# Türkçe Haber Metinlerinde Şartlı Rastgele Alanlar ile Varlık İsmi Tanıma: Özniteliklerin Gözden Geçirilmesi[1]

# Named Entity Recognition with Conditional Random Fields on Turkish News Dataset: Revisiting the Features

*R.Fırat ÇEKİNEL*[*], *Mustafa AĞRIMAN*[*], *Pınar KARAGÖZ*[*], *Burcu YILMAZ*[†]
[*]*Orta Doğu Teknik Üniversitesi Bilgisayar Mühendisliği*, [†]*Gebze Teknik Üniversitesi Bilgisayar Mühendisliği*
*rfcekinel@ceng.metu.edu.tr*, *mustafa.agriman@ceng.metu.edu.tr*, *karagoz@ceng.metu.edu.tr*, *byilmaz@gtu.edu.tr*

***Özetçe* —Varlık ismi tanıma, farklı tipteki metinlerden kişi, yer, kurum, tarih, saat, para ve yüzde gibi varlık isimlerinin işaretlenmesini amaçlayan bir doğal dil işleme problemidir. Bu dalda yapılan çalışmalarla, konum tahmini, olay zamanı tahmini, metinde geçen önemli kişilerin belirlenmesi gibi farklı işlemler mümkün olabilmektedir. Türkçe metinler üzerine bu alanda yapılan çalışmalar İngilizce'ye göre oldukça sınırlıdır. Bu çalışmamızda Türkçe haber metinlerde varlık ismi tanıma için *Şartlı Rastgele Alanlar* tekniğinin kullanımı gözden geçirilerek yeni özniteliklerin model doğruluğuna etkisi analiz edilmiştir. Çeşitli veri kümeleri üzerinde yapılan deneylerle benzer çalışmalarla karşılaştırmalı olarak sonuçlar sunulmuştur.**

***Anahtar Kelimeler*—*Varlık İsmi Tanıma; Şartlı Rastgele Alanlar; Türkçe.***

***Abstract*—Named entity recognition is a natural language processing problem that aims to mark entity names, such as person, place, organization, date, time, money and percentage, from different types of text. Various applications such as location estimation, event time estimation, determination of important people in the text can be possible with the solutions to this problem. The number of named entity recognition studies on Turkish texts is quite limited compared to those on English. In this study, the use of the technique Conditional Random Fields for the named entity recognition in Turkish news texts has been reviewed and the effect of new attributes on the model accuracy has been analyzed. The results of the experiments on different data sets are compared with the similar studies and presented.**

***Keywords*—*Named Entity Recognition; Conditional Random Fields; Turkish.***

## I. GİRİŞ

Varlık İsmi Tanıma (VİT) problemi, makina çevirisi, duygu analizi, sözdizimsel parçalama gibi doğal dil işleme işlerinin önemli bir aşamasını oluşturmaktadır [1]. Günümüzde İngilizce [1] başta olmak üzere birçok dil için bu alanla ilgili çalışmalar yürütülmekte ve yüksek başarılar elde edilmekteyken Türkçe gibi biçimbilimsel olarak karmaşık dil yapısına sahip dillerde zorlayıcı bir problem olmaya devam etmektedir.

Varlık ismi tanıma çalışmalarında farklı yöntemler kullanılmaktadır. İlk geliştirilen sistemlerde istatistiksel ve kural tabanlı çalışmalar ağırlıktayken günümüzde daha sıklıkla makine öğrenmesi ve derin öğrenme tabanlı sistemler geliştirilmektedir. Türkçe varlık ismi tanıma üzerine yapılan ilk çalışma 1999 yılında Cucerzan and Yarowski tarafından [17] Türkçe'nin de aralarında olduğu 5 farklı dil için geliştirilen sistemdir. 2003 yılında yayınlanan [3] çalışmada ise haber metinleri üzerinde Saklı Markov Modelleri kullanılarak isim, kurum ve yer isimleri işaretlenmiştir. Türkçe için ilk kural tabanlı sistem Küçük ve Yazıcı [4] tarafından sunulan çalışmada geliştirilmiş olup, daha sonradan Küçük ve Yazıcı [5], bu kural tabanlı sistemi daha da geliştirilerek hibrit bir sistem haline getirmişlerdir. Tatar ve Çiçekli [14]'nin çalışmasında, VİT kurallarını sözcüklerin biçimbilim, imla ve bağlam özelliklerine göre otomatik olarak öğrenen kural tabanlı bir sistem geliştirilmiştir.

Türkçe için Şartlı Rastgele Alanlar (CRF) tabanlı çeşitli VİT çalışmaları bulunmaktadır. Yeniterzi [6] tarafından geliştirilen sistemde biçimbilimsel yapının varlık ismi tanıma üzerine etkisi incelenmiştir. Ayrıca kelime ve biçimbilimsel tabanlı iki farklı model üzerinden ölçümler yapılmış ve biçimbilimsel tabanlı modelde belirgin bir iyileşme gözlemlenmemiştir. Özkaya ve Diri [7] tarafından geliştirilen sistemde ise akademik, kurumsal ve kişisel e-posta içeriklerinde yer alan kişi, kurum ve yer adları çıkarılmıştır. En başarılı sonuçlara %87 doğru tanıma oranıyla kurumsal e-postalarda ulaşılmıştır. Şeker ve Eryiğit [8] tarafından geliştirilen çalışmada ise haber metinlerinde kişi, yer, kurum varlık tanımada %91.94 (CoNLL F1-ağırlıklı) başarı elde edilmiştir. Daha sonradan bu çalışmadaki varlık isimlerine ek olarak tarih, saat, para ve yüzde işaretlemeleri eklenmiştir. Genişletilmiş modelin başarı oranı %92.34 (CoNLL F1-ağırlıklı) olarak belirtilmiştir [9].

Demir ve Özgür [15] tarafından yapılan çalışmada, yapay sinir ağları ile yarı gözetimli öğrenme (semi-supervised learning) yöntemiyle VİT sistemi geliştirilmiştir. Sistem %91.85

[1]Bu çalışma TÜBİTAK tarafından 117E566 no'lu proje kapsamında desteklenmiştir.

F1 puanına sahiptir. 2018 yılında Güneş ve Tantuğ [16] tarafından DBLSTM yapay sinir ağları modelini haber metinleri veri kümesinde uygulayarak %93.69 F1 puanına ulaşmışlardır. 2018 yılında Devlin ve diğerleri [10] tarafından geliştirilen derin öğrenme tabanlı sistem (Bert), farklı doğal dil işleme problemlerinde kullanılabilmektedir. Çift taraflı eğitilen bu dil modeli VİT için de uygulanabilmektedir. CoNLL İngilizce 2003 verisi üzerinde %92.80 F1 puanına ulaşılmıştır. Ayrıca Türkçe için de dil desteği sağlanmaktadır.

Bu çalışmamızda, Türkçe haber metinleri üzerinde çalışılmış olup varlık isimlerinin tanınmasına yönelik bir sistem geliştirilmiştir. Şartlı Rastgele Alanlar kullanılarak geliştirilen bu sistemde farklı özniteliklerin varlık ismi tanıma üzerindeki etkisi ve geliştirilen sistemin sonuçları ayrıntılı olarak sunulmuştur. Önceki çalışmalardan farklı olarak sistemimizde sözlük kullanılmamış ve kelime kökü yerine kelimenin konumsal öznitelikleri modele eklenmiştir. Elde edilen sonuçlar karşılaştırmalı olarak sunulmuştur. Ayrıca Türkçe haber veri kümesi kullanılarak Bert'te VİT sisteminin başarısı ölçülmüştür.

Bildirimizin kalan kısmı şu şekilde düzenlenmiştir: Bölüm II'de kullanılan yöntem, veri kümesi ve kullanılan öznitelikler sunulmaktadır. Bölüm III'te deneysel sonuçlar sunulup Bölüm IV'te çalışmamızın sonuçları özetlenmiştir.

## II. VARLIK İSMI TANIMA

### A. Şartlı Rastgele Alanlar

Şartlı Rastgele Alanlar [2], sıralı verileri işaretlemek için kullanılan ayırt edici (discriminative) olasılıksal bir modeldir. Öznitelik kümesi kullanılarak koşullu olasılık dağıtımı modellenmektedir. Üretici (generative) bir model olan Saklı Markov Modellerinin aksine farklı sınıflar arasındaki sınır optimize edilmeye çalışılır. Ayrıca Saklı Markov Modellerine göre daha zengin bir işaret dağıtım kümesi de modelleyebilir. Bunun sebebi ise Saklı Markov Modelleri'nde her kelimenin işareti sadece kendisine ve bir önceki kelimenin işaretine göre belirlenirken Şartlı Rastgele Alanlarda daha esnek bir yapı sunulmasıdır.

Çalışmamızda L-BFGS Şartlı Rastgele Alanlar kullanılmıştır. Bu modelde $P(y|x)$ yani $y$ olarak belirtilen işaret için $x$ olarak belirtilen öznitelikler kullanılarak Eşitlik 1'de belirtilen şekilde hesaplamalar yapılır.

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

$f_k(y_{t-1}, y_t, x_t)$ fonksiyonu, verilen $x_t$ özniteliklerine göre $y_{t-1}$ durumundan $y_t$ durumuna geçme olasılığını hesaplamak için kullanılmaktadır. Model eğitilirken $\theta_k$ optimize edilir. $z_\theta(x)$ ise normalizasyon faktörü olarak kullanılmaktadır. [11] Çalışmamızda CRFSuite'in [12] Python programlama diline uyumlanmış sklearn-crfsuite[1] kütüphanesi kullanılmıştır.

[1]https://sklearn-crfsuite.readthedocs.io/en/latest/

| Kelime | Tür | Sınıf |
|---|---|---|
| İstanbul | İsim | YER |
| yüzde | İsim | YÜZDE |
| 2013 | Sayı | TARİH |
| Meclis | İsim | KURUM |
| lira | İsim | PARA |
| şimdi | Zarf | ZAMAN |

TABLO I: Örnek kelimelerin işlenmeye hazır hali

### B. Veri Kümesi

Çalışmamızda Tür ve diğerleri [3] tarafından işaretlenmiş haber veri kümesi kullanılmıştır. Veri kümesi yaklaşık 500K kelime içermektedir. Haber metinlerinde kişi, yer ve kurum isimleri işaretlenmiştir. [6]'da sunulan çalışmada belirtilen yöntemle veri kümesinin %90'ını modeli eğitmek için %10'unu ise test etmek kullanılmıştır (Bundan sonra bu çalışmada eğitim veri kümesi TRAIN3, test veri kümesi ise TEST3 olarak anılacaktır).

[9] tarafından yapılan çalışmada ise aynı veri kümesindeki varlık ismi çeşitliliği arttırılmıştır. Haber metinlerindeki kişi, kurum ve yer isimlerinin yanı sıra tarih, saat, para ve yüzde isimleri de işaretlenmiştir (Bundan sonra bu çalışmada eğitim veri kümesi TRAIN7_ALL olarak anılacaktır). İlgili çalışmanın sadece eğitim veri kümesine erişebildiğimiz için bu veri kümesindeki doğru tanıma oranı 10-katlı çapraz doğrulama ile değerlendirilmiştir.

Geliştirdiğimiz modelde kullandığımız özniteliklerin modelin başarısına katkısını ölçmek için TRAIN7_ALL veri kümesinde rastgele olarak ayırma işlemi yapılmış, verinin %90'ı ile eğitim veri kümesi, %10'u ile de test veri kümesi oluşturulmuştur (Bundan sonra bu çalışmada eğitim veri kümesi TRAIN7, test veri kümesi ise TEST7 olarak anılacaktır).

Haber metinlerinin yanısıra [9] tarafından işaretlenen sosyal medya metinleri de modelin başarısını değerlendirmek için test veri kümesi olarak kullanılmıştır. (Bundan sonra bu çalışmada bu veri kümesi IWT7 olarak anılacaktır.)

### C. Geliştirilen Sistem ve Kullanılan Öznitelikler

VİT sistemi Python uygulaması olarak yazılmıştır. Ön işleme aşamasında veri kümesindeki her haber metni kelimelere göre ayrılmıştır. Ayrıca kelimenin içerisinde ayraç geçiyorsa ayraçtan itibaren ek kısmı yeni bir kelime olarak yazılmıştır. Daha sonra Zemberek projesi [13] kullanılarak sözcük türleri (POS Tags) bulunmuştur. Tablo I'de örnek kelimelerin VİT sistemine uygun hale getirilmiş hali gösterilmiştir.

Haber metinleri VİT sistemine uygun hale getirildikten sonra aşağıda açıklanan öznitelikler kullanılarak Şartlı Rastgele Alanlar modeli oluşturulmuştur. Her kelime için kelimenin kendisi ve iki kelime öncesine ve iki kelime sonrasına kadar olan kelimeleri de içerecek şekilde modele öznitelikler verilmiştir. Cümlenin başında ve sonunda bulunan kelimeler için öncesi ve sonrası kelimeler için eklenen öznitelikler çıkarılmış bu özniteliklerin yerine kelimenin cümlenin sonu ve başı olduğunu gösteren öznitelikler eklenmiştir.

*1) Temel Öznitelikler:*

- **Sözcük Yüzey Formu:** VİT çalışmalarında [6] belirtildiği gibi sözcüklerin yüzey formu kullanılmaktadır.

Ancak Türkçe gibi biçimbilimsel açıdan zengin dillerde veri seyrekliği problemiyle karşılaşmamak adına kelimeler küçük harfe çevrilmeli ve kelimelerin yüzey formunun yanında başka öznitelikler de kullanılmalıdır.

- **Cümlenin Başlangıcı:** Kelime cümlenin başındaysa öznitelik kümesine ilk kelime olduğunu ifade eden *"bos" (beginning of sentence)* etiketi eklenmiştir.
- **Cümlenin Sonu:** Kelime cümlenin sonundaysa öznitelik kümesine son kelime olduğunu ifade eden *"eos" (end of sentence)* etiketi eklenmiştir.

*2) Konumsal Öznitelikler:*

- **Kelime Başı (KB):** Şartlı Rastgele Alanlar ile varlık ismi tanıma sistemi geliştirilen birçok çalışmada [6], [8], [9] biçimbilimsel analizler ile elde edilen kelime kökü modellere öznitelik olarak sunulmaktadır. Çalışmamızda ise kelimenin başındaki ilk dört ve ilk beş karakterler modele öznitelik olarak verilmiştir.
- **Kelime Sonu (KS):** Bazı varlık isimlerinin son ekleri kelimenin ne olabileceğine dair çıkarımda bulunmamıza yardımcı olur. Örneğin "stan" söz öbeği ülke isimlerinde sıkça geçmektedir. Bulgaristan, Hindistan ve Sırbistan gibi birçok ülke isminin sonunda bu tip ekler bulunmaktadır. Bu tip durumlarda modelin daha başarılı işaretleyebilmesi için kelimelerin son üç ve son dört karakterleri modele öznitelik olarak verilmiştir.

*3) Biçimbilimsel Öznitelikler:*

- **Sözcük Türleri (Part-of-Speech Tags)(POS):** Kelimelerin sözdizimsel özelliklerini gösteren ögeler dile özel geliştirilen yazılımlarla çıkarılabilir. Çalışmamızda Zemberek projesi [13] kullanılarak sözcüklerin türü (POS Tags) modele eklenmiştir.

*4) Yazımsal Öznitelikler (YÖ):*

- **Harfler:** Kelimede geçen harflerin küçük harfte veya büyük harfte yazılmaları da işaretlemeleri etkilediği yapılan çalışmalarda [6] gösterilmiştir. Bu nedenle, kelime tamamen küçük harflerden oluşuyorsa 0, tamamen büyük harflerden oluşuyorsa 1, sadece ilk harf büyükse 2, büyük ve küçük harfler karışık olarak bulunuyorsa 3 modelimize öznitelik olarak verilmiştir.
- **Yüzde İfadeleri:** Kelimede "%" işareti veya "yüzde", "onda" veya "binde" gibi bir ifadenin geçip geçmediği belirtilmiş ve modele öznitelik olarak verilmiştir.
- **Saat İfadeleri:** Kelimenin içerisinde "saat" ifadesinin geçip geçmediği belirtilmiş ve modele öznitelik olarak verilmiştir.
- **Sayısal Öznitelikler:** Kelimenin içerisinde rakam geçip geçmemesinin VİT sistemlerinde başarıyı etkilediği yapılan çalışmalarda [9] gösterilmiştir. Bu nedenle her kelime için içerisinde herhangi bir rakam olup olmaması kontrol edilip modele öznitelik olarak verilmiştir. Ayrıca kelimenin tamamen sayısal değerlerden oluşup oluşmadığı kontrol edilip modele öznitelik olarak verilmiştir.

| | Kişi | Kurum | Yer | Genel |
|---|---|---|---|---|
| Yeniterzi [6] | 89.32 | 83.50 | 92.15 | 88.94 |
| Şeker-Eryiğit [9] Sözlüksüz | - | - | - | 89.70 |
| Şeker-Eryiğit [9] | 92.94 | 88.77 | 92.93 | 91.94 |
| Önerilen Model | 90.85 | 88.16 | 90.71 | 90.15 |

TABLO II: TRAIN3/TEST3 üzerinde doğru tanıma oranları

| | Temel | +KB | +KS | +YÖ | +POS |
|---|---|---|---|---|---|
| Kişi | 88.72 | 91.19 | 92.41 | 93.81 | 94.09 |
| Kurum | 86.59 | 88.98 | 88.85 | 89.44 | 90.08 |
| Yer | 92.88 | 94.01 | 93.97 | 94.56 | 94.22 |
| Tarih | 81.39 | 82.40 | 82.25 | 80.87 | 81.20 |
| Para | 75.79 | 77.55 | 75.00 | 77.89 | 74.47 |
| Yüzde | 90.32 | 89.60 | 91.94 | 94.31 | 95.93 |
| Saat | 75.00 | 75.00 | 80.00 | 81.48 | 81.48 |

TABLO III: TRAIN7/TEST7 üzerinde, önerilen modelde kullanılan özniteliklerin modelin başarısına etkisi

## III. Deneysel Sonuçlar

Geliştirilen sistem şu veri kümelerinde değerlendirilmiştir: (1) TRAIN3 veri kümesiyle model eğitilip TEST3 veri kümesi üzerinde test yapılmıştır. (2) TRAIN7_ALL veri kümesi 10 katlı çapraz doğrulama ile test yapılmıştır. (3) TRAIN7 veri kümesi ile farklı özniteliklerin eklendiği modeller eğitilmiş ve TEST7 veri kümesi üzerinde test yapılmıştır. (4) TRAIN7_ALL veri kümesinde model eğitilip IWT7 veri kümesi üzerinde test yapılmıştır.

Tablo II'de TRAIN3 veri kümesinde daha önce yapılan çalışmaların doğru tanıma başarıları ve önerilen modelin başarısı CoNLL yöntemiyle F1 ölçüsü cinsinden sunulmuştur. Buna göre en başarılı sonuç %91.94 ile Şeker ve Eryiğit [9] tarafından geliştirilen model tarafından elde edilmiştir. Ancak bu sistemde sözlük kullanılmaksızın elde edilen başarı oranı %89.70 olarak sunulmuştur. Geliştirdiğimiz sistemde sözlük olmadığı dikkate alındığında sözlüksüz geliştirilen önceki Şartlı Rastgele Alanlar tabanlı sistemlere göre daha iyi bir başarı oranı yakalandığı sonucuna varılabilir.

Şeker-Eryiğit [9] veri kümesine tarih, saat, para ve yüzde işaretlerini de eklemiştir. Elimizde ilgili çalışmanın sadece eğitim veri seti olduğu için geliştirdiğimiz sistemi 10 katlı çapraz doğrulama kullanarak test ettiğimizde ise ortalama F1 ölçüsü 91.74, standart hata ise $\pm 0.13$ olarak ölçülmüştür. Aynı veri kümesini kullanan [9] çalışmasının başarısının (91.53 $\pm 0.50$) olduğu göz önünde bulundurulursa oldukça yakın bir başarı oranı yakalanmıştır.

Önerilen modelde kullandığımız özniteliklerin sistemin başarısına etkisini ölçmek için sırayla öznitelikler eklenmiş, TRAIN7 veri kümesi ile model oluşturulmuş ve sistemin başarısı TEST7 veri kümesi kullanılarak ölçülmüştür. Tablo III'e göre, modele eklediğimiz özniteliklerin modelin doğru tanıma başarısına olumlu etki yapmıştır. Özellikle kelimenin kökü yerine ilk 4 ve ilk 5 harfine bakılan KB özniteliği sistemin başarısını 2 puana yakın arttırmıştır. Kelimenin büyük, küçük harf içermesine ve saat, yüzde gibi özelliklerine baktığımız YÖ özniteliği sayesinde modelimizde saat ve yüzde isimlerinde olumlu bir artış sağlanmıştır. Modelimizde biçimbilimsel öznitelik olarak kullandığımız sözcüğün türü (POS) ise özellikle kişi ve kurum isimlerini sınıflandırmada modelin başarısına olumlu etki yapmıştır.

Geliştirilen sistemi test ettiğimiz bir diğer veri kümesi

| | Şeker-Eryiğit [9] | Önerilen Model |
|---|---|---|
| Kişi | 67.22 | 58.01 |
| Kurum | 77.17 | 48.18 |
| Yer | 53.87 | 78.66 |
| Tarih | 70.31 | 53.19 |
| Saat | 80.00 | 87.50 |
| Para | 27.45 | 58.46 |
| Yüzde | 50.00 | 80.00 |

TABLO IV: IWT7 üzerinde modellerin doğru tanıma oranları

IWT7'dir. Sözlük kullanmadan geliştirdiğimiz sistemin sonuçları ile bu veri kümesinde daha önce yapılan çalışmanın sonuçları Tablo IV'te sunulmuştur. Tablo IV'teki başarı oranlarına bakıldığında genel başarı oranı olarak en başarılı modelin gerisinde kalındığı görülmektedir. Ancak bu farkın oluşmasında en büyük neden kurum isimlerini tanımada geliştirdiğimiz modelin başarı oranının oldukça geride kalmasıdır. TRAIN7_ALL veri kümesinin 1990'lı yıllara ait veriler içerirken IWT7 veri kümesinde son birkaç yılda kurulan "Facebook", "Twitter" gibi kurumların sıkça geçmesi nedeniyle önerdiğimiz sistem bu varlık isimlerini doğru tanıyamamıştır. Bu gibi durumlarda sözlük kullanımının modelin başarısını arttıracağı düşünülmektedir. Bununla birlikte, kelimelerin ilk harflerine ve son harflerine bakarak geliştirdiğimiz sistem yer, para ve yüzde olarak işaretlenen varlıklarda en iyi sisteme göre daha başarılı olmuştur. Saat olarak işaretlenen varlıklarda da önemli bir ilerleme sağlanmıştır.

Günümüzde VİT sistemlerinde derin öğrenme modelleri sıkça kullanılmaktadır. 2018 yılında [10] tarafından geliştirilen önceden eğitimli (pre-trained) BERT'in doğal dil işleme alanına getirdiği en büyük yenilik transformatörlerin (Transformer) iki yönlü olarak eğitilebilmesinin sağlanmasıdır. Bert, soru cevaplama (SQUAD) ve VİT gibi farklı doğal dil işleme alanlarında kullanılabilmektedir. Çalışmamızda Bert'in temel dil modeli üzerine sınıflandırma katmanı eklenerek [2] Türkçe için VİT ölçümleri yapılmıştır. Çalışmada TRAIN3/TEST3 veri kümeleri kullanılmıştır. Her kelimenin (token) çıktı vektörleri sınıflandırma katmanına girdi olarak kullanılarak kelimeleri kişi, kurum, yer veya hiçbiri olarak işaretlenmiştir. Güneş ve Tantuğ tarafından geliştirilmiş olan Türkçe için derin öğrenme tabanlı çözümde %93.69 F1-ölçüsü raporlanmıştır. Bert ile geliştirilen VİT çözümü için %87.92 F1-ölçüsü elde edilmiştir. Geliştirdiğimiz sisteme [16]'da belirtilen öznitelikler eklenerek ve Bert'in daha kapsamlı dil modelinden faydalanılarak daha yüksek başarı oranı potansiyeli bulunmaktadır.

## IV. Sonuç

Çalışmamızda[3], Şartlı Rastgele Alanlar yöntemiyle haber metinlerinde varlık ismi tanıma problemi üzerinde çalışılmış, modelde kullanılan özniteliklerin sistemin genel başarısına etkisi analiz edilmiş ve kullanılan özniteliklerin genel başarıya olumlu katkı sağladığı görülmüştür. Yapılan ölçümlere göre Türkçe gibi biçimbilimsel olarak zengin ve karmaşık yapıdaki dillerde de konumsal ve yazımsal özellikler ile rekabetçi sonuçlar elde edilebileceği sonucuna varılmıştır. Şartlı Rastgele Alanlar kullanılarak geliştirilen VİT sistemlerinde yaygın bir şekilde sözlük kullanıldığı göz önüne alındığında sözlük kullanılmadan geliştirilen model ile diğer çalışmaların sözlüksüz sistemlerine göre daha başarılı sonuçlar elde edilmiştir. Çalışmanın devamında Şartlı Rastgele Alanlar kullanarak geliştirilen sisteme sözlük eklenerek performansının incelenmesi hedeflenmektedir.

[2]https://github.com/kyzhouhzau/BERT-NER

[3]https://github.com/teghub/ner

## Kaynaklar


[1] B. Sundheim, Overview of results of the MUC-6 evaluation, Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995, pages 13–31, 1995. doi: 10.1145/1072399.1072402.

[2] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In ICML, pages 282–289, 2001.

[3] G. Tür, D. Hakkani-Tür, K. Oflazer, A statistical information extraction system for Turkish, Natural Language Engineering, 9(2):181–210, 2003. doi: 10.1017/S135132490200284X.

[4] D. Küçük, A. Yazıcı, Named entity recognition experiments on Turkish texts, International Conference on Flexible Query Answering Systems, LNCS, 5822:524–535, 2009.

[5] D. Küçük, A. Yazıcı, A hybrid named entity recognizer for Turkish, Expert Systems with Applications, 39(3): 2733-2742, 2012.

[6] R. Yeniterzi, Exploiting morphology in Turkish named entity recognition system, In The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Student Session, pages 105–110, 2011. URL http://www.aclweb.org/anthology/P11-3019.

[7] S. Özkaya, B. Diri, Named entity recognition by conditional random fields from Turkish informal texts, In Proceedings of the IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011), pages 662–665, Antalya, Turkey, 2011.

[8] G. A. Şeker, G. Eryiğit, Initial explorations on using CRFs for Turkish named entity recognition, In Martin Kay and Christian Boitet, editors, COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India, pages 2459–2474, 2012. URL http://aclweb.org/anthology/C/C12/C12-1150.pdf.

[9] G. A. Şeker, G. Eryiğit, Extending a CRF-based named entity recognition model for Turkish well formed text and user generated content of Turkish, Semantic Web Journal, 8, pages 625-542, 2017. doi:10.3233/SW-170253

[10] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[11] C. Sutton, A. McCallum, An Introduction to Conditional Random Fields, Foundations and Trends in Machine Learning, 4:267–373, August 2012.

[12] J. Wijffels, N. Okazaki, Conditional Random Fields for Labelling Sequential Data in Natural Language Processing based on CRFsuite: a fast implementation of Conditional Random Fields (CRFs). 2018. https://github.com/bnosac/crfsuite

[13] A. A. Akın, Zemberek-NLP, January 22, 2019. Retrieved February 04, 2019, from https://github.com/ahmetaa/zemberek-nlp

[14] S. Tatar, I. Cicekli, Automatic rule learning exploiting morphological features for named entity recognition in Turkish, Journal of Information Science,37(2):137–151,2011.

[15] H. Demir, A. Özgür, Improving named entity recognition for morphologically rich languages using word embeddings, In 13th International Conference on Machine Learning and Applications, ICMLA 2014, pages 117–122, Detroit, MI, USA, December 3-6, 2014. doi: 10.1109/ICMLA.2014.24

[16] A. Güneş, A. C. Tantuğ, Turkish named entity recognition with deep learning, 26th Signal Processing and Communications Applications Conference (SIU), pages 1-4, 2018. doi:10.1109/siu.2018.8404500.

[17] S. Cucerzan, D. Yarowsky, Language independent named entity recognition combining morphological and contextual evidence, In Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, pp. 90-99, 1999.