Türkçe için Wikipedia Tabanlı Varlık İsmi Tanıma Sistemi

Doğan KÜÇÜK, Nursal ARICI
Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü
(Geliş / Received : 18.12.2015 ; Kabul / Accepted : 26.03.2016)

ÖZ

Varlık ismi tanıma, doğal dil işleme araştırma alanında bir problemdir ve genellikle doğal dildeki metinlerden kişi, yer ve kurum isimlerinin otomatik çıkarılması olarak tanımlanmaktadır. Bu çalışmada, Türkçe için geliştirdiğimiz Wikipedia tabanlı bir varlık ismi tanıma sistemi tanıtılmıştır. Wikipedia gibi internet kullanıcıları tarafından oluşturulan kaynakların varlık ismi tanıma gibi konular için oldukça faydalı oldukları bilinmektedir. Öncelikle, Türkçe Wikipedia'dan otomatik olarak geniş bir insan ismi listesi derlenmiştir. Daha sonra, bu liste ile birlikte yine Türkçe Wikipedia'dan ve Türkçe için kural-tabanlı bir sistemden elde edilmiş kişi, yer ve kurum ismi listelerini de kullanarak Türkçe için Wikipedia-tabanlı bir varlık ismi tanıma sistemi geliştirilmiştir. Sistemimiz değişik veri kümeleri üzerinde test edilerek değerlendirilmiş ve umut verici sonuçlar elde edilmiştir. Türkçe metinlerde bilgi çıkarımı üzerinde yapılmış kısıtlı sayıda çalışma olduğundan bizim sistemimiz bu konuda önemli bir katkı teşkil etmektedir.

Anahtar Kelimeler: Varlık ismi tanıma, bilgi çıkarımı, Türkçe, otomatik metin işleme

Wikipedia-based Named Entity Recognition System for Turkish

ABSTRACT

Named entity recognition is a problem in the research area of natural language processing and is usually defined as the automatic extraction of the names of people, locations, and organizations in natural language texts. In this study, a Wikipedia-based named entity recognition system for Turkish is introduced. It is well-known that resources like Wikipedia, which are created by internet users, are considerably important for topics like named entity recognition. We have first automatically compiled a large list of person names from Turkish Wikipedia. Then, we have developed a Wikipedia-based named entity recognition system for Turkish which utilizes this large list with other lists of person, location and organization named obtained from Turkish Wikipedia and a former rule-based named entity recognizer for Turkish. We have evaluated our system on different types of datasets and obtained promising results. Our system is a significant contribution to information extraction on Turkish texts since there are limited number of related studies carried out so far.

Keywords: Named entity recognition, information extraction, Turkish, automatic text processing

1. GİRİŞ (INTRODUCTION)

Türkçe'de doğal dil işleme ve özellikle bilgi çıkarımı konusunda yapılan çalışmalar; İngilizce, Almanca ve Fransızca gibi diğer dillerdeki metinler üzerinde yapılan çalışmalara göre sayısal olarak ve kapsam olarak oldukça sınırlıdır. Bununla beraber, her gün Türkçe dilinde de hatırı sayılır miktarda doküman oluşturulmaktadır ve bu dokümanlardaki bilgilerin otomatik olarak çıkarılması otomatik bilgi erişimi, soru cevaplama ve özetleme gibi sistemler için oldukça fazla önem arz etmektedir.

Varlık ismi tanıma, doğal dildeki metinlerden insan, yer ve kurum ismi gibi isimleri çıkarma ve bunları türlerine göre sınıflandırma şeklinde tanımlanmaktadır [1]. Türkçe metinler üzerinde de İngilizce gibi diller üzerinde yapılan çalışmalarla karşılaştırılınca kısıtlı olsa da yapılmış birtakım çalışmalar vardır. Türkçe üzerindeki varlık ismi tanıma çalışmalarının bilgimiz dâhilindeki ilkinde dilden bağımsız bir sistem sunulmuş ve bu sistem Türkçe ile birlikte Hintçe, İngilizce, Rumence ve Yunanca dilleri üzerinde de denenmiştir [2]. Bu konuda yapılan ilk çalışmalardan bir diğerinde Saklı Markov Modelleri (Hidden Markov Models) kullanılmıştır [3]. Daha sonra yapılan bir çalışmada, Türkçe için kural-tabanlı bir varlık ismi tanıma sistemi tanıtılmıştır [4]. Sonrasında farklı kural-tabanlı varlık ismi tanıma sistemleri de sunulmuştur [5]. Bir diğer çalışmada, Türkçe'de varlık isimlerini tanımak için otomatik olarak kural öğrenen bir sistem tanıtılmıştır [6]. Şartlı Rasgele Alanlar (Conditional Random Fields) adlı makine öğrenmesi yönteminin diğer dillerde varlık ismi tanımada da başarılı olduğu gösterilmiştir [7]. Bu nedenle, Türkçe metinlerde varlık ismi tanımada da Şartlı Rastgele Alanlar yöntemini kullanan çalışmalar olmuştur [8,9,10]. Daha önce sunulmuş kural-tabanlı sistemini, basit bir öğrenme yöntemi olan ezber yöntemiyle (rote learning) birleştirerek geliştirilmiş melez (hybrid) bir varlık ismi tanıma sistemi de

** Sorumlu Yazar (Corresponding Author)*
e-posta: nursal@gazi.edu.tr
Digital Object Identifier (DOI) : 10.2339/2016.19.3 325-332

mevcuttur [11]. İlgili bir diğer güncel çalışmada da Türkçe metinlerde varlık ismi tanıma için yapay sinir ağları kullanılmıştır [12].

Yukarıda bildirilen çalışmaların tümü çoğunlukla dilbilgisi ve yazım hatalarının bulunmadığı haber metinleri için geliştirilmiş ve bu türdeki metinler üzerinde değerlendirilmişlerdir. Yakın zamanda ise, sosyal medya metinlerinin otomatik olarak işlenmeleri gerekliliğinin ortaya çıkmasından dolayı tweet metinleri gibi çoğunlukla dilbilgisi ve yazım hatalarının bulunduğu metinler üzerinde varlık ismi tanıma çalışmalarına başlanmıştır [13]. Haber metinleri için geliştirilmiş veya bunlar üzerinde eğitilmiş sistemlerin direk olarak tweet metinlerinde oldukça kötü değerlendirme sonuçları aldıkları görülmüştür ve dolayısıyla bu yeni metin türü için kendine özgü sistemlerin geliştirilmesi gerektiği vurgulanmıştır [13]. Bu nedenle, yakın dönemde Türkçe tweet'ler üzerinde de varlık ismi tanıma değerlendirmeleri ile bu metinler üzerinde başarım değerlendirme sonuçlarını artırmaya yönelik çalışmalar yapılmıştır [14,15,16].

Bu çalışmada, Türkçe varlık ismi tanıma için Türkçe Wikipedia [17] kaynaklarının kullanılması sağlanmıştır. Türkçe varlık ismi tanımada Wikipedia'dan yararlanılması konusunda bilgimiz dâhilinde literatürde sadece bir çalışmaya [18] rastlanmıştır. Bu çalışmada Türkçe Wikipedia başlıklarının yirmide biri varlık isimleriyle elle işaretlenmiş, ardından bu küme eğitim kümesi olarak kullanılarak en yakın k komşu (k-nearest neighbors) algoritması kullanılarak kişi, yer ve kurum isimleri elde edilmiştir. Bizim çalışmamız bu çalışmadan hem yöntem olarak hem de elde edilen kaynakların boyutları açısından farklıdır. Ayrıca, bizim çalışmamızda bu çalışmada [18] elde edilen kaynaklar da kullanılmıştır.

Öncelikle Wikipedia'dan otomatik olarak öğrenilen büyük boyutlu bir kişi ismi listesi kullanılarak Wikipedia tabanlı bir kişi ismi tanıma sistemi geliştirilmiştir. Daha sonra ise, bu liste diğer Wikipedia'dan öğrenilmiş kişi, yer ve kurum ismi listeleri [18] ve önceden geliştirilmiş olan kural-tabanlı sistemin [4] sözlüksel kaynakları ile beraber kullanılarak Türkçe Wikipedia tabanlı varlık ismi tanıma sistemi geliştirilmiştir. Sistemlerimiz daha önceki sistemler tarafından da kullanılan çok çeşitli türlerdeki veri kümeleri üzerinde değerlendirilmiş ve umut verici sonuçlar elde edilmiştir.

Çalışmamızın ilgili literatüre katkıları aşağıda maddeler halinde sıralanmıştır:

1. Çalışmamız kapsamında tam otomatik bir yöntemle Türkçe Wikipedia'dan hatırı sayılır boyutta bir kişi ismi listesi oluşturulmuş ve bu kaynağı kullanarak yüksek doğrulukta sonuçlar veren bir Türkçe kişi ismi tanıma sistemi geliştirilmiştir. Söz konusu kişi ismi tanıma sistemi Türkçe Wikipedia'yı varlık ismi tanıma kaynağı olarak kullanan öncül çalışmalardandır.
2. Wikipedia'dan otomatik olarak farklı bir çalışma kapsamında [18] çıkarılan kişi, yer ve kurum ismi listeleriyle, önceden sunulan kural-tabanlı bir sistemin kişi, yer ve kurum ismi listeleri de bu sistemin bünyesine katılarak Türkçe için Wikipedia tabanlı tam bir varlık ismi tanıma sistemi geliştirilmiştir.
3. Her iki sistem başarımları da çeşitli boyut ve özelliklere sahip farklı veri kümeleri üzerinde değerlendirilmiş ve umut verici sonuçlar elde edilmiştir. Bu deneme veri kümelerinin türleri ve toplam boyutları dikkate alındığında, mevcut çalışmamızın şu anda kadar Türkçe varlık ismi tanıma konusunda yapılmış en kapsamlı değerlendirmeyi içerdiği söylenebilir.

2. WIKIPEDIA TABANLI KİŞİ İSMİ TANIMA SİSTEMİ (WIKIPEDIA BASED PERSON NAME RECOGNITION SYSTEM)

Türkçe metinlerde kişi ismi tanıma probleminin, yer ve kurum ismi gibi diğer varlık isimlerinin tanınmalarına göre daha zor bir problem olduğu bilinmektedir. Var olan birçok çalışmada, kişi ismi tanıma performansının diğer türlerin tanınma performanslarından daha düşük olduğu belirtilmektedir.

Türkçe metinlerde kişi ismi tanımada yaşanan zorluğun en büyük nedeni, yaygın kişi isimlerinin Türkçe'deki nesne isimleriyle eş sesli olmasıdır. Örnek olarak; *Onur, Bilge, Ela, Yağmur, Damla, Deniz, Savaş, Barış, Kartal, Şahin* gibi isimler yaygın kişi isimlerindendir ancak aynı zamanda Türkçe'deki nesne isimleri ve sıfatlarla da eş seslidirler. Bunun dışında; *Nisan, Eylül, Ekim, Kasım* gibi yaygın kişi isimleri de Türkçe'deki ay isimleriyle eş seslidirler.

Dolayısıyla, Türkçe'deki kişi isimlerini içeren uzun bir listedeki elemanların verilen bir metinde taranarak art arda en az bir kelimeden oluşan çakışmaların kişi ismi olarak işaretlenmesi fazladan yanlış işaretlemelere neden olmaktadır.

Wikipedia gibi internet kullanıcılarının ortak olarak düzenledikleri bilgi tabanlarında haberlerde geçen birçok ünlü kişiye ait başlıklar mevcuttur. Wikipedia'nın başlık isimleri ve makalelerinin metinleri açık olarak mevcuttur ve araştırma amaçlarıyla kullanılabilmektedir.

Türkçe Wikipedia'ya Java programlama dili ile erişebilmek için bu dili kullanan ve erişimi sağlayan JWPL (Java Wikipedia Library) adlı kütüphane [19] kullanılmıştır.

Öncelikle, Türkçe kişi isimlerine karşılık gelen makale başlıklarının genellikle; X bir yıl olmak üzere "*X doğumlular*" ve "*X yılında ölenler*" desenlerine sahip Wikipedia kategorileri altında yer aldıkları gözlemlenmiştir. Örneğin "*1900 doğumlular*" kategorisi altında 1900 yılında doğmuş olan ünlü kişilere ait Wikipedia makaleleri yer almaktadır. Benzer şekilde "*1900 yılında ölenler*" kategorisi altında 1900 yılında

ölmüş ünlü kişilerin makaleleri yer almaktadır. Bu iki kategoriye ait örnek bir Türkçe Wikipedia kategori sayfası aşağıda verilmiştir:

https://tr.wikipedia.org/wiki/Kategori:1926_do%C4%9Fumlular

(1926 yılında doğan kişilerin sayfalarının bağlantılarının bulunduğu kategori sayfası)

https://tr.wikipedia.org/wiki/Kategori:1998_y%C4%B1l%C4%B1nda_%C3%B6lenler

(1998 yılında ölen kişilerin sayfalarının bağlantılarının bulunduğu kategori sayfası)

Bizim önerdiğimiz desen-tabanlı yaklaşım X değeri 0 ile 2014 tarihleri arasında değişecek şekilde; "*X doğumlular*" ve "*X yılında ölenler*" desenlerine uyan toplam 4.030 kategorinin altındaki makale başlıklarını otomatik olarak tespit etmiştir. Bu işlem sonucunda yaklaşık 42.500 adet Türkçe ve yabancı kişi ismi tespit edilmiştir. Bu liste içerisindeki tek kelimelik isimler diğer varlık ismi türleriyle, nesne veya sıfatlarla eş sesli olabileceğinden, listeden çıkarılmıştır.

Özellikle sosyal medya metinleri gibi metinlerde daha sık olmak üzere; Türkçe'deki ç, ğ, ı, ö, ş, ü yerine sırasıyla c, g, i, o, s, u harfleri kullanılabilmektedir. Örneğin, "*Barış Manço*" yerine "*Baris Manco*" yazılabilmektedir. Verilen bir Türkçe metin girdisinde, bu şekilde yazılan kişi isimlerinin de tespit edilebilmesi için; elde edilen kişi isimlerinden yukarıda belirtilen ç, ğ, ı, ö, ş, ü harflerinden en az birini içeren isimlerin her biri için bu harflerin sırasıyla karşılık gelen harflerle (c, g, i, o, s, u) değiştirildikleri halleri oluşturulmuş ve bunlar da kişi ismi listesine eklenmiştir. Bu işlemden sonra kişi ismi listesindeki isim sayısı yaklaşık 55.000 olmuştur.

Yukarıda tanıtılan kişi ismi listesi kullanılarak aşağıdaki özelliklere sahip bir Türkçe kişi ismi tanıma sistemi Java programlama diliyle geliştirilmiştir:

1. Sistem, verilen girdi metninde sırasıyla art arda gelen 7 kelimelik tüm dizileri, sonra 6, sonra 5, sonra 4, sonra 3, sonra da 2 kelimelik dizileri kişi ismi listesinde aramakta, eğer bulursa bunları Message Understanding Conference (MUC) [20] konferans serisinde varlık isimlerini işaretlemek için kullanılan ENAMEX etiketiyle işaretlemektedir. Örneğin iki kelimelik diziler kontrol edilirken "*Barış Manço*" dizisi bulunduğunda bu dizi aşağıdaki şekilde işaretlenecektir.

<ENAMEXTYPE="PERSON">Barış Manço</ENAMEX>

2. Sistem iç içe işaretlemeler yapmamakta, yukarıda belirtilen sıra takip edildiği için en fazla 7 kelimeden oluşmak üzere en uzun diziyi işaretlemektedir. Örneğin, eğer listede hem "*Ali Sami Yen*", hem de "*Sami Yen*" elamanları varsa ve girdi metinde "*Ali Sami Yen*" geçiyorsa, sistem bu kısmı aşağıdaki şekilde işaretleyerek dönmektedir.

<ENAMEX TYPE="PERSON">Ali Sami Yen</ENAMEX>

Bir metinde bir varlık ismi yer aldıktan bir süre sonra yine o isme metin içinde yer verilmesi sık karşılaşılan bir durumdur. Bu şekilde metnin farklı kısımlarında gözlenen varlık isimlerinin aynı şekilde işaretlenmeleri gereklidir. Örneğin daha önce de verilen ve ODTÜ Türkçe Derlemi'nden [21] alınan aşağıdaki metinde koyu renkle gösterilen kişi: isimleri aynı gerçek kişiye karşılık gelmektedir

***Osman Bölükbaşı**, uzun kollarını iki tarafa açmış, ikişer üçer kişiyi kollarına almış olarak, gece boyunca bir şeyler anlattı durdu. **Bölükbaşı** anlattıkça, gruptakiler yüksek sesle kahkahalar atıyorlardı.*

Yukarıdaki örnekte koyu ile yazılmış bu varlık isimleri arasındaki ilişkiye benzer ilişkilere yerel olmayan bağımlılıklar (non-local dependencies) adı verilmektedir ve varlık ismi tanıma işlemi sırasında önemli bir ipucu olarak kullanılabilmektedirler [22]. Örneğin, bu amaçla en son metindeki 1000 kelime için yapılan sınıflandırmaların kaydedildiği ve kullanıldığı *genişletilmiş tahmin tarihi (extended prediction history)* adlı yöntemin başarılı sonuçlar verdiği gözlemlenmiştir [22].

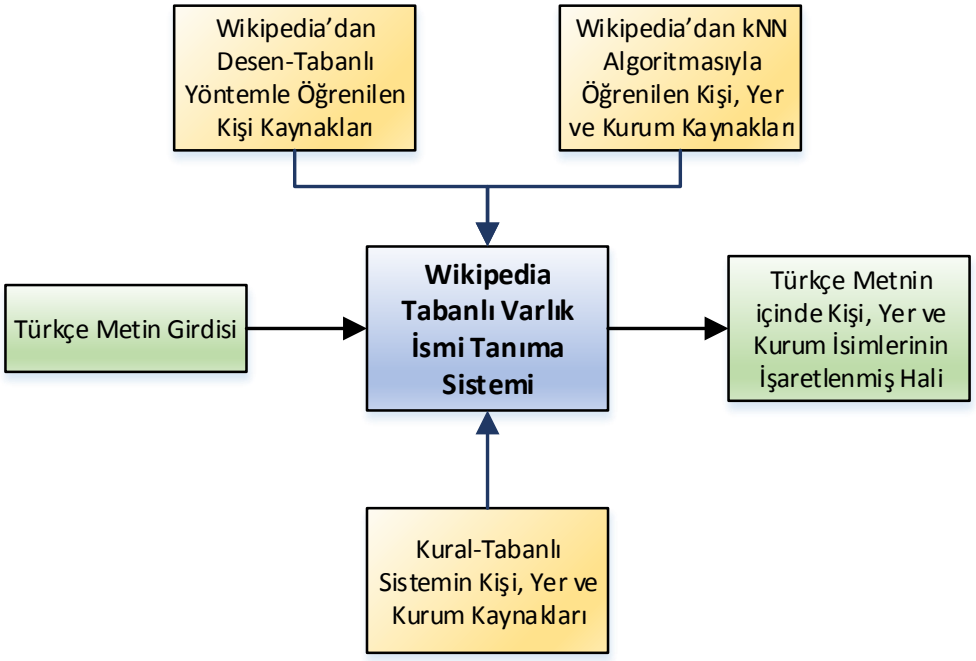
Biz de geliştirdiğimiz Türkçe kişi ismi tanıma sistemimizin kapsamını, yani tanıyabildiği kişi ismi sayısını, arttırabilmek amacıyla yerel olmayan bağımlılıkları da kullandık. Sistemimizin mevcut hali tek kelimeleri isim adayı olarak almıyordu ve dolayısıyla işaretlemiyordu. Yukarıdaki örnek metinde sadece "*Osman Bölükbaşı*" dizisi kişi ismi olarak sistem tarafından işaretlenmekte, ikinci cümledeki "*Bölükbaşı*" kelimesi sistem tarafından işaretlenmemektedir. Bu yeni sürümünde ise, sistem bir kişi ismini tanıdıktan sonra bu kişinin soyadı bir listeye ekleniyor ve bundan sonra gelen tek kelimeler de bu listede mevcutlarsa kişi ismi olarak etiketleniyorlar. Sistemin bu yeni sürümü, yukarıdaki örnek metin üzerinde çalıştırıldığında, önce "*Osman Bölükbaşı*" dizisini kişi ismi olarak işaretlemekte, ardından "*Bölükbaşı*" kelimesini soyisim listesine eklemekte, sıra ikinci cümledeki "*Bölükbaşı*" kelimesine geldiğinde bu kelime listede mevcut olduğundan olması gerektiği gibi kişi ismi olarak işaretlemektedir.

3. WIKIPEDIA TABANLI VARLIK İSMİ TANIMA SİSTEMİ (WIKIPEDIA BASED NAMED ENTITY RECOGNITION SYSTEM)

Bir önceki bölümde tanıttığımız kişi ismi tanıma sistemimizin Wikipedia kaynağını; daha önce tanıtılmış olan kural-tabanlı bir sistemin [4] ve önceki bir çalışmada [18] Wikipedia'dan elde edilen diğer bazı kaynaklarla kullanarak Türkçe için tam bir varlık ismi tanıma sistemini Java programlama dili ile geliştirdik.

Şekil 1'de geliştirdiğimiz sistemin genel mimarisi sunulmuştur.

da Jaccard benzerliği formülü kullanılmıştır [23]. Bu benzerlikte, benzerliği hesaplanacak ifadeler A ve B adlı iki kelime dizisi olarak alınır ve dizilerdeki kelimelerin kesişim kümesinin dizilerin birleşim kümesine oranı



Şekil 1. Wikipedia Kaynaklarını Kullanan Türkçe Varlık İsmi Tanıma Sistemi (Named Entity Recognition System for Turkish Utilizing Wikipedia Resources)

Sunulan bu yeni varlık ismi tanıma sisteminin bilgi kaynaklarının detayları aşağıda verilmiştir.

1. *Wikipedia'dan Desen Tabanlı Yöntemle Öğrenilen Kişi Kaynakları:* Bu kaynaklar bir önceki bölümde anlatılan ve desen-tabanlı bir yöntemle Wikipedia'dan otomatik olarak elde edilmiş olan yaklaşık 55,000 kişi isminden oluşan kişi ismi listesidir.

Wikipedia'dan kNN Algoritmasıyla Öğrenilen Kişi, Yer ve Kurum Kaynakları: Bu kaynaklar, ilgili çalışmada [18] açıklanan yöntemle öğrenilen kişi, yer ve kurum isimleri listeleridir. İlgili çalışmada, Türkçe Wikipedia'daki başlıkların yirmide biri elle kişi, yer veya kurum ismi olarak işaretlenmiş, daha sonra kalan başlıklar bu işaretlenmiş küme eğitim kümesi olarak kullanılarak k-nearest-neighbor (k-en yakın komşu - kNN) algoritması kullanılarak sınıflandırılmaya çalışılmıştır. Çalışmada, sınıflandırma her bir yeni aday için 0.2'den büyük benzerliğe sahip en az 5 komşu varsa, bunlardan en fazla elemana sahip sınıfa adayı atamak şeklindedir. Bu şartlar sağlanmıyorsa sınıflandırma yapılmamış böylelikle düşük güvenlikli kaynaklar oluşturmaktan kaçınılmıştır. Benzerlik olarak aşağıdaki formülle hesaplanır:

$|(A \cap B)/(A \cup B)|$

Kural-tabanlı Sistemin Kişi, Yer ve Kurum Kaynakları: Söz konusu kural-tabanlı sistem [4] çalışmasında tanıtılmıştır. İnternette açık olarak mevcut kaynaklardan elle toplanmış kişi, yer ve kurum isimleri listelerini içermektedir. Kural-tabanlı sistem bu kaynaklarla birlikte bazı kurallar da kullanmakta, böylelikle bu kaynaklarda mevcut olmayan isimler de girdi olarak verilen metinden çıkarılabilmektedir.

Söz konusu kaynaklarda kesişen elemanlar bulunabileceğinden, bu yinelenen elemanların son sistemin kullanacağı kaynaklardan silinmesi için kısa bir program yazılmış ve yinelenen elemanların sayısı bire düşürülmüştür. Çizelge 1'de sistemin kaynaklarının bu silinme işlemi öncesi ve sonrası sayıları verilmektedir.

Çalışmalarımız kapsamında geliştirdiğimiz Türkçe kişi ismi tanıma ve Türkçe varlık ismi tanıma sistemlerimizi yedi farklı veri kümesi üzerinde test ettik. Söz konusu veri kümeleri ayrıca, Türkçe varlık ismi tanıma konusunda yapılmış literatürdeki çalışmalarda da deneme veri kümesi olarak kullanılmış kümelerdir. Bu

Çizelge 1. Wikipedia-tabanlı Türkçe Varlık İsmi Tanıma Sistemi'nin Kaynaklarının Dağılımı (The Distribution of the Resources of the Wikipedia Based Named Entity Recognition System for Turkish)

	Wikipedia'dan Desen-tabanlı Yöntemle Elde Edilen Kişi İsmi Listesinde	*Wikipedia'dan kNN ile Otomatik Öğrenilen Kaynaklarda*	*Türkçe için Kural-tabanlı Sistemin Kaynaklarında*	***TOPLAM (Yinelenenlerle Birlikte)***	***TOPLAM (Yinelenenler Silindikten Sonra)***
Kişi İsmi Sayısı	54872	17003	67	**71942**	**60832**
Yer İsmi Sayısı	0	4893	5465	**10358**	**10185**
Kurum İsmi Sayısı	0	2388	1043	**3431**	**3414**
TOPLAM	**54872**	**24284**	**6575**	**85731**	**74431**

veri kümelerinden üç tanesi genel haber metinlerinden, bir tanesi finans haberi metinlerinden iki tanesi bir sosyal medya metni türü olan tweet'lerden, bir tanesi de tarihi metinlerden oluşmaktadır. Dolayısıyla, veri kümeleri hem haber metinleri gibi hatasız metin türündeki verileri, hem de tweet'ler gibi yazım ve dilbilgisi hataları da içerebilen metin türlerinden oluşmaktadır.

Veri kümeleriyle ilgili istatistiksel bilgiler Çizelge 2'de verilmiştir. Çizelge 2'de görüldüğü üzere, veri kümelerinde yaklaşık 343.000 kelime bulunmakta ve toplam 22.941 adet de işaretlenmiş varlık ismi bulunmaktadır. Diğer kümelerden farklı olarak, Finans Haberleri Kümesi'nde yer isimleri işaretlenmemiş, sadece kişi ve kurum isimleri işaretlenmiştir.

Türkçe metinlerde varlık ismi tanıma konusunda yapılmış çalışmaların büyük bir bölümü sadece haber metni kümeleri üzerinde denenmiştir. Bununla birlikte, az sayıda da olsa farklı metin türleri üzerinde denemeler yapmış çalışmalar mevcuttur. Ancak, hem bu şekilde farklı türde veri kümeleri üzerinde denemelerin yapıldığı hem de söz konusu deneme veri kümelerinin bizim kullandığımız kümelerin toplam büyüklüğüne (hem kelime sayısı, hem de işaretlenmiş varlık ismi sayısı olarak) eriştiği herhangi bir çalışma bilgimiz dâhilinde değildir. Bu nedenlerle, kullandığımız veri kümelerinin büyüklükleri ve çeşitliliği dikkate alınırsa, bu çalışmamız Türkçe'de varlık ismi tanıma konusundaki en kapsamlı değerlendirmeyi sunmaktadır.

Çalışmalarımızın değerlendirilmesinde de varlık ismi tanıma sistemlerini değerlendirmede yaygın olarak kullanılan precision (P), recall (R) ve F-Measure (F) ölçütleri kullanılmıştır. Aşağıda bu ölçütlerin formülleri yüzde şeklinde verilmiştir:

*Precision (P) = (Sistemin doğru tespitleri) * 100 / (Sistemin tüm sonuçları)*

*Recall (R) = (Sistemin doğru tespitleri) * 100 / (Sistemin tanıması gereken varlık isimleri)*

*F-Measure (F) = (2 * Precision * Recall) / (Precision + Recall)*

Bizim çalışmamızda; bu ölçütler, bir varlık isminin ancak hem sınırları hem de türü cevap kümesindekiyle tamamen aynı ise onu doğru kabul edecek şekilde ve dolayısıyla, literatürde sıklıkla kullanılan CoNLL konferans serisinde [24] verildiği şekilde hesaplanmaktadır.

Çizelge 2. Değerlendirmede Kullanılan Veri Kümeleriyle ilgili Bilgiler (Information on the Datasets Used During the Evaluation)

			Varlık İsmi Sayıları				
Veri Kümesi Adı	*Kullanıldığı Çalışma*	*Kelime Sayısı (Yaklaşık)*	*Kişi İsmi*	*Birden Fazla Kelime İçeren Kişi İsmi Sayısı*	*Yer İsmi*	*Kurum İsmi*	***TOPLAM***
Haber Veri Kümesi-1	Küçük ve Yazıcı (2009)	20.000	398	169	571	456	**1.425**
Haber Veri Kümesi-2	Tür ve ark. (2003), Şeker ve ark. (2012)	48.000	1.596	743	1.091	863	**3.550**
Haber Veri Kümesi-3	Küçük (2015)	100.000	3.288	1.488	2.470	3124	**8.882**
Finans Haberleri Kümesi	Küçük (2015)	84.000	1.115	468	0	4.521	**5.636**
Tarihi Metin Kümesi	Küçük ve Yazıcı (2009)	20.000	387	217	585	122	**1.094**
Tweet Kümesi-1	Küçük ve ark. (2014)	21.000	457	149	282	241	**980**
Tweet Kümesi-2	Çelikkaya ve ark. (2013)	50.000	774	190	191	409	**1.374**
TOPLAM		**343.000**	**8.015**	**3.424**	**5.190**	**9.736**	**22.941**

Türkçe için Wikipedia'dan desen tabanlı yöntemle elde ettiğimiz listeyi kullanan ve 2. bölümde açıklanan kişi tanıma sisteminin başarım değerlendirme sonuçları, yukarıda verdiğimiz ölçütler kullanılarak aşağıda Çizelge 3'te verilmiştir.

Çizelgede ikinci kolonda verilen precision değerleri incelenirse; bu değerlerin oldukça yüksek olduğu, yani sistemimizin precision türünde başarımının yüksek olduğu görülmektedir. Sistemin precision değerleri %92,11 ile %98.11 arasında değişmektedir. En yüksek precision değeri *Haber Veri Kümesi-1* üzerinde, en düşük precision değeri ise *Tarihi Metin Kümesi* üzerinde elde edilmiştir. Bu başarılı sonuçların bir nedeni Wikipedia'dan otomatik elde edilen isimlerin oldukça güvenilir olmaları, bir nedeni de hatalı işaretlemelere neden olacak tek kelimelik isimlerin listede bulunmamasıdır, yani sistemde kullanılmadan önce listeden temizlenmiş olmalarındandır.

Çizelge 3. Wikipedia Tabanlı Kişi İsmi Tanıma Sistemi'nin Başarım Değerlendirme Sonuçları (Performance Evaluation Results of the Wıkipedia Based Person Name Recognition System)

Veri Kümesi Adı	*P*	*R*	*F*	*R (Sadece Birden Fazla Kelimelik İsimler için)*	*F (Sadece Birden Fazla Kelimelik İsimler için)*
Haber Veri Kümesi-1	98,11	26,13	41,27	61,54	75,64
Haber Veri Kümesi-2	95,54	20,11	33,23	43,2	59,5
Haber Veri Kümesi-3	97,25	21,5	35,22	47,51	63,84
Finans Haberleri Kümesi	95,16	10,58	19,05	25,21	39,86
Tarihi Metin Kümesi	92,11	9,04	16,47	16,13	27,45
Tweet Kümesi-1	96,67	12,69	22,44	38,93	55,5
Tweet Kümesi-2	95,45	5,43	10,27	20,69	34,01

Çizelgenin üçüncü kolonundaki recall değerlerininse oldukça düşük olduğu görülmektedir. Yani sistemimizin kullandığı kişi adları listesinde bulamadığından, tanıyamadığı birçok kişi ismi kalmıştır. Bu recall değerlerinin %5,43 ile %26,13 arasında değiştiği gözlemlenmiştir. En yüksek recall değeri yine *Haber Veri Kümesi-1* üzerinde, en düşük recall değeri ise *Tweet Kümesi-2*'de elde edilmiştir. Özellikle genel haber verisi kümelerinde (*Haber Veri Kümesi-1*, *Haber Veri Kümesi-2* ve *Haber Veri Kümesi-3*), recall değerlerinin daha yüksek olduğu gözlenmektedir. Bunun önemli bir nedeni, haber metinlerinde ismi geçen kişilerin bir kısmının kendilerine ait Wikipedia sayfalarının bulunması, dolayısıyla da bizim sistemimizin otomatik olarak elde edip kullandığı isim listesinde yer almalarıdır.

Bununla birlikte, sistemimiz tek kelimelik isimleri (listesinde bu boyutta isimler bulunmadığından) hiç dikkate almadığı için recall değerlerinin Çizelge 3'ün beşinci kolonunda olduğu gibi birden fazla kelimeden oluşan isimler üzerinden hesaplanması ve dikkate alınması daha doğru olacaktır. Çizelgedeki bu recall değerlerine baktığımızda; bu değerlerin normal recall değerlerine göre oldukça yüksek olduğu görülmektedir. Bu yeni recall değerleri %16,13 ile %61,54 arasında değişmektedir ve çizelgenin son kolonundaki F-Measure değerlerinin hesaplanmasında da bu yeni recall değerleri kullanılmaktadır. Bu son kolondaki F-Measure değerleri de çizelgenin dördüncü kolonundaki F-Measure değerlerine göre oldukça yüksektir.

Özetlemek gerekirse, kişi tanıma sistemimizin tanıdığı isimlerin çoğunun doğru şekilde tanındığı tüm veri kümeleri üzerinde %92 ile %98 civarı precision değerleriyle kendini göstermiştir. Daha önceden geliştirilmiş olan kural-tabanlı sistemin [4] son halinin *Haber Veri Kümesi*-1 üzerinde kişi ismi tanımada precision değeri %57.4 olarak bulunmuştur. Dolayısıyla sistemimizin kişi ismi tanımada recall değerleri düşük olsa da precision değerlerinin daha önceki sistemlerle karşılaştırıldığında oldukça yüksek olduğu gözlemlenmiştir.

Çalışmamızın 2. bölümünde kişi ismi tanıma sistemimizin yerel olmayan değişkenleri de kullanarak kapsamı-nın arttırılabileceği belirtilmiş ve başlangıçtaki kişi ismi tanıma sisteminin kişi ismi listesi ile birlikte bu yerel olmayan bağımlılıkları da kullanan bir sürümü oluşturul-muştur. Bu sürümünün başarım değerlendirme sonuçları Çizelge 4'te, çizelgenin 2., 3. ve 4. sütunlarında verilmiş-tir.

Çizelgenin 5., 6. ve 7. sütunlarında da bu yeni elde edilen sonuçlarla, sistemin ilk sürümünün yani yerel olmayan bağımlılıkları kullanmayan ilk halinin Çizelge 5.2'de verilen sonuçları arasındaki farklar verilmiştir.

Çizelgedeki sonuçlar incelendiğinde ilk dört veri kümesi için recall değerlerinin oldukça fazla artış gösterdiği gözlenmiştir. Bu sonuçlar; yerel olmayan bağımlılıkların dikkate alınmasının kişi ismi tanıma performansını oldukça arttıracağını ortaya koymuştur. Tüm veri kümelerindeki değerlendirmelerde, sistemin precision değerlerinin çok az azaldığı görülmüştür. Bu

Çizelge 4. Wikipedia Tabanlı Kişi İsmi Tanıma Sistemi'nin Yerel Olmayan Bağımlılıkları da Kullanan Sürümünün Başarım Değerlendirme Sonuçları (Performance Evaluation

Veri Kümesi Adı	*P*	*R*	*F*	*ΔP*	*ΔR*	*ΔF*
Haber Veri Kümesi-1	96,9	31,41	47,44	↓ 1,21	↑ 5,28	↑ 6,17
Haber Veri Kümesi-2	93,27	35,59	51,52	↓ 2,27	↑ 15,48	↑ 18,29
Haber Veri Kümesi-3	93,96	39,75	55,87	↓ 3,29	↑ 18,25	↑ 20,65
Finans Haberleri Kümesi	97,23	28,34	43,89	↓ 2,07	↑ 17,76	↑ 24,84
Tarihi Metin Kümesi	87,5	9,04	16,39	↓ 4,61	↔	↓ 0,08
Tweet Kümesi-1	93,94	13,57	23,71	↓ 2,73	↑ 0,88	↑ 1,27
Tweet Kümesi-2	93,33	5,43	10,25	↓ 2,12	↔	↓ 0,02

da beklenen bir sonuçtur, çünkü sisteme dâhil edilen tek kelimelik soyadları listesindeki elemanlardan bazıları nesne isimleri, sıfatlar ve fiiller gibi kişi ismi olmayan kelimelerle eş sesli olabilmekte, bunlar da az da olsa sistemin bu işaretlenmemesi gereken kelimeleri de kişi ismi olarak işaretlemesine neden olmaktadır.

Çalışmamızın 3. bölümünde açıkladığımız çoğunluğu Wikipedia'dan olmak üzere çok kaynaklı varlık ismi tanıma sistemimizin başarım değerlendirme sonuçları da yine aynı ölçütler kullanılarak Çizelge 5'te verilmiştir.

Çizelge 5. Wikipedia-tabanlı Türkçe Varlık İsmi Tanıma Sistemi'nin Kişi, Yer ve Kurum İsmi Tanımadaki Başarım Değerlendirme Sonuçları (Performance Evaluation Results of the Wikipedia Based Turkish Named Entity Recognition System on the Recognition of Person, Location and Organization Names)

Veri Kümesi Adı	*P*	*R*	*F*
Haber Veri Kümesi-1	93,04	74,11	82,5
Haber Veri Kümesi-2	78,62	44,76	57,05
Haber Veri Kümesi-3	89,19	56,29	69,02
Finans Haberleri Kümesi	84,35	23,14	36,31
Tarihi Metin Kümesi	74,07	43,6	54,89
Tweet Kümesi-1	70,52	24,9	36,8
Tweet Kümesi-2	58,63	19,29	29,03

Çizelge 5'teki sonuçlar incelendiğinde özellikle haber metinlerinde, onların içinde de özellikle Haber Veri Kümesi-1 üzerinde sistemin tüm ölçütlerde başarım değerlendirme sonuçlarının oldukça iyi olduğu gözlenmiştir. Sistemin herhangi bir kural listesi kullanmadan veya öğrenen sistemlerin gerektirdiği hatırı sayılır miktarda işaretli öğrenme kümesine ihtiyaç duymadan sadece çoğunluğu Wikipedia'dan otomatik yollarla elde edilmiş kaynaklarla bu değerlere ulaşmış olması oldukça önemlidir.

Ayrıca bu sistem sadece bu şekilde Wikipedia kaynaklarından beslenen Türkçe varlık ismi tanıma için geliştirilmiş olan ilk çalışmalardan biridir. Daha önceki ilgili sadece bir çalışma mevcuttur ve o da bizim de kaynaklarını kullandığımız [18] çalışmasında sunulmuştur. O çalışmada bizim çalışmamızda da kullandığımız Türkçe Wikipedia'daki başlıklardan kNN algoritması kullanılarak elde edilen kişi, yer ve kurum ismi listeleri kural-tabanlı varlık ismi tanıma sistemine ek kaynak olarak eklenmiş ve sistemin başarım değerlendirme sonuçlarını iyileştirmiştir. Bizim çalışmamızda ise çoğunluğu Wikipedia'dan elde edilen tüm kişi, yer ve kurum isimleri birleştirilip fazlalıklar çıkarıldıktan sonra tam bir Türkçe varlık ismi tanıma sisteminin gerçekleştiriminde kullanılmışlardır. Yani, bizim sistemimiz sadece çoğunluğu Wikipedia'dan çıkarılmış sözlüksel kaynaklara dayanmaktadır.

Çalışmamızın üzerine yapılabilecek ileri çalışmalarda;

1. Türkçe yer ve kurum isimleri de bu çalışmamızda kişi isimleri için yaptıklarımıza benzer şekilde otomatik olarak elde edilebilir. Böylelikle, elde edilecek yer ve kurum isimlerini de içerecek yeni bir Türkçe varlık ismi tanıma sistemi geliştirilebilir ve bu geliştirilecek sistemin varlık ismi tanımadaki başarımı daha yüksek olabilir.
2. Recall değerlerini arttırmaya yönelik çeşitli iyileştirmeler yapılabilir. İlk olarak tek kelimeden oluşan ve Wikipedia'dan otomatik çıkarılan kişi isimlerinin de sunulan sistemlere dâhil edilmesi üzerine çalışılabilir. Wikipedia gibi başka Web kaynaklarından da varlık ismi listeleri otomatik olarak çıkarılabilir ve kullanılabilir. Bunun dışında farklı çalışmalarda özellikle recall değerlerini arttıran metotlar araştırılarak bunlar mevcut sistemlerimize entegre edilebilir.
3. Türkçe üzerine yapılmış çalışmalar dışında, değişik özelliklere sahip Arapça [25] ve Hint dilleri [26] üzerinde yapılmış çalışmalar daha detaylı olarak incelenebilir ve bu çalışmalarda kullanılan ve Türkçe'ye de uygulanabilecek başarım arttırıcı metotlar mevcut sistemlerimize entegre edilebilir.
4. Son olarak, mevcut sistemlerimiz; tıp, sağlık, tarih ve sosyal bilimler gibi farklı alanlarda karşılaşılan varlık ismi türlerini de kapsayacak şekilde genişletilebilir, böylece sistemlerimiz daha geniş bir uygulama alanı kazanmış olur.

5. SONUÇLAR (CONCLUSIONS)

Varlık ismi tanıma, doğal dil işleme alanının konularından olan bilgi çıkarımının önemli araştırma alanlarından birisidir. Bu çalışmada, Türkçe için önce Wikipedia tabanlı bir kişi ismi tanıma sistemi, sonrasında da yine Türkçe için Wikipedia tabanlı tam bir varlık ismi tanıma sistemi sunulmuştur. Her iki sistem de Türkçe için mevcut çok farklı özelliklere sahip veri kümeleri üzerinde değerlendirilmiş ve oldukça umut verici başarım değerlendirme sonuçları elde edilmiştir. Değerlendirmelerimizde kullandığımız bu veri kümelerinin büyüklükleri ve çeşitliliği dikkate alınırsa, bu çalışmamız Türkçe'de varlık ismi tanıma konusundaki en kapsamlı değerlendirmeyi sunmaktadır.

Bu çalışmamız üzerinde yapılabilecek ileri çalışmalar şunlardır: Sistemlerin recall değerlerinin de arttırılabilmesi için Web üzerinden otomatik olarak elde

edilecek başka kaynakların da sistemlere entegre edilmesi sağlanabilir. Ayrıca, sistemlerin kaynakları; tıp, sağlık, tarih, sosyal bilimler gibi diğer önemli alanlardaki varlık ismi türlerini de kapsayacak şekilde genişletilebilir ve sistemler bu alanlar için de uygulanabilir hale getirilebilir.

KAYNAKLAR (REFERENCES)

1. Nadeau, D. ve Sekine, S. "A survey of named entity recognition and classification". ***Lingvistica Investigationes***, 30(1): 3-26, (2007)
2. Cucerzan S, Yarowsky D. "Language independent named entity recognition combining morphological and contextual evidence". ***Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora***, 90-99, (1999)
3. Tür, G., Hakkani-Tür, D. ve Oflazer, K. "A statistical information extraction system for Turkish". ***Natural Language Engineering***, 9(2): 181-210, (2003)
4. Küçük, D. ve Yazıcı, A. "Named entity recognition experiments on Turkish texts". ***International Conference on Flexible Query Answering Systems,*** LNCI 5822: 524-535, (2009)
5. Özger, Z. B. ve Diri, B. "Türkçe dokümanlar için kural tabanlı varlık ismi tanıma". ***Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi***, 6(6): 91-101, (2012)
6. Tatar, S. ve Çiçekli, İ. "Automatic rule learning exploiting morphological features for named entity recognition in Turkish". ***Journal of Information Science*** 37(2): 137-151, (2011)
7. McCallum, A. ve Li, W. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". ***Seventh Conference on Natural Language Learning at HLT-NAACL***, 188-191, (2003)
8. Yeniterzi, R. "Exploiting morphology in Turkish named entity recognition system". ***ACL Student Session***. 105-110, (2011)
9. Özkaya, S. ve Diri, B. "Türkçe metinlerde şartlı rasgele alanlarla varlık ismi tanıma". ***IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı***, 662-665, (2011)
10. Şeker, G.A. ve Eryiğit, G. "Initial explorations on using CRFs for Turkish named entity recognition". ***International Conference on Computational Linguistics***. 2459-2474, (2012)
11. Küçük, D. ve Yazıcı, A. "A hybrid named entity recognizer for Turkish". ***Expert Systems with Applications*** 39(3): 2733-2742, (2012)
12. Demir H, Özgür A. "Improving named entity recognition for morphologically rich languages using word embeddings". ***International Conference on Machine Learning and Applications***, 117-122, (2014)
13. Ritter, A., Clark, S., ve Etzioni, O. "Named entity recognition in tweets: an experimental study". ***Conference on Empirical Methods in Natural Language Processing***, 1524-1534, (2011)
14. Çelikkaya, G., Torunoğlu, D., Eryiğit, G. "Named entity recognition on real data: A preliminary investigation for Turkish". ***International Conference on Application of Information and Communication Technologies***, (2013)
15. Küçük, D., Jacquet, G., Steinberger, R "Named entity recognition on Turkish tweets". ***Language Resources and Evaluation Conference***. 450-454, (2014)
16. Küçük, D., ve Steinberger, R. "Experiments to improve named entity recognition on Turkish tweets". ***Workshop on Language Analysis for Social Media (LASM) of EACL***, 71-78, (2014)
17. Vikipedi: Özgür Ansiklopedi, https://tr.wikipedia.org/wiki/Ana_Sayfa
18. Küçük, D. "Automatic compilation of language resources for named entity recognition in Turkish by utilizing Wikipedia article titles". ***Computer Standards & Interfaces***, 1-9, (2015)
19. Zesch, T., Müller, C., ve Gurevych, I. "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary". ***Language Resources and Evaluation Conference***, 1646-1652, (2008)
20. Grishman, R. ve Sundheim, B. "Message understanding conference-6: A brief history". ***16th International Conference on Computational Linguistics***, 466-471, (1996)
21. Say, B., Zeyrek, D., Oflazer, K. ve Özge, U. "Development of a corpus and a treebank for present-day written Turkish". ***11th International Conference of Turkish Linguistics***, (2002)
22. Ratinov, L. ve Roth, D. "Design challenges and misconceptions in named entity recognition". ***Thirteenth Conference on Computational Natural Language Learning***, 147-155, (2009)
23. Cohen, W., Ravikumar, P. ve Fienberg, S. "A comparison of string metrics for matching names and records". ***KDD Workshop on Data Cleaning and Object Consolidation***, 73-78, (2003)
24. Tjong Kim Sang EF, De Meulder F. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". ***Seventh Conference on Natural Language Learning at HLT-NAACL***, 142-147, (2003)
25. Shaalan, K. "A survey of Arabic named entity recognition and classification". ***Computational Linguistics***, 40(2): 469-510, (2014)
26. Sasidhar, B., Yohan, P. M., Babu, A. V., & Govardhan, A. "A survey on named entity recognition in Indian languages with particular reference to Telugu". ***International Journal of Computer Science Issues***, 8: 438-443, (2011)