsmi Tanıma için en çok kullanılan yöntemlerden biri kural tabanlı yaklaşımdır. Kural tabanlı yaklaşımı tek başına kullanmak artık yetersiz kalmaktadır. Günümüzde doküman türüne ve içeriğine bağlı olarak geliştirilen kurallar doküman türünden bağımsız herhangi bir alanda kullanıldığında verimli sonuçlar verememektedir. Kural tabanlı yaklaşımda en büyük sorun sınıflandırma iken bu sorunu aşabilmek için çok fazla kural çıkartılması gerekmektedir. Ancak tanımlanan her kural her dokümana uyarlanamadığından beklenen sonuç elde edilememektedir.

Kural tabanlı yaklaşımda elde edilen başarılar ve gözlemlenen eksiklikler göz önüne alınarak destekleyici yöntem olarak kullanıldığında daha verimli sonuç elde edilmektedir. Varlık İsmi Tanıma’ya yönelik geliştirilen Şartlı Rastgele Alanlar yöntemi destekleyici olarak kural tabanlı yaklaşım ile zenginleştirildiğinde sonuçlar daha verimli olabilmektedir. Böylece kural tabanlının yetersiz kaldığı sınıflandırma aşamaları Şartlı Rastgele Alanlar ile aşılrken, Türkçe diline özgü günümüze kadar çıkartılan kurallar ile etiketle işleminde daha yüksek başarı elde edilmiştir.

Kural tabanlı yaklaşımda her bir varlık için o türe ait belirleyici durumlar araştırılıp kural haline getirilmiştir. Varlık İsmi Tanıma için tanımlanmış kurallar dört grup altında ele alınmıştır.

- Kişi İsimleri
- Kurum İsimleri
- Yer İsimleri
- Tarih Varlıkları

3.4.1 Kişi İsimleri

Türkçe bir dokümanda en çok bulunan varlıklar genelde kişi isimleri olmaktadır. Bir sistemi diğer sistemlerden ayıran en büyük özelliklerden biri kişi ismi belirlemedeki başarısıdır. Bu nedenle kural tabanlı yaklaşımlarda geliştirilen sistemlerde kişi ismi belirleme için daha çok kural tanımlanmaktadır. Tez çalışması için geliştirilen sistemde kişi isimlerinin tespiti için tanımlanan kurallar şöyle sıralanabilir;

- Anahtar kelimelerin sonrası ve öncesi incelenen dört kelimenin her birinin büyük harf ile başlaması. *“Sayın Prof. Dr. Hasan Eren konu hakkında açıklama yaptı”* cümlesinde *“Sayın ”* kelimesinden sonra ve önce ilk dört kelimeye bakılmaktadır. *“Sayın”* kelimesi Şartlı Rastgele Alanlar da özel kelime kullanımına girmekte olduğundan bundan sonra gelen kişi isminin özellik etiketine atanmaktadır. Ancak, bu kural ile birlikte *“Sayın”* kelimesinden sonra gelen ilk dört kelimenin büyük harf ile başlamasına bakılmaktadır.
- Kişi ismi birden fazla kelimedden oluşuyorsa bunları bir bütün olarak değerlendirebilmek için aralarında noktalama işareti olup olmadığına, *“ve”*, *“ile”*, *“la”*, *“le”* gibi bağlaçların aralarında olup olmadığına bakılarak kişi isminin bir den fazla kelimedden oluşup oluşmadığı belirlenir.
- Kişi ismi olarak etiketlenmiş bir kelimedden önce ya da sonra gelen *“ve”*, *“ile”*, *“la”*, *“le”* gibi bağlaçlardan önce ya da sonra yazılan kelime büyük harf ile başlıyorsa kişi ismi olarak etiketlenmiştir. *“Çiğdem’le Serap beraber büyüdüler”* cümlesinde *“Çiğdem”* kişi ismi olarak belirlendiyse *“Serap”* kelimesinin kişi ismi olama olasılığına bakılır. Bunun içinde büyük harf ile başlama, noktalama işaretlerinin kullanılması ve *“le”* bağlacı ile bir önceki kelime ile ilişkilendirilmesi etiketlerine bakılmaktadır.

- Doküman içerisinde ad veya soyad tek başına geçerse her birinin bulunması için önceden tespit edilen ad-soyad ikilisi dokümanın sonraki kısımlarında ayrı ayrı da bakılmaktadır. Böylece ad, soyad tek başına geçerse her birinin tespit edilmesi sağlanmaktadır.

3.4.2 Kurum İsimleri

Türkçe bir dokümanda kurum isimlerini belirlemek için yapılan Varlık İsmi Tanıma sistemlerinde ülkemizdeki kurum, organizasyon, vakıf, dernek gibi yapıların temel isimlendirilme özelliklerine bakılmaktadır. Kurum isimlerini belirlemede kural tabanlı yaklaşımlarda en önemli nokta kişi isimlerini belirlemede kullanılan kurallar ile çatışmamasıdır. Varlık İsmi Tanıma için geliştirilen sistemde kurum isimlerini belirlemek için tanımlanan kurallar;

- Türkçe’de kurum isimleri çoğunlukla kısaltmalarla ifade edilmektedir. Ancak kısaltma denilince; bir kelimenin, terimin veya özel adın, içerdiği harflerden biri veya birkaçı ile daha kısa olarak ifade edilmesi olduğundan kısaltma olarak verilen kurum isimlerinin diğer varlık türlerinden ayırt edilmesi işlemi gerekmektedir. Bunun içinde kurum isimlerini niteleyen, kurum isimlerinden sonra gelen herhangi bir özel kelime kullanılmış mı diye bakılır. Eğer kurum için tanımlanan özel kelime kullanılmış ise ve kurum isminin bütün harfleri büyük harf ile yazılmış ise bu kısaltma kurum ismi olarak etiketlenmektedir. Örneğin aşağıdaki cümlede “TOG” kurum kısaltması olarak belirtilmiştir. Bu kelimenin kurum ismi olduğunu belirtebilmek için bütün harflerin büyük harf olması, özel kelime kullanılması (“vakıf”), cümlenin ilk kelimesi olması gibi özellikler etiketlenerek kurum ismi olduğuna karar verilmiştir.

“Toplum Gönüllüleri Vakfı hizmete başlamıştır. TOG vakfı gençlerin kişisel gelişimlerine katkıda bulunmaktadır.”

- Kurum isimleri yazılırken çoğunlukla parantez içerisinde kurum ismi kısaltması ya da kurumun açık ismi yazılmaktadır. Bu yazım özelliği dikkate alındığında kurum ismi tespitinde, büyük harf ile başlayan kelimelerden sonra parantez kullanma özelliğine bakılmaktadır. Örneğin;

“Türk Dil Kurumu (TDK) Türk dilinin güncel sorunlarıyla ilgilenererek çözüm yolları araştırmaktadır.”

- Kurumun açık ismi bazen “ve” bağlacı ile bağlanmaktadır. Bir anahtar kelimedenden önce gelen kelimeler içerisinde “ve” bağlacı varsa ve bu kelimeler büyük harfler ile başlamış ise, kelimeler kurum ismi olarak ele alınmaktadır.

“Bankacılık Düzenleme ve Denetleme Kurumu”

3.4.3 Yer İsimleri

Türkçe bir dokümanda yer isimleri kişi isimleri kadar çok sık kullanılmamaktadır. Varlık İsmi Tanıma için yapılan çalışmalarda genellikle yer isimleri anahtar kelimelerle tespit edilmektedir. Geliştirilen sistemde yer isimlerini belirlemek için tanımlanan kurallar;

- Yer isimlerinde kullanılan -de, -da, -den, -dan, -te, -ta hal ekleri yer isimlerini belirlemede kullanılabilir. Bu hal eklerini alan bir kelime büyük harf ile başlıyorsa bu kelime yer ismi olarak etiketlenmektedir.
- Yer isimleri bazen birleşik isim olabilmektedir. Birleşik isim olan yer isimlerini belirlerken deniz, doğu, köy gibi sistemde tanımlanan özel kelimeler içeriyorsa ve büyük harf ile kelime başlıyorsa yer ismi olarak etiketlenmektedir. Örneğin;

“Karadeniz ve Ortadoğu bölgelerinden yakın ilişkiler içinde olan ülkeler.”

- Türkçe’de bazen ülke isimlerini belirtmek için kullanılan “-istan, -ye, -ya” heceleri yer isimlerini belirlemede kullanılmaktadır. Örneğin;

“Özbekistan, Kırgızistan, Rusya, Avusturya, Malatya”

BÖLÜM 4

GELİŞTİRİLEN SİSTEM

Varlık İsmi Tanıma'ya yönelik oluşturulan sistemde resmi olmayan bir dilde yazılmış, belli bir çalışma alanından bağımsız olan tüm Türkçe dokümanlar içerisinde, varlık isimlerini bulmak ve bunları kişi, kurum ve yer ismi olarak sınıflandırmak amaçlanmıştır.

Varlık İsmi Tanıma için geliştirilen sistem makine öğrenme tekniklerinden Şartlı Rastgele Alanlar yöntemi esas alınarak oluşturulmuştur.

Sistem ilk önce yarı yapılandırılmış doküman formatına sahip olan, Türkçe e-postalar için tasarlanmıştır. Ancak, daha sonrasında doküman formatından ve içeriğinden bağımsız yapılandırılmamış tüm dokümanlar hedef alınarak gelişmiş bir sistem tasarlanmıştır.

Bu bölümde ilk önce; Türkçe e-postalar içerisinde Şartlı Rastgele Alanlar kullanılarak varlık isimlerinin bulunması için geliştirilen sistem anlatılacaktır. İkinci aşamada da doküman türü ve alanından bağımsız dokümanlar için geliştirilen genişletilmiş sistem tanıtılacaktır.

4.1 Veri Seti Kullanımı

Varlık İsmi Tanıma için geliştirilen sistemde veri seti oluşturulması iki aşamada yapılmıştır. İlk çalışmada e-postalar için veri seti oluşturulmuştur. Sonraki aşamada doküman türünden ve alanından bağımsız dokümanlar için veri seti oluşturulmuştur.

Türkçe e-postalarda özel anlamlı varlıkları tespit edebilmek için yapılan çalışma üç farklı alanda toplanan e-postalar üzerinde uygulanmıştır. Akademik alanda staj yapmak

isteyen öğrencilerden toplanan 50 adet e-posta, kurumsal şirketlerde yapılan yazışmalardan elde edilen 50 adet e-posta ve kişisel yazışmalardan elde edilen 50 adet e-posta olmak üzere toplam 150 adet e-postadan oluşturulan bir veri seti ile çalıştırılmıştır.

Çizelge 4.1 E-posta veri seti özellikleri

E-posta Veri Seti Kaynağı	Toplam Adet Sayısı	Toplam Kelime Sayısı
Akademik E-postalar	50	6,320
Kurumsal E-postalar	50	3,762
Kişisel E-postalar	50	8,750

Doküman türüne ve alanına bağlı olmadan geliştirilen Varlık İsmi Tanıma sistemini eğitebilmek için eğitim veri seti hazırlanmıştır. Eğitim veri seti doküman türüne bağlı olmayıp resmi olmayan bir dilde yazılmış dokümanlar için yapılmıştır. Oluşturulan eğitim veri setinde kişi, kurum ve yer ismi Şekil 4.1’de gösterildiği gibi MUC-6 ve MUC-7 standartlarına uygun bir şekilde etiketlenmiştir.

```

1 Su tek defnenin altına.
2 <ENAMEX TYPE="LOCATION">Kargacak yolu</ENAMEX>na mı?
3 Yok canım, <ENAMEX TYPE="LOCATION">Balık Kayaası</ENAMEX>'ndan yana. Onlar değil mi? Toplamışlar bu kez salyangozları?
4 Herhalde öyle olacak. Bizi atlatmak için hep. Ulan <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX>, dusersin elime sen.
5 Bir süre konuşmadan durdular. Daha doğrusu konuşacak halleri kalmamıştı. Bir süre daha çit çıkarmadan beklediler. Ağızlarını bıçak açmıyordu.
6 Tam önlerinden geçerken <ENAMEX TYPE="PERSON">Vural</ENAMEX>, kesti yollarını: Hey! dedi. Durun orada. Kimildamayın!
7 İkisi birden atladılar defnenin alt dallarından.
8 Ne var sepetlerinizde? Salyangoz mu?
9 Ne salyangoz? dedi <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX>. Düş mü görüyorsun sen? .
10 Düşüm mu gösterirsin sana. Ne var diyorum sepetlerinizde?
11 <ENAMEX TYPE="LOCATION">Balık Kayaası</ENAMEX>'ndan gelenlerin sepetinde ne olabilir, dedi <ENAMEX TYPE="PERSON">Tekin</ENAMEX>.
12 Gerişim <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX> getirdi: Evet, dedi. <ENAMEX TYPE="LOCATION">Balık Kayaası</ENAMEX>'ndan gelenlerin sepetinde k
13 da mezzit.
14 Üzerine örtülen defne yapıklarını eliyle
15 topladı.
16 <ENAMEX TYPE="PERSON">İbrahim</ENAMEX>, bir ara saskinlıkla, Demek salyangoz yokmuş, dedi.
17 Ne salyangozu be, dedi <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX>. Salyangoz da nereden çıktı?
18 <TIMEX TYPE="DATE">Dün</TIMEX> sen Bayan <ENAMEX TYPE="PERSON">Mari</ENAMEX>'ye ne verdin, salyangoz vermedin mi?
19 Hadi ordan sen de. Çitir çitir ince kabuklu ceviz verdin, yoldan dönüp geldim de. Öyle değil mi <ENAMEX TYPE="PERSON">Tekin</ENAMEX>?
20 Pekii, dedi <ENAMEX TYPE="PERSON">İbrahim</ENAMEX>, Öyleyse <TIMEX TYPE="DATE">bugün</TIMEX> neden getirmedin?
21 İnsan her gün baklava börek yese bıkar be. <TIMEX TYPE="DATE">Bugün</TIMEX> de istavrit isterim diye tutturdu. Biz de kıramadık, indik <ENAMEX TY
22 Bu konuşmaları duyan <ENAMEX TYPE="PERSON">Lui</ENAMEX>'yle <ENAMEX TYPE="PERSON">Pol</ENAMEX> de kosarak gelmişlerdi. Balık sepetlerini görünce,
23 Bayan <ENAMEX TYPE="PERSON">Mari</ENAMEX> sesleniyordu uzaktan onlara: Heeey! <ENAMEX TYPE="PERSON">Tekin</ENAMEX>! <ENAMEX TYPE="PERSON">Bahri</
24 <ENAMEX TYPE="PERSON">Tekin</ENAMEX>'le <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX> <ENAMEX TYPE="LOCATION">Çandır Köyü</ENAMEX>'nın altındaki kumsalda
25 <ENAMEX TYPE="PERSON">Tekin</ENAMEX>, Bir dalıp çıkalım istersen, dedi
26 arkadaşlarına.
27 <ENAMEX TYPE="PERSON">Bacaksız</ENAMEX>'in aklı başka yerlerdeydi. <ENAMEX TYPE="PERSON">Jülyet</ENAMEX> çok iyi kız, dedi. Beni <ENAMEX TYPE="
28 Fransızca bilmiyorsun ki. Ne isin var
29 <ENAMEX TYPE="LOCATION">Paris</ENAMEX>'te?
30 Babam Almanca biliyor mu? Orada
31 çalısıp para kazanıyor.

```

Şekil 4.1 Etiketlenmiş eğitim veri seti

Eğitim veri seti oluşturulurken ilk olarak daha önceden yapılan çalışmalarda kullanılan eğitim veri setleri ilk önce toplanmıştır. Daha sonra elle farklı alanlardan derlenen

dokümanlar eğitim veri setine eklenmiştir. Sistemi eğitebilmek için oluşturulan bu eğitim veri setindeki özel anlamlı varlık sayıları Çizelge 4.2’de gösterilmiştir.

Çizelge 4.2 Eğitim veri seti özellikleri

Eğitim Veri Seti Kaynağı	Toplam Kelime Sayısı	Toplam Kişi İsmi Sayısı	Toplam Kurum İsmi Sayısı	Toplam Yer İsmi Sayısı
Diğer Çalışmalardan Derlenen Eğitim Veri Seti [10] (Tur et al., 2003)	530,031	24,101	16,105	13,540
Manuel Oluşturulan Eğitim Veri Seti	406,984	6,280	2,817	3,150

4.2 Türkçe E-Posta için Geliştirilen Sistem

E-postaları diğer doküman türlerinden ayıran en büyük özellik kendilerine özgü bir formata sahip olmalarıdır. Metin içeriğinin anlaşılabilmesi için iyi tanımlanmış bir başlık bilgisi, e-postanın kimden geldiği ya da kime gönderileceği bilgilerinin belirtildiği alanlar, e-posta iletilisinin sonunda kurum ya da kişi isminin belirtildiği imza kısmı bir e-posta iletilisinin temel özellikleridir. Ancak, bunlara rağmen Türkçe’de Varlık İsmi Tanıma için yapılan çalışmalar arasında e-postalar üzerine yapılan çalışmalar nerdeyse bulunmamaktadır.

4.2.1 Kelime Nesnesi Oluşturma

Varlık İsmi Tanıma için geliştirilen sistem Java programlama dili ve özellikleri kullanılarak geliştirilmiştir. Varlık İsmi Tanıma ve nesne tabanlı yaklaşımların özellikleri düşünülerek e-posta içerisindeki her kelime bir Java nesnesi olarak sistemde oluşturulmuştur. Her nesne kelimeye özgü özellikleri içerisinde barındırmaktadır.

4.2.2 N-Gram Çıkarımı

Varlık İsmi Tanıma için geliştirilen sistemde bazı son ekleri belirleyebilmek için n-gram çıkarımına ihtiyaç duyulmuştur. N-gram, bir karakter katarının n adet karakter dilimidir [17]. Geliştirilen sistemde özellik çıkarımında kullanmak üzere n-gram'ın farklı birkaç uzunluğu olarak 1-gram, 2-gram, 3-gram'lar çıkartılmıştır.

4.2.3 E-Posta Özellikler Çıkartılması

E-postaların yapısına özgü özelliklere bakılarak bir kelimenin temel özelliklerinin bir kısmı bu aşamada belirlenir. E-posta başlık bilgisi ile başladığından bir çok bilgi bu alanda bulunmaktadır. *From, To, Cc, Bcc, Send, Sender, In-Reply-to, Followup-to, Reply-to, Forwarded by* alanlarına bakılmaktadır. Ayrıca, e-posta konu başlık kısmında da özel anlamlı kelimeler olabileceğinden bu alana da bakılmaktadır.

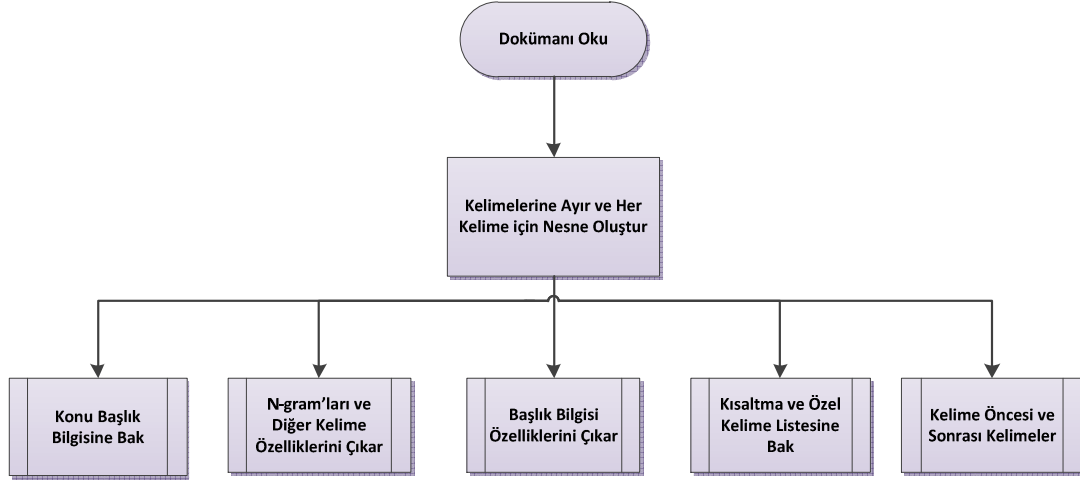
4.2.4 Etiketlenebilen Özelliklerin Çıkartılması

E-posta için geliştirilen sistemde Şartlı Rastgele Alanlar kullanılarak olasılığın belirlenebilmesi için belirlenen özelliklere bakılmaktadır. Kelimenin sahip olduğu her özellik kelime için oluşturulan Java nesnesinde etiketlenerek tutulmaktadır. Buna göre kelime nesnesinde tutulan özellikler;

- Kelimenin kendisi
- Kelimenin uzunluğu
- Kelimenin geçme sıklığı
- Kelimenin kısaltma olup olmadığı
- Cümlelerin ilk kelimesi olup olmadığı
- Kelimenin büyük harf ile başlaması
- Kelimenin 1-gram, 2-gram, 3-gram'ları
- Başlık bilgisinde yer alması (From, To, Cc, Bcc, Send, Sender, vd.)
- Konu başlık bilgisinde yer alması
- Büyük harf ile başlaması

- Noktalama işareti ile birlikte yazılması
- Ünvan ile başlaması
- Özel kelime listesindeki kelimelerin biriyle başlaması
- Kelimeden önce ve sonra gelen kelimeler

Temel Özelliklerin belirlenmesi için yapılan işlemler Şekil 4.2’de kısaca belirtilmiştir.

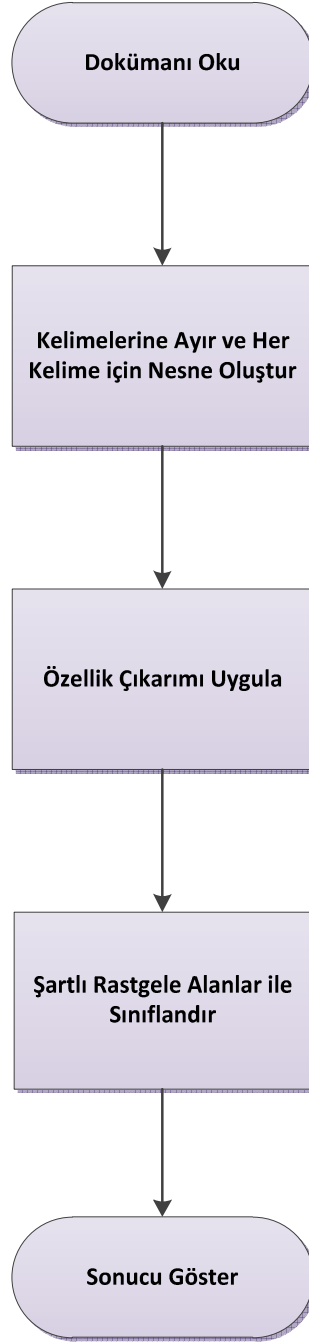


Şekil 4.2 Özellik çıkarımı

4.2.5 Sınıflandırma

Türkçe e-posta için geliştirilen sistemde yukarıda belirlenen özellik çıkarımı yapıldıktan sonra etiketleme işlemine geçilir. Veri setinden çıkartılan özellikler kullanılarak sınıflandırma için optimizasyon parametreleri belirlenir. Şartlı Rastgele Alanlar yöntemi kullanılarak kelime için etiket sınıfı bulunur.

Şekil 4.3’de gösterildiği gibi özellik çıkarımında elde edilen değerler Şartlı Rastgele Alanlar da olasılık hesabı ve etiket sınıfını belirleme için kullanılmaktadır. Bir kelimeye ait ne kadar çok özellik etiketlenebilirse Şartlı Rastgele Alanlar yöntemiyle ait olduğu özel anlamlı varlık sınıfını belirleme daha sağlıklı olmaktadır.

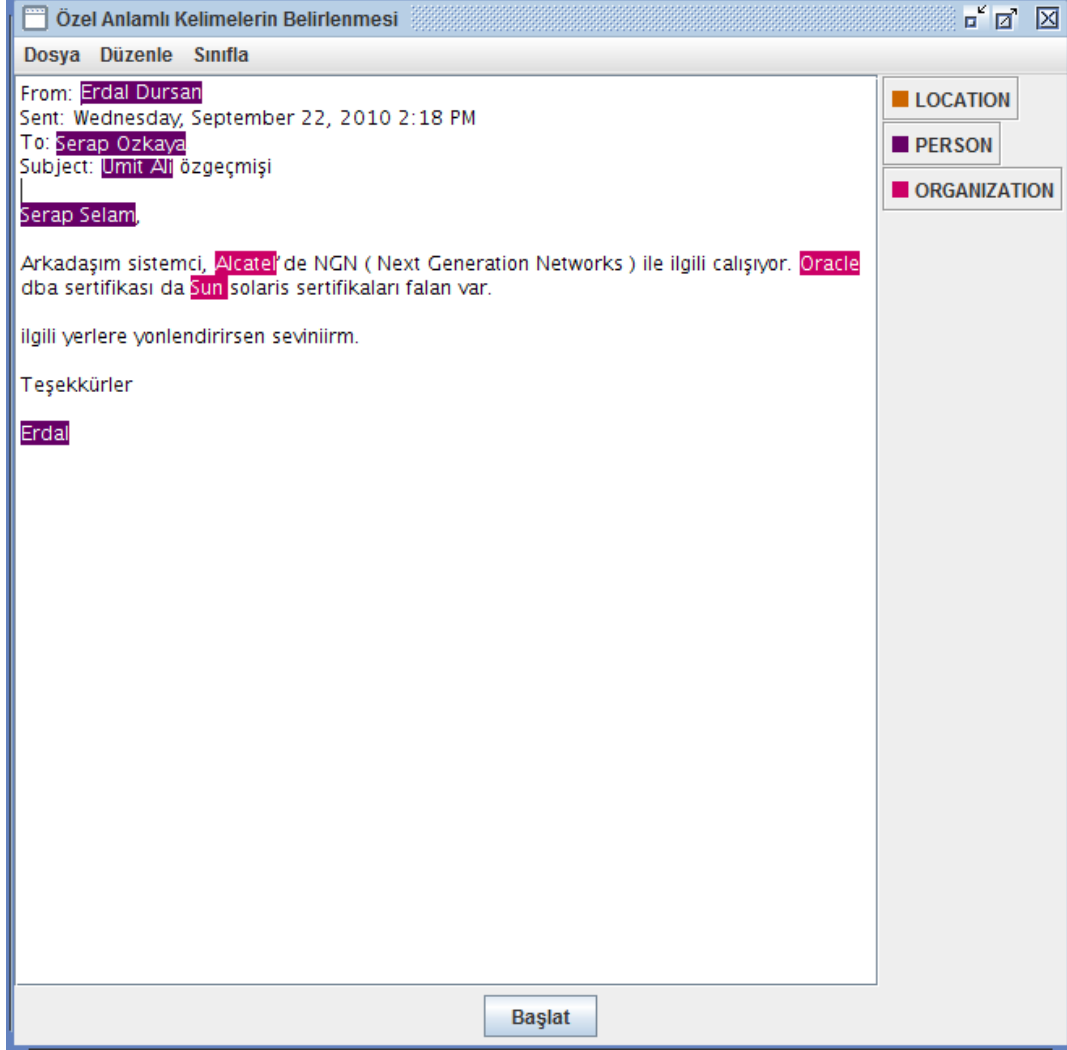


Şekil 4.3 Uygulama akışı

4.2.6 Uygulamanın Çalıştırılması

Varlık İsmi Tanıma da Türkçe e-posta için geliştirilen uygulamada Şekil 4.3 de gösterilen akış uygulandığında herhangi bir Türkçe e-posta için kişi, kurum ve yer isimleri

belirlenmektedir. Uygulamanın çalıştırılmasına örnek olarak Şekil 4.4’de verilmiştir. Uygulama çalıştırıldığı zaman çıkan kullanıcı arayüz ekranında özel anlamlı varlıklar renk kodları ile gösterilmektedir.



Şekil 4.4 Uygulamanın çalıştırılması

4.3 Yapılandırılmamış Dokümanlar için Geliştirilen Sistem

Varlık İsmi Tanıma için doküman türünden ve alanından bağımsız olarak sistem yeniden tasarlanıp, yeni bir uygulama geliştirilmiştir. Türkçe e-posta için yapılan çalışmada makine öğrenmesi tekniklerinden Şartlı Rastgele Alanlar kullanılsa da doküman yapısına bağlı olarak uygulama hazırlanmıştır. Bu çalışmada öğrenilen bilgi ve tecrübe ile daha gelişmiş bir sistem hazırlanmıştır. Yapılandırılmamış Türkçe bir dokümanda

Varlık İsmi Tanıma için Şartlı Rastgele Alanlar kural tabanlı yaklaşım ile zenginleştirilerek kullanılmıştır.

Ayrıca, sistem tasarlanırken Şartlı Rastgele Alanlar yarı eğitici makine öğrenmesi tekniği ile uygulanmıştır. Bu yöntem ile istatistiksel bir model olan Şartlı Rastgele Alanlar'da olasılık hesabı daha başarılı belirlenmiştir.

Türkçe e-posta çalışmasında olduğu gibi Şartlı Rastgele Alanlar'da kullanılmak üzere özellik çıkarımları gerçekleştirilmiştir. Elde edilen değerler kural tabanlı yaklaşım ile çıkartılan kurallar ile genişletilmiştir.

4.3.1 Sistemin Eğitilmesi

Şartlı Rastgele Alanları yarı eğitici makine öğrenmesi tekniği ile uyguladığımızdan test veri setinden önce sistemi eğitmek gerekmektedir. Sistemi eğitebilmek için daha önceden hazırlanan eğitim veri seti ile sistem eğitilmiştir.

Çizelge 4.2 de verilen eğitim veri seti değerleri ile sisteme kişi, kurum ve yer isimleri bilgileri tanıtılmıştır. Sisteme tanıtılan eğitim veri seti miktarına bağlı olarak Şartlı Rastgele Alanlar yöntemiyle belirlenen olasılık değerleri değişkenlik göstermektedir. Bu nedenle eğitim veri seti ne kadar doğru ve miktarı ne kadar çok olursa sistemin başarı oranı o ölçüde etkilenmektedir.

4.3.2 Özel Kelime Kullanılması

Özel anlamlı varlıkları belirlemede kullanabilmek için daha önceden belirlenen özel kelimeler listesi, eğitim veri setinden sonra sisteme okutulmaktadır. Kişi, kurum ve yer isimleri için ayrı özel kelime listesi oluşturulmuş olup, bu veriler sisteme bir kez okutulmaktadır. Bir kelimedenden önce veya sonra belirtilen ünvan, makam ya da kısaltma özel kelime listesinde ise kelimenin özel kelime kullanımı özelliği etiketlenmektedir.

4.3.3 Özellik Çıkarımı

Özellik çıkarımı geliştirilen sistemin temel noktalarından biridir. Kelimenin karakteristik yapısına göre çıkartılan özellikler Şartlı Rastgele Alanlar'da kullanabilmek için

etiketleme yöntemi ile belirtilmektedir. Yapılandırılmamış dokümanlar için tanımlanan temel özellikler başlıca;

- Kelimenin kendisi
- Kelimenin geçme sıklığı
- Büyük harf ile başlaması
- Kelime uzunluğu
- Kelimenin kısaltma olup olmadığı
- Cümlelerin ilk kelimesi olup olmadığı
- Noktalama İşareti ile birlikte yazılması
- Numerik değer içermesi
- Özel kelime listesindeki kelimelerin biriyle yazılması
- Kelimeden önce ve sonra gelen kelimeler

4.3.4 Şartlı Rastgele Alanların Kullanımı

Varlık İsmi Tanıma için geliştirilen sistemde Şartlı Rastgele Alanlar yöntemi ile örüntü tanıma için olasılık değerleri belirlenmektedir. Olasılık değerlerini belirleyebilmek için ihtiyaç duyulan etiket dizisi eğitim veri setinden elde edilmektedir. Tanımlanan özellik işlevlerinin her biri etiketlenmiş eğitim veri setindeki değerlere göre belirlenmektedir. Ayrıca, Şartlı Rastgele Alanlar yönteminin uygulanması için [17]'de verilen kütüphane özellikleri kullanılmıştır.

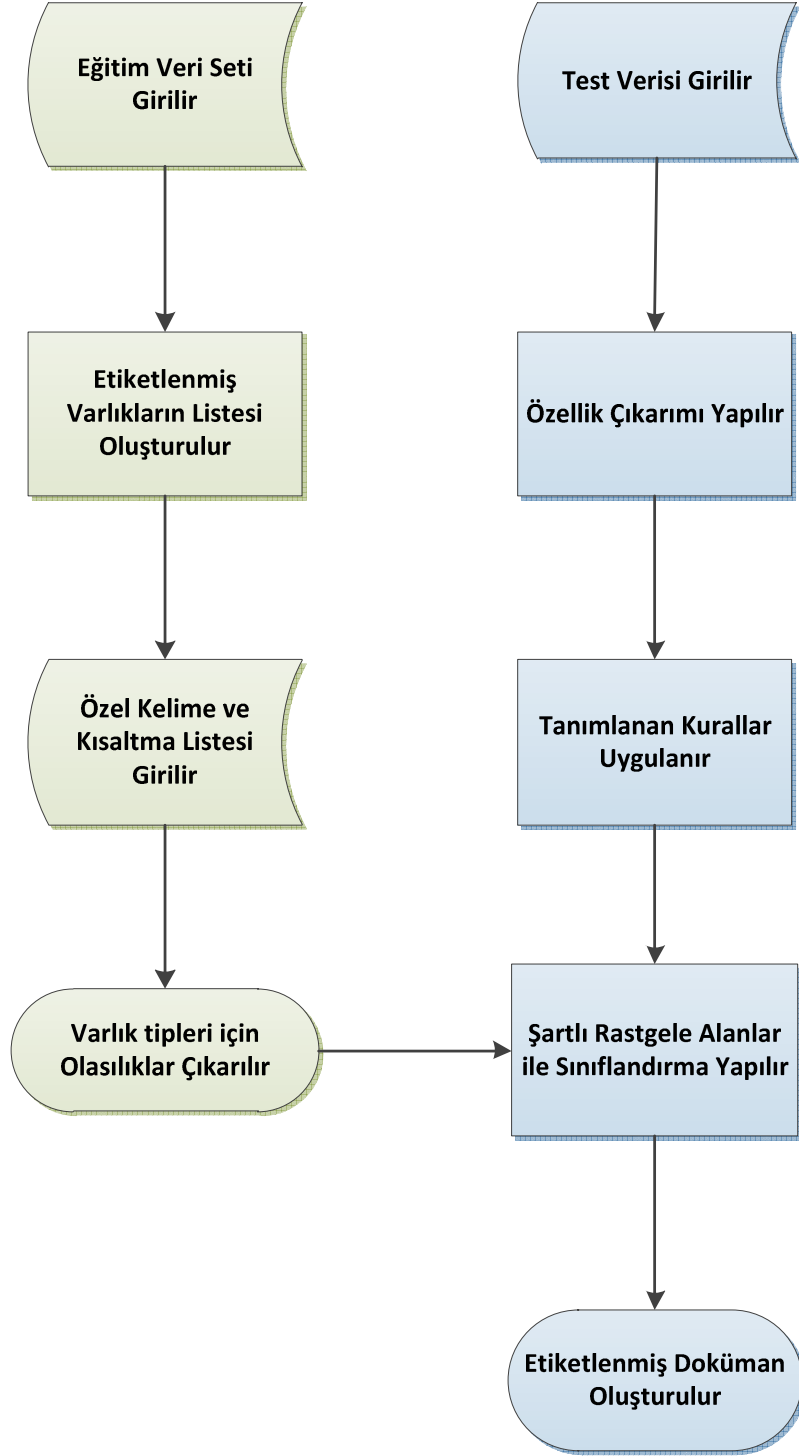
4.3.5 Kural Tabanlı Yaklaşım

Varlık İsmi Tanıma için geliştirilen sistemde Şartlı Rastgele Alanlar yöntemi ile kelimenin ait olabileceği varlık sınıfı, olasılık üzerine belirlenmektedir. Şartlı Rastgele Alanlar yöntemine ek olarak kural tabanlı yaklaşım yöntemi ile sistem genişletilmiştir. Kişi, kurum ve yer ismi için ayrı ayrı tanımlanan kurallar özellik çıkarımı dışında uygulanmaktadır. Kural tabanlı yaklaşım geliştirilen sisteme sonradan eklenilmiş olup, deneysel sonuçlarda karşılaştırmalı olarak etkisi belirtilmiştir.

4.3.6 Uygulamanın Çalıştırılması

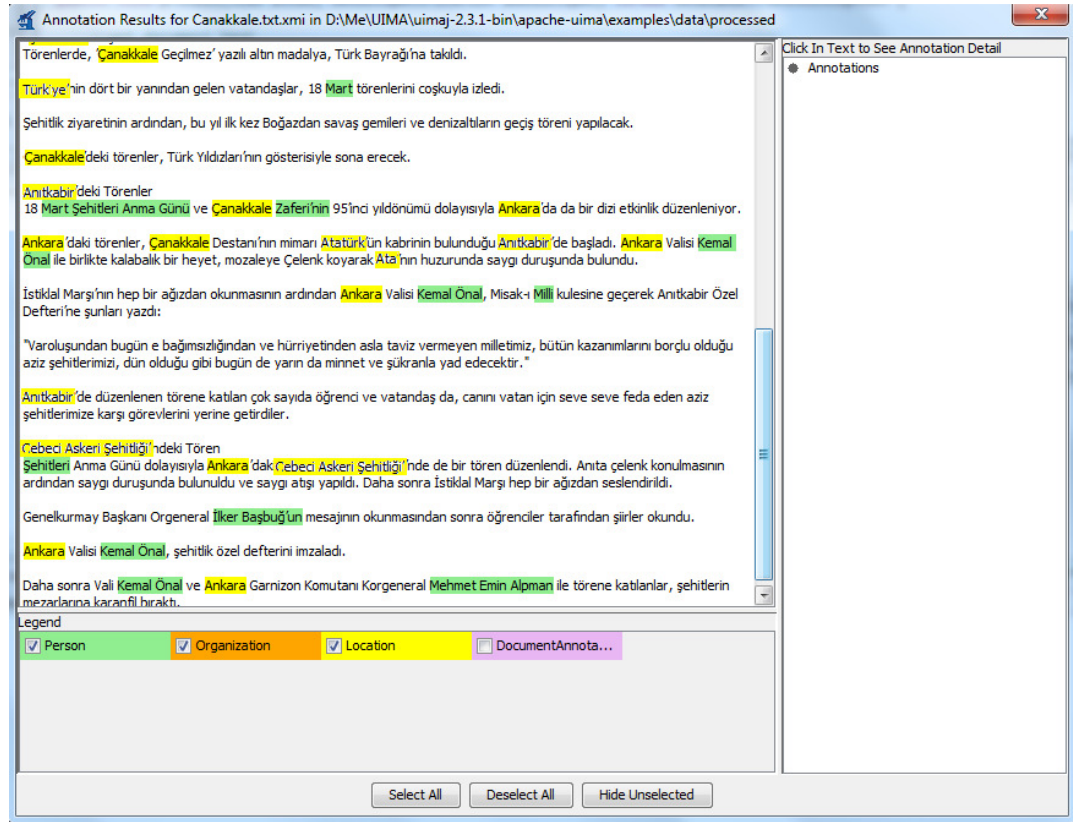
Geliştirilen sisteme eğitim veri seti ile özel kelime listesi tanıtıldıktan sonra varlık sınıflarını belirleyebilmek için eğitim veri setinden olasılık hesabı yapılmaktadır. Böylece sistem herhangi bir dokümanı işleyip, varlık isimlerini belirleyebilmek için hazır olmuş olacaktır. Test verisi girilip varlık isimleri belirlenmek istenildiğinde; ilk adım olarak sistemde tanımlanmış özellik çıkarımı test verisi için uygulanır. Test verisinin sahip olduğu özellikler etiketlenerek saklanır. İkinci adım olarak kural tabanlı yaklaşım ile belirlenen kuralların özelliklerine bakılır. En son adım olarak etiketlenmiş özellikler kullanılarak Şartlı Rastgele Alanlar yöntemi uygulanır. Bu aşamada eğitim veri setinden gözlemlenen olasılık değerleri kullanılmaktadır. Bütün bu işlemler sonucunda test verisinin sahip olduğu varlık isimleri etiketlenerek belirtilir.

Şekil 4.5’de doküman türü ve alanından bağımsız olarak geliştirilen Varlık İsmi Tanıma uygulamasının akışı anlatılmaktadır.



Şekil 4.5 Yapılandırılmamış dokümanlar için Varlık İsmi Tanıma

Yapılandırılmamış dokümanlar için geliştirilen uygulamanın çalıştırılmasına örnek olarak Şekil 4.6 verilmiştir. Uygulama tarafından tespit edilen özel anlamlı varlıklar kullanıcı arayüz ekranında özel renk kodları ile gösterilmektedir.



Şekil 4. 6 Uygulamanın çalıştırılması

DENEYSEL SONUÇLARIN DEĞERLENDİRİLMESİ

Varlık İsmi Tanıma için gerçekleştirilen çalışmanın temel amacı varlıkların bulunup tanınması ve bu varlıkların doğru olarak sınıflandırılmasıdır. Günümüzde Türkçe için yapılan çalışmaların çoğunda kullanılan yerel dilbilgisi yaklaşımı veya kural tabanlı yaklaşım ile elde edilen başarı oranları beklenen başarıyı sağlamamaktadır. Varlık İsmi Tanıma’da kullanılabilen Şartlı Rastgele Alanlar, yapılandırılmış veya yapılandırılmamış doküman türleri için kullanılabilme özelliğinden dolayı çok daha yüksek başarı oranı sunmaktadır. Temelinde istatistiksel yöntemler olan Şartlı Rastgele Alanlar eğitici ya da yarı eğitici yöntemler ile birlikte kullanıldığında verimli sonuçlar vermektedir.

Bu çalışmada Şartlı Rastgele Alanlar yarı eğitici makine öğrenmesi tekniğiyle kullanılıp, yarı yapılandırılmış ve yapılandırılmamış dokümanlar için ayrı ayrı çalıştırılmış ve sonuçları değerlendirilmiştir.

Ayrıca, çalışmanın son aşamasında yapılandırılmamış dokümanlar için sistem kural tabanlı yaklaşım ile desteklenmiştir. Bu yöntem ile elde edilen sonuçlar da ayrı şekilde değerlendirilmiştir.

Geliştirilen sistemin başarısını ölçümlemek için standart F-ölçüm yöntemi kullanılmıştır.

F-ölçüm (F-measure) Eşitlik 5.1’deki Tutturma (Precision) ve Eşitlik 5.2’deki Bulma (Recall) değerlerinin harmonik ortalamasıdır [18].

$$\text{Tutturma (P)} = \frac{\text{Doğru tespit edilen varlık sayısı}}{\text{Tespit edilen varlık sayısı}} \quad (5.1)$$

$$\text{Bulma (R)} = \frac{\text{Doğru tespit edilen varlık sayısı}}{\text{Test kümesindeki toplam varlık sayısı}} \quad (5.2)$$

Tutturma ve Bulma değerleri ile F-ölçüm değerinin belirlenmesini sağlayan formül [18], [24] ise;

$$\text{F-Ölçüm} = \frac{2 * P * R}{P + R} \quad (5.3)$$

5.1 Test Veri Seti

Yarı-yapılandırılmış doküman türü olarak Türkçe e-postalar üzerinde çalışma yapılmış olup geliştirilen sistemin başarısını belirleyebilmek için resmi olmayan Türkçe akademik, kurumsal ve kişisel e-postalardan veri seti oluşturulmuştur. E-postalar ortalama 158 kelime uzunluğunda olup, veri seti akademik, kurumsal ve kişisel e-postalardan eşit sayıda toplanmış 150 adet e-postadan oluşmaktadır.

Yapılandırılmamış dokümanlar için geliştirilen Varlık İsmi Tanıma sisteminde doküman alanına ve türüne bağlı olmaksızın internet ortamından toplanmış kaynaklar kullanılmıştır. Bu kaynakların sayılarından ziyade uzunlukları esas alınıp, içerisinde bulunan kelime sayısına bakılmıştır. Yapılandırılmamış dokümanlar da sistemin başarısını ölçebilmek için toplamda 50 sayfalık 23,369 kelimeye sahip test veri seti kullanılmıştır.

Çizelge 5.1 Test veri seti özellikleri

Test Veri Seti Tipi	Toplam Adet	Toplam Kelime Sayısı
Yarı-yapılandırılmış Dokümanlar	150 E-Posta	18,832
Yapılandırılmamış Dokümanlar	50 Sayfa	23,369

5.2 Yarı-yapılandırılmış Dokümanlar için Deneysel Sonuçlar

Türkçe e-posta için yapılan çalışma akademik, kurumsal ve kişisel olmak üzere üç farklı alanın her biri için uygulanmış olup, başarı oranları bu alanlara göre ayrı ayrı ölçülmüştür.

Çizelge 5.2’de üç farklı doküman türünde yer alan toplam metin sayıları, her alan içerisindeki isim, yer ve kurum varlık sayıları, sistem tarafından tespit edilen varlık ismi sayısı, sistem tarafından doğru tespit edilen varlık ismi sayısı ve sistemden bağımsız elle yapılan varlık ismi tespitleri gösterilmektedir. Uygulamanın doğru tespit sayısı, uygulama tarafından tespit edilen varlık ismi sayılarından hatalı tespitler çıkartılarak oluşturulmuştur.

Çizelge 5.2 Tespit edilen varlık sayıları
(UT, Uygulamanın Tespiti, DT, Uygulamanın Doğru Tespiti, GT, Gerçek Tespit)

E-Posta Türü	Akademik UT/DT/GT	Kurumsal UT/DT/GT	Kişisel UT/DT/GT
Kişi İsmi	159/146/150	175/172/180	160/141/145
Yer İsmi	32/17/18	54/35/35	36/22/26
Kurum İsmi	37/27/28	47/37/41	17/14/14

E-posta veri seti için Çizelge 5.2’de elde edilen sonuçlara göre F-ölçüm değeri hesaplanmıştır. Akademik, kurumsal ve kişisel e-postalar için ayrı ayrı kişi, yer ve kurum varlık ismi belirleme F-ölçüm değerleri sırasıyla Çizelge 5.3, Çizelge 5.4 ve Çizelge 5.5’de gösterilmiştir.

Çizelge 5.3 Akademik E-Posta F-ölçüm değerleri

Akademik E-Posta Değerleri	Tutturma (P)	Bulma(R)	F-Ölçüm
Kişi İsmi	0.91	0.97	0.94
Yer İsmi	0.53	0.94	0.68
Kurum İsmi	0.73	0.96	0.83

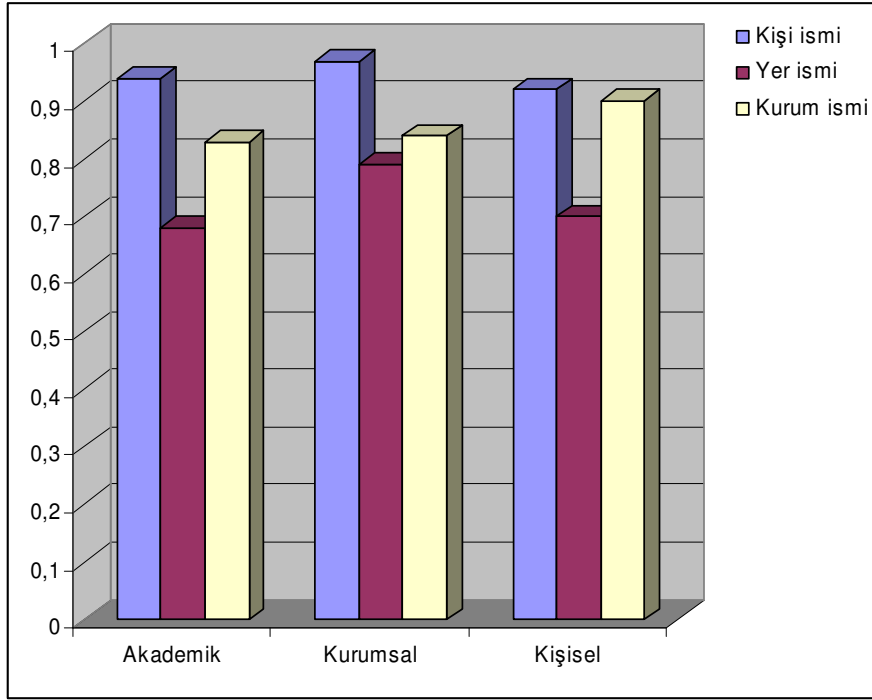
Çizelge 5.4 Kurumsal E-Posta F-ölçüm değerleri

Kurumsal E-Posta Değerleri	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.98	0.95	0.97
Yer İsmi	0.64	1.0	0.78
Kurum İsmi	0.78	0.90	0.84

Çizelge 5.5 Kişisel E-Posta F-ölçüm değerleri

Kişisel E-Posta Değerleri	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.88	0.97	0.92
Yer İsmi	0.61	0.84	0.70
Kurum İsmi	0.82	1.0	0.90

Şekil 5.1’de üç farklı e-posta türünde kişi, yer ve kurum varlık isimlerinin tanınmadaki başarısı F-ölçüm cinsinden verilmektedir. Geliştirilen kurallar ile en başarılı şekilde bulunan varlık kişi ismi olmuştur. İkinci sırada ise kurum ismi gelmektedir. En düşük performans yer isimleri olmuştur.



Şekil 5.1 Varlık isimlerinin farklı alanlardaki F-ölçüm değerleri

5.3 Yapılandırılmamış Dokümanlar için Deneysel Sonuçlar

Yapılandırılmamış dokümanlar için geliştirilen Varlık İsmi Tanıma sistemi, doküman türü ve alanından bağımsız olarak geliştirildiğinden Türkçe e-postalarda olduğu gibi ayrı türlerde hazırlanmış veri seti bulunmamaktadır.

Geliştirilen Varlık İsmi Tanıma sisteminde eğitim ve test için kullanılan veri seti, alandan bağımsız olarak işlenmiştir.

Geliştirilen sisteme daha sonrasında kural tabanlı yaklaşım eklendiğinden deneysel sonuçlar adım adım çıkartılmıştır. İlk aşamada, geliştirilen sistemde sadece Şartlı Rastgele Alanlar kullanılarak varlıkların belirlenmesi işlemi ölçümlenmiştir. İkinci aşamada ise Şartlı Rastgele Alanlar, kural tabanlı yaklaşım ile birlikte uygulandığında varlıkların belirlenmesi ayrı ölçümlenmiştir.

Kullanılan iki yaklaşım ile tespit edilen varlık sayıları Çizelge 5.6'da gösterilmiştir. Test veri seti içerisinde bulunan yer, kurum ve isim varlık sayıları, sistem tarafından tespit edilen varlık ismi sayısı, sistem tarafından doğru tespit edilen varlık ismi sayısı ve sistemden bağımsız elle yapılan etiketleme ile varlık ismi tespitleri gösterilmektedir.

Çizelge 5.6 Yapılandırılmamış dokümanlarda tespit edilen varlık sayıları
(UT, Uygulamanın Tespiti, DT, Uygulamanın Doğru Tespiti, GT, Gerçek Tespit)

Uygulama Türü	Şartlı Rastgele Alanlar UT/DT/GT	Şartlı Rastgele Alanlar + Kural Tabanlı Yaklaşım UT/DT/GT
Kişi İsmi	568/501/531	556/509/531
Yer İsmi	289/234/269	289/241/269
Kurum İsmi	129/104/135	135/111/135

Şartlı Rastgele Alanlar yarı eğitimci makine öğrenmesi tekniği ile uygulandığında elde edilen F-ölçüm değerleri kişi, yer ve kurum ismi bazında Çizelge 5.7’de verilmiştir.

Çizelge 5.7 Şartlı Rastgele Alanlar kullanıldığında F-ölçüm değerleri

Şartlı Rastgele Alanlar Kullanımı	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.88	0.94	0.91
Yer İsmi	0.80	0.86	0.83
Kurum İsmi	0.81	0.77	0.78

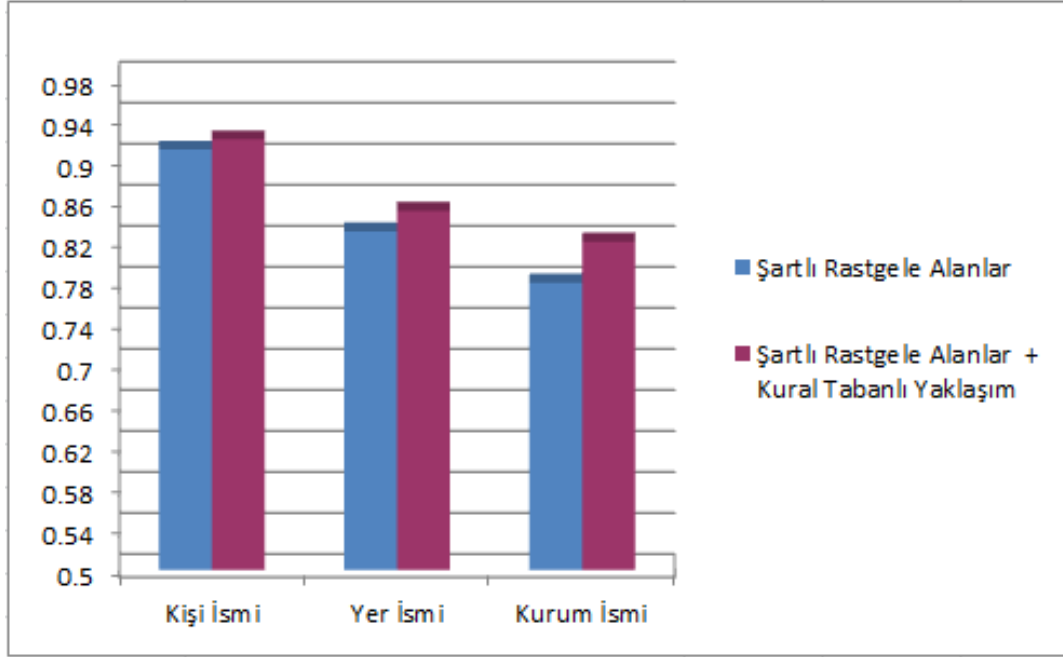
Şartlı Rastgele Alanlar yöntemi kural tabanlı yaklaşım ile desteklendiğinde elde edilen F-ölçüm değerlerinin kişi, yer ve kurum ismi bazında sonuçları ise Çizelge 5.8’de gösterilmiştir.

Çizelge 5.8 Kural tabanlı yaklaşım eklenildiğinde F-ölçüm değerleri

Şartlı Rastgele Alanlar + Kural Tabanlı Yaklaşım Kullanımı	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.91	0.95	0.92
Yer İsmi	0.83	0.89	0.85
Kurum İsmi	0.82	0.82	0.82

Şartlı Rasgele Alanlar, Kural Tabanlı yaklaşım ile desteklendiğinde kişi varlık isimlerini bulmada başarı %91'den %92'ye yükselmiştir. Yer isminde ise %83'ten %85'e, kurum isminde de %78'den %82'ye yükselmiştir.

Yarı-yapılandırılmamış dokümanlarda yer varlık ismi diğerlerine göre daha başarısız iken yapılandırılmamış dokümanlarda kurum isimleri daha başarısız olmuştur (Şekil 5.2).



Şekil 5.2 Varlık isimlerinin F-ölçüm değerleri

5.3.1 Eğitim Veri Setinin Etkisi

Varlık İsmi Tanıma için geliştirilen sisteme eğitim veri setinin etkisini belirleyebilmek için ayrı bir çalışma yapılmıştır. Şartlı Rasgele Alanlar yöntemi ile sistemi eğitebilmek için hazırlanan eğitim veri setinin dörtte biri ile sistem eğitildiğinde elde edilen başarı, eğitim veri setinin yarısı ile sistem eğitildiğinde elde edilen başarı ve eğitim veri setinin dörtte üçü ile sistem eğitildiğinde elde edilen başarı ayrı ölçümlenmiştir. Böylece eğitim veri seti artırıldıkça eğitici sistemlerde başarı oranının etkisi gözlemlenmiştir.

Çizelge 5.9'da eğitim veri setinin dörtte biri olan 234,200 kelime ile sistem eğitildiğinde elde edilen F-ölçüm değerleri gösterilmiştir.

Çizelge 5.9 Eğitim veri setinin dörtte biri ile sistem eğitildiğinde

Şartlı Rastgele Alanlar	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.82	0.84	0.82
Yer İsmi	0.76	0.79	0.77
Kurum İsmi	0.79	0.77	0.77

Çizelge 5.10'da eğitim veri setinin yarısı olan 468,500 kelime ile sistem eğitildiğinde elde edilen F-ölçüm değerleri, Çizelge 5.11'de ise eğitim veri setinin dörtte üçü olan 702,000 kelime ile sistem eğitildiğinde elde edilen F-ölçüm değerleri kişi, kurum ve yer ismi bazında gösterilmiştir.

Çizelge 5.10 Eğitim veri setinin yarısı ile sistem eğitildiğinde

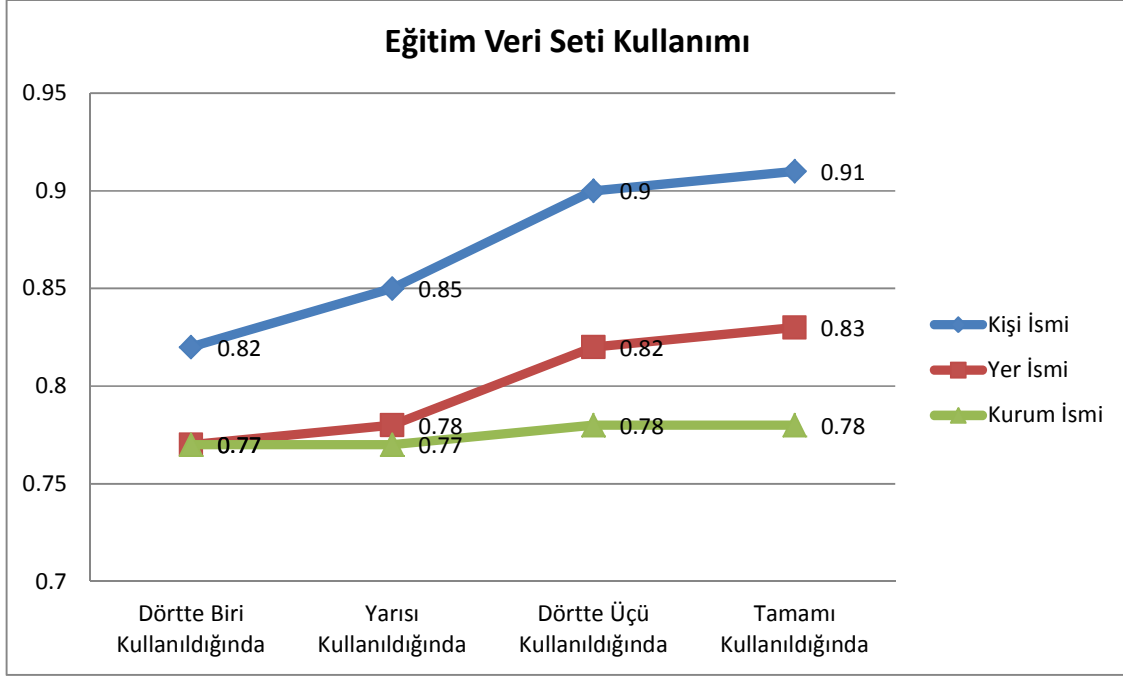
Şartlı Rastgele Alanlar	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.84	0.86	0.85
Yer İsmi	0.74	0.85	0.78
Kurum İsmi	0.80	0.76	0.77

Çizelge 5.11 Eğitim veri setinin dörtte üçü ile sistem eğitildiğinde

Şartlı Rastgele Alanlar	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.89	0.93	0.90
Yer İsmi	0.81	0.84	0.82
Kurum İsmi	0.81	0.77	0.78

Eğitim veri setinin etkisini belirleyebilmek için yapılan bu çalışmada elde edilen sonuçlar Şekil 5.3'te grafik olarak gösterilmiştir. Eğitim veri seti artırıldıkça başarı oranının da yükseldiği fark edilmiştir. Ancak sistem doyum noktasına ulaştığında ise bu

noktadan sonra eğitim veri seti miktarının artırılmasının başarıya etkisi pek bulunmamaktadır.



Şekil 5.3 Eğitim veri seti kullanımı ve F-ölçüm değerleri

5.3.2 Test Veri Setinin Etkisi

Yapılandırılmamış dokümanlar için geliştirilen sistemin başarısını ölçebilmek için şu ana kadar 23,369 kelimelik bir test veri seti kullanılmıştır. Bir sistemin başarısını ölçmede kullanılan test veri miktarı ve niteliği de etkilediğinden, test veri seti miktarı artırılarak ayrı bir ölçümleme de yapılmıştır. Çalışmanın bu adımında manuel olarak oluşturulan eğitim veri setinin etiketsiz hali test veri seti olarak kullanılmıştır. Sistemin eğitilmesinde ise diğer çalışmalardan derlenen eğitim veri seti [10] (Tur et al., 2003) sadece kullanılmıştır. Varlık İsmi Tanıma için geliştirilen sistem 406,984 kelimeye sahip veri seti ile test edildiğinde elde edilen varlık ismi tespitleri kişi, yer ve kurum ismi bazında Çizelge 5.12’de gösterilmiştir.

Çizelge 5.12 Test veri seti miktarı artırıldığında tespit edilen varlık sayıları
(UT, Uygulamanın Tespiti, DT, Uygulamanın Doğru Tespiti, GT, Gerçek Tespit)

Uygulama Türü	Şartlı Rastgele Alanlar UT/DT/GT	Şartlı Rastgele Alanlar + Kural Tabanlı Yaklaşım UT/DT/GT
Kişi İsmi	6,322/5,713/6,280	6,357/5,789/6,280
Yer İsmi	3,320/2,723/3,150	3,336/2,763/3,150
Kurum İsmi	2,870/2,325/2,817	2,925/2,338/2,817

Test veri seti miktarı artırılıp Şartlı Rastgele Alanlar kullanıldığında elde edilen F-ölçüm değerleri Çizelge 5.13’de gösterilirken, kural tabanlı yaklaşım eklenildiğinde elde edilen F-ölçüm değerleri Çizelge 5.14’de gösterilmiştir. Çalışmanın bu adımında sistemin eğitilmesi için daha az eğitim datası kullanılmasına rağmen, elde edilen sonuçlara göre test veri seti miktarı artırıldığında başarı oranlarıda değişmektedir.

Çizelge 5.13 Test veri seti miktarı artırıldığında F-Ölçüm değerleri

Şartlı Rastgele Alanlar	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.90	0.90	0.90
Yer İsmi	0.82	0.86	0.83
Kurum İsmi	0.81	0.82	0.81

Çizelge 5.14 İki yöntem birlikte kullanıldığında F-ölçüm değerleri

Şartlı Rastgele Alanlar + Kural Tabanlı Yaklaşım Kullanımı	Tutturma (P)	Bulma (R)	F-Ölçüm
Kişi İsmi	0.91	0.92	0.91
Yer İsmi	0.82	0.87	0.84
Kurum İsmi	0.79	0.82	0.80

5.3.3 Deneysel Sonuçların Karşılaştırılması

Yapılandırılmamış dokümanlar için geliştirilen Varlık İsmi Tanıma sisteminde iki ayrı test veri seti ile sistemin başarıları ölçümlenmiştir. Bu ölçümlmelerde her ne kadar aynı miktarda eğitim veri seti ile sistem eğitilmediysede elde edilen sonuçlar karşılaştırmalı olarak değerlendirilebilmektedir.

Eğitim veri seti artırıldıkça sistemin başarı oranı yükselmektedir. Aynı şekilde test için kullanılan test veri miktarı artırıldığında elde edilen sonuçlarda da başarı oranı yükselmektedir. Buradaki en büyük etki, veri setinde bulunan varlık miktarının daha fazla olmasıdır.

Türkçe’de Varlık İsmi Tanıma için daha önceden yapılan çalışmalar arasında [10] (Tur et al., 2003) daki eğitim veri setini kullanan çalışmaların başarı oranlarına göre sistemin başarıları değerlendirildiğinde %92 ile kişi isimlerinde en iyi başarı elde edilmektedir.

Yeniterzi’nin morfolojik özelliklerin kullanılmasıyla gerçekleştirdiği çalışmada [11] en iyi başarı sonucu %88 iken, [10] Tur et al.’in yaptığı çalışmada en iyi başarı sonucu %91 dir.

Ancak aynı veri setini kullanarak sistem başarısını ölçen bir çalışma olmadığından tam bir karşılaştırma yapılamamaktadır.

SONUÇ VE ÖNERİLER

Varlık İsmi Tanıma, bilgi çıkarımı ve bilgiye erişim alanında kendisine yer edinse de, dile bağlı olarak geliştirilen yöntemler ile doğal dil işleme altında da kendisine yer bulmuştur. Varlık İsmi Tanıma üzerine yapılan çalışmalar Türkçe’de sınırlı sayıda olup, çalışmaların çoğu alana bağlıdır. Bu çalışmada resmi olmayan yarı-yapılandırılmış ve yapılandırılmamış dokümanlar için Şartlı Rastgele Alanlar kullanılarak kişi, kurum ve yer varlık isminin tanınması gerçekleştirilmiştir.

Yapılandırılmış doküman olarak kullanılan Türkçe e-postalarda başarının en yüksek olduğu varlık kişi ismidir. İkinci sırada ise kurum ismi gelmektedir. Tanımada başarının en düşük olduğu varlık, yer isimleridir. E-postalar içerisinde yer isimlerinin yazılmasında, yer isminden sonra gelen iyelik, durum ve bildirme eklerinde kesme işaretinin kullanılmadığı gözlemlenmiştir. Bu durumda yer isimlerinin hatalı etiketlenmesine sebep olmaktadır.

En başarılı e-postalar kurumsal içerikli olup, ortalama %87 doğrulukta tanıma yaparken, en az hata ile tanınan varlık ismi %95 ile kişi ismi olmuştur. En başarısız da ortalama %72 ile yer varlık ismidir. Türkçe üzerine aynı veri setini kullanan bir çalışma olmadığından sistemin başarısı farklı çalışmalarla karşılaştırılamamıştır.

Yapılandırılmamış dokümanlar için elde edilen deneysel sonuçlar incelendiğinde Şartlı Rastgele Alanlar yöntemi ile başarının en yüksek olduğu varlık kişi ismidir. Bu yöntem de en az başarı elde edilen varlık sınıfı ise kurum isimleridir.

Varlık İsmi Tanıma işleminde sadece Şartlı Rastgele Alanlar yöntemi yarı eğitici makine öğrenmesi tekniği ile uygulandığında en başarılı ortalama %92 ile kişi ismi olurken, yer ismi %85 , kurum ismi ise %82 başarı ile tespit edilebilmektedir.

Eğitim ve test verisi detaylı incelendiğinde başarıya etki eden faktörler belirlenebilmektedir. En önemli faktör eğitim datasında etiketlenmiş varlık miktarıdır. Eğitim verisi ne kadar çok etiketlenmiş varlık içerirse başarı oranı da o ölçüde etkilenmektedir. Ayrıca, eğitim verisinde bulunan varlıkların nitelikleride başarıyı etkileyebilmektedir. Bir varlık sınıfına ait etiketlenmiş veri miktarı çok tekrar içermesi durumunda olasılık hesabı başarılı yapılamamaktadır.

Kural tabanlı yaklaşım kullanılarak uygulama genişletildiğinde ise başarı oranında artış gözlemlenebilmektedir. Bu yaklaşım ile başarı oranında en büyük artış kurum isimlerinde elde edilmektedir. Buradaki en büyük etken eğitim verisindeki etiketlenmiş kurum ismi miktarı özellikle kişi ismine göre daha az bulunmaktadır. Şartlı Rastgele Alanlar yöntemi ile tespit edilemeyen kurum isimlerinin bazıları kural tabanlı yaklaşım ile tespit edilebilmiştir.

Geliştirilen sistemde kural tabanlı yaklaşımın kişi ve yer ismi için başarı oranına etkisi kurum isimleri kadar etkili olmamaktadır.

Sistemi test ederken kullanılan test verisi içerisinde tekrar eden varlık miktarının çok olması geliştirilen sistemin başarı oranını ölçerken etkili olmuştur. Bu nedenle Türkçe için geliştirilen Varlık İsmi Tanıma uygulamaları için aynı test verileri kullanılırsa sistemin başarısı diğer yaklaşımlara göre daha net ölçülebilmektedir. Bu çalışmada deneysel sonuçlar çıkartılırken aynı test verisi ile bütün senaryolar çalıştırılmıştır.

Varlık İsmi Tanıma için geliştirilen sisteme test için kullanılan veri uygulama çalışması sonrasında elle kontrol edilerek hatalı kısımlar düzeltilip eğitim verisine aktarılabilir. Böylece her defasında eğitim verisi genişletilerek daha yüksek başarı oranı elde edilebilir.

KAYNAKLAR

- [1] Nadeau, D. ve Sekine, S., (2007). "A survey of named entity recognition and classification", *Linguistica Investigationes*, 30:3-26.
- [2] Rau, L.F., (1991). "Extracting Company Names from Text", In *Proc. Conference on Artificial Intelligence Applications of IEEE*.
- [3] Nadeau, D., (2007). *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*, Doktora Tezi, Ottawa-Carleton Institute for Computer Science School Of Information Technology and Engineering University of Ottawa, Canada.
- [4] Alfonseca, E. ve Manandhar, S., (2002). "An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery", *International Conference on General WordNet*.
- [5] Brin, S., (1999). "Extracting Patterns and Relations from the World Wide Web", *The World Wide Web and Databases, LNCS*, 1590:172-183.
- [6] Cucerzan, S., ve Yarowsky, D., (1999). "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence", In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- [7] Bayraktar, Ö. ve Temizel, T.T., (2008). "Person Name Extraction From Turkish Financial News Text Using Local Grammar-Based Approach", *ISCIS Computer and Information Sciences*, 16 Aralık 2008
- [8] Küçük, D. ve Yazıcı A., (2009). "Rule-based Named Entity Recognition from Turkish Texts", *International Symposium on Innovations in Intelligent Systems and Applications*, 29 June-1 July 2009, Trabzon.
- [9] Dalkılıç, F. E., Gelişli, S. ve Diri, B., (2010). "Türkçe Kural Tabanlı İsim Varlık Tanıma", *Signal Processing and Communications Applications Conference*, 918-920.
- [10] Tür, G., Hakkani-Tür, D. ve Oflazer, K., (2003). "A Statistical Information Extraction System for Turkish", *Natural Language Engineering*, 9:181-210.

- [11] Yeniterzi, R., (2011). "Exploiting Morphology in Turkish Named Entity Recognition System", Student Session ACL 2011, Portland, Oregon.
- [12] Johnson, C. S. (1967). "Hierarchical Clustering Schemes", *Psychometrika*, 2, 241-254.
- [13] Anderson, J. R., Michalski, R. S., Carbonell, J. G., ve Mitchell, T. M. (1985). *Machine Learning: An Artificial Intelligence Approach*, 1.
- [14] MacQueen, J. B., (1967). "Some Methods for classification and Analysis of Mutivariate Observations", *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281-297., Berkeley.
- [15] Ekbal, A. ve Bandyopadhyay, S., (2009). "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi", *Linguistic Issues in Language Technology-LiLT*, 2(1).
- [16] Douthat, A., (1998). "The Message Understanding Conference Scoring Software User's Manual", in *The Message Understanding Conference*.
- [17] Sarawagi, S., (2003). Conditional Random Field Project Page, <http://crf.sourceforge.net/>, 15 Ekim 2010.
- [18] Sasaki, Y. ve Fellow, R., (2007). "The truth of the F-measure", University of Manchester MIB.
- [19] Wallach, H. M., (2004). "Conditional Random Fields: An Introduction", University of Pennsylvania Technical Report MS-CIS-04-21, 24 February 2004.
- [20] Cormen, T. H., Leiserson, C.E., ve Rivest, R.L., (1990). "Introduction to Algorithms", Third Edition, MIT Press/McGraw-Hill, New York.
- [21] Ratnaparkhi, A., (1997). "A simple introduction to maximum entropy models for natural language processing", Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, 1997, Washington D.C
- [22] Rabiner, L. R., (1989). "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, 77(2):257-285.
- [23] Berger, A. L., Della Pietra S. A., ve Della Pietra V. J., (1996). "A maximum entropy approach to natural language processing", *Computational Linguistics*, 22(1):39-71.
- [24] Chinchor, N., (1992). "MUC-4 Evaluation Metrics", *Proc. of the Fourth Message Understanding Conference*, 23-25 August 1992, San Diego.
- [25] Lafferty, J., McCallum, A. ve Pereira, F., (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data", *18th International Conference on Machine Learning*, 282-289.
- [26] Şeker, G. A. ve Eryiğit, G., (2012). "Initial explorations on using CRFs for Turkish Named Entity Recognition", *24th International Conference on Computational Linguistics*, 8-15 December 2012, Mumbai, India.

- [27] Tatar, S. ve Çiçekli, I., (2011). "Automatic rule learning exploiting morphological features for named entity recognition in Turkish", Journal of Information Science, 37(2):137-151.
- [28] Küçük, D. ve Yazıcı, A., (2012). "A hybrid named entity recognizer for Turkish", Expert System with Applications, 39(3), 2733-2742.

ÖZEL KELİMELER

Varlık İsmi Tanım'ya yönelik oluşturulan sistemde temel özelliklerin çıkartılmasında yardımcı olan kısaltmalar, ünvanlar ve bazı özel kelimeler varlık isimlerini belirlemede kullanılmıştır. Türkçe'de özellikle kişi ve kurum isimleri ile birlikte kullanılan ünvan, makam ve kısaltmalar özel kelime listesine eklenmiştir.

A-1 Özel Kelime Listesi

ABD	Amerika Birleşik Devletleri(USA)
AGİK	Avrupa Güvenliği ve İşbirliği Konferansı
Adr.	Adres
Alb.	Albay
ANAP	Anavatan Partisi
Ank.	Ankara
Apt.	Apartman
Ar.Gör.	AraştırmaGörevlisi
As.	Asker, askerî, askerlik (Askerî terim) Asistan
Astsb.	Astsubay
AŞ,A.Ş.	Anonim Şirketi
Atğm.	Asteğmen
Atm.	Atmosfer
AÜ	Ankara Üniversitesi
BAC, B.A.C.	Birleşik Arap Cumhuriyeti
BBC, B.B.C.	British Broadcasting Cooperation (İngiliz Radyosu)
Bçvş.	Başçavuş
BEA	British European Airways (İngiliz Avrupa Hava Yolları)
BJK, B.J.K.	Beşiktaş Jimnastik Kulübü
Bkz. (bkz.)	Bakınız
Bl.	Bölük (Askerî terim)
Bl.K., Bl.K.lığı	Bölük Komutanı, Bölük Komutanlığı (Askerî terim)

Bn.	Bayan
Bnb.	Binbaşı
Bşk. Bşk.lık	Başkan, Başkanlık
Bş. Öğ.	Başöğretmen
Bştbp.	Baştabip
Bul.	Bulvar
BÜ	Boğaziçi Üniversitesi
Cad.	Caddesi
C.C.	Celle Celâlühü (Uludur, Yücedir.)
CHP, C.H.P.	Cumhuriyet Halk Partisi
CIA, C.I. A.	Central Intelligence Agency (ABD Merkezî Haberalma Teşkilâtı)
C.G.S.	Santimetre, gram, saniye birim sistemi
CIF, c.i.f.	Mâliyet, sigorta, navlun (Cost. in surarance, freight)
cm.	Sa ntimetre
Cmh. Bşk.	Cumhurbaşkanı
ÇEK, Ç. E.K.	Çocuk Esirgeme Kurumu
Çev.	Çeviri, çeviren
ÇNAEM	Çekmece Nükleer Araştırma ve Eğitim Merkezi
Çvş.	Çavuş
d.	Dakika
DDT	Dichlor-Diphenyl-Trichlor-aethan (Pire, sinek vb. yok etme ilâcı)
DHMI	Devlet Hava Meydanları İşletmesi
Dipl.	Diploma, Diplomalı
DM	Deutsche Mark (Alman Markı)
DMO	Devlet Malzeme Ofisi
DNA	Deoksiribo Nükleik Asit
Doç.	Doçent
DPT	Devlet Planlama Teşkilâtı
Dr.	Doktor
DSİ	Devlet Su İşleri
DTCF	Dil ve Tarih-Coğrafya Fakültesi
Dz. Ataş.	Deniz Ataşesi
Dz. HO	Deniz Harp Okulu
Dz. K.	Deniz Komutanı
Dz. Kuv.	Deniz Kuvvetleri
E	East, Est (Doğu)
EBK	Et ve Balık Kurumu
Ecz.	Eczane
Ed. F.	Edebiyat Fakültesi
Ef.	Efendi, Efrat (As.)
EFTA	European Free Trade Association (Avrupa Serbest Ticaret Ortaklığı)
EGO	Elektrik Gaz Otobüs İşletmeleri (Ankara Belediyesi)
EKG	Elektrokardiyografi
EOKA	Ethnike Organosis Kypriotikes Apelentberoseos (Kıbrıs'ta Rum Çete Örgütü)
ESHOT	Elektrik Su Havagazı Otobüs Tramvay İşletmeleri (İzmir Belediyesi)

FAO	Food and Agriculture Organization (Birleşmiş Milletler Gıda ve Tarım Örgütü)
FB., F. B.	Fenerbahçe Spor Kulübü
FBI	Federal Bureau of Investigation (Amerikan Millî Emniyeti)
FIFA	Federation Internationale de Football Association (Milletlerarası Futbol Federasyonu)
FIS	Federation Internationale de Ski: (Milletlerarası Kayakçılar Federasyonu)
FKB	Fizik, Kimya, Biyoloji
FOA	Foreign Operations Administration (Amerika'nın yabancı devletlere yardım fonu)
Gal. Galon	
Gmr.	Gümrük
Gn. Kur.	Genel Kurmay
Gön.	Gönderen
g.	Gram
Gram.	Gram Dilbilgisi
GS., G.S.	Galatasaray Spor Kulübü
H.	Hidrojen, Hicrî
HO, H.O.	Harp Okulu
HP	Beygircü
Hrk. Bşk., Bşk.lığı	Harekât Başkanı, Başkanlığı
Hrp. T. D. Bşk., Bşk.lığı	Harp Tarihi Dairesi Başkanı, Başkanlığı
Hst.	Hastâne
Hst. Bş. Hek.	Hastane Başhekimî, Başhekimliği
Hv.	Hava
Hv. Kuv.	Hava Kuvvetleri
Hız.	Hazret veya hazretleri
ICAO	International Civil Aviation Organization (Milletlerarası Sivil Havacılık Teşkilâtı)
IMCO	Inter-Governmental Maritime Consultative Organization (Milletlerarası Denizcilik Danışma Teşkilâtı)
IMF	International Monetary Fund (Milletlerarası Para Fonu)
INS	International News Service (Milletlerarası Haber Servisi)
IRO	International Refugee Organization (Milletlerarası Mülteci Teşkilâtı)
İHA	İhlas Haber Ajansı
İÜ	İstanbul Üniversitesi
İTÜ	İstanbul Teknik Üniversitesi
İmp.	İmparator
Kd.	Kıdemli
kg.	Kilogram
Khz.	Kilohertz
KİT	Kamu İktisâdî Teşekkülleri
KKK	Kara Kuvvetleri Komutanı
K. Kuv.	Kara Kuvvetleri
Kl.	Kilolitre
KLM	Koninklijke Luchtvaart Maatschappij (Hollanda Kraliyet Hava Yolları)

km.	Kilometre
Koll.Şti.	Kollektif Şirketi
Kom.	Komisyon
Kom.Şti.	Komandit Şirket
Kor.	Kolordu
Kora.	Koramiral
Korg.	Korgeneral
krş	Kuruş, Karşılaştırınız
KS, K.S.	Kuddise Sirruh
Kur.	Kurmay
Kur. Bşk.	Kurmay Başkanı
Ltd. Şti.	Limited Şirketi
Ltd.	Limited
lt.	Litre
Log.	Logaritma
Lv.	Levazım
m., m	Metre
mad.	Madde
Md.	Müdür
M. E. B.	Milli Eğitim Bakanlığı
mg.	Miligram
MİT	Millî İstihbarat Teşkilâtı
Mm	Milimetre
MHP	Milliyetçi Hareket Partisi
Mrşl.	Mareşal
MS, M.S.	Milattan Sonra
MTA	Mâden Tetkik Arama (Enstitüsü)
MTTB	Millî Türk Talebe Birliği
Mu.	Muhabere
Müh.	Mühendis
Mv. Hast.	Mevkî Hastanesi
N	North (Kuzey)
NASA	National Aeroautics And Space Administialion (ABD Millî Havacılık ve Feza Dâiresi)
NATO	North Atlantic Treaty Organization (Kuzey Atlantik Paktı Teşkilâtı)
No.	Numara
O.W.	Ouest West: Batı
ODTÜ	Orta Doğu Teknik Üniversitesi
Onb.	Onbaşı
Opr.	Operatör
Or.	Oramiral
Ora.	Oramiral
Ord.	Ordinaryüs
Ord. Prof.	Ordinaryüs Profesör
Org.	Orgeneral
Ort.	Ortaklık

Oto.	Otomobil
Öğ.	Öğretmen
ölm.	Ölüm, Ölümü
Ör.	Örnek
P.	Polis
PK, P.K.	Posta Kutusu
Prof.	Profesör
Pş.	Paşa
PTT	Posta Telgraf Telefon (İdaresi)
Rd.	Radar
RNA	Ribo Nükleik Asit
r.	Yarıçap
r.anh, r.anha	Radıyallahü anh, anha (Allah ondan razı olsun)
r.aleyh	Rahmetullahi aleyh (Allahü teâlânın rahmeti onun üzerine olsun)
S	South (Güney)
s.	Sayfa
sa.	Saat
SABENA	Société Anonyme Belge D'Exploitation de la Navigation Aérienne (Belçika Hava Yolları)
s.a.v	Sallallahü aleyhi ve sellem (Ona hayır dualar olsun)
Sb.	Subay
Sin.	Sinüs
SSCB	Sovyet Sosyalist Cumhuriyetler Birliği
SEATO	Sout East Asia Treaty Organization (Güneydoğu Asya Antlaşması Teşkilâtı)
sm., cm.	Santimetre
s.	Saniye
Sok.	Sokak
SOS	Save Our Souls (Denizde milletlerarası acele yardım isteme imdat işareti)
Şb.	Şube
Şrt.	Şirket
TAO	Türk Anonim Ortaklığı
TASS	Telgrafneo Agenstsvo Sovetskovo Soyuzu (Rus Haber Ajansı)
TAŞ	Ticaret Anonim Şirketi
Tb.	Tabur
Tbp.	Tabip
TBMM	Türkiye Büyük Millet Meclisi
TC	Türkiye Cumhuriyeti
TCDD	Türkiye Cumhuriyeti Devlet Demiryolları
TCK	Türk Ceza Kânunu
Tel.	Telefon
Telg.	Telgraf
TGRT	Türkiye Gazetesi Radyo Televizyonu
Tğm.	Teğmen
THA	Türk Haberler Ajansı THK Türk Hava Kurumu, Türk Hukuk Kurumu
THY	Türk Hava Yolları
TKİ	Türkiye Kömür İşletmeleri

TL	Türk Lirası
TMMOB	Türkiye Mühendis ve Mimar Odaları Birliği
TMO	Toprak Mahsülleri Ofisi
TNT	Trinitrotoluen (Dinamit)
TRT	Türkiye Radyo Televizyon Kurumu
TPAO	Türkiye Petrolleri Anonim Ortaklığı
Tug.	Tugay
Tuğg.	Tuğgeneral
Tümg.	Tümgeneral
TV	Televizyon
UFO	Unidentified Flying Object (Bilinmeyen Uçan Cisim)
UNESCO	United Nations Educational Scientific and Cultural Organization (Birleşmiş Milletler Eğitim Bilim ve Kültür Teşkilâtı)
UNICEF	United Nations International Children's Emergency Fund (Birleşmiş Milletler Milletlerarası Çocuklara Acil Yardım Fonu)
UP	United Press (Associations) (Amerikan Haberler Ajansı)
UPU	Universal Postal Union, Union Postale Universelle (Milletlerarası Posta Birliği)
USA	United States of America (Amerika Birleşik Devletleri)
Uzm.	Uzman
Üçvş.	Üçşavuş
Ütgm.	Üsteğmen
V	Volt
vb.	ve benzeri, ve bunun gibi, ve başkaları
vd.	ve devamı
Vet.	Veteriner
vs.	ve saire
W	West (batı), Watt
WB	World Bank (Dünya Bankası)
WHO	World Health Organization (Dünya Sağlık Birliği)
WHO	World Meteorological Organization (Dünya Meteoroloji Birliği)
Yd.	Yedek
Yd. Sb.	Yedek Subay
Y. Müh.	Yüksek Mühendis
Yb.	Yarbay
yy.	Yüzyıl
Yzb.	Yüzbaşı
Yrd. Doç.	Yardımcı Doçent

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Serap ÖZKAYA
Doğum Tarihi ve Yeri : 1982 Malatya
Yabancı Dili : İngilizce
E-posta : serap.ozkaya@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Lisans	Bilgisayar Mühendisliği	Ege Üniversitesi	2005
Lise	Fen Bilimleri	Atatürk Üniversitesi	1999

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2010	Avea İletişim Hizmetleri	Uzman Yazılım Mühendisi
2005	Vodafone Teknoloji Hizmetleri	Yazılım Mühendisi

YAYINLARI

Bildiri

1. Türkçe Metinlerde Şartlı Rastgele Alanlarla Varlık İsmi Tanıma, IEEE 19. Sinyal İşleme ve İletişim Uygulamaları Kurultayı – SUI 2011