Türkçe Dokümanlar İçin Kural Tabanlı Varlık İsmi Tanıma (Named Entity Recognition for Turkish Text)

Zeynep Banu ÖZGER
Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü
zeynep.banu@hotmail.com

Banu DİRİ
Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü
banu@ce.yildiz.edu.tr

Özetçe

Varlık İsmi Tanıma, Doğal Dil İşleme biliminin çalışma alanlarından biri olup, dokümanlarda geçen varlık isimlerini kişi, yer ve organizasyon olarak ayırmanın yanı sıra formül, tarih ve parasal ifadeleri de bulabilmeyi hedefleyen, son yıllarda farklı dillerde çalışmaların devam ettiği bir alandır. Kural Tabanlı Varlık İsmi Tanıma ise, birtakım sözlüksel kaynaklar ile kurallar oluşturup, yüksek doğrulukla Varlık İsmi Tanıma işleminin gerçeklenmesidir.

Bu makalede farklı doküman türleri için tasarlanmış, Türkçe Kural Tabanlı bir Varlık İsmi Tanıma çalışmasından bahsedilmektedir. Varlıkları sınıflama ve etiketleme işlemi kişi, kurum ve yer isimleri ile tarih, para ve saat varlıkları olmak üzere toplam 6 farklı tür için gerçekleştirilmiştir. Varlık isimlerinin bulunup etiketlenebilmesi amacıyla her bir varlık türü için küçük boyutlu sözlükler kullanılarak kurallar oluşturulmuştur. Yapılan çalışmanın sonucunda kurum isimlerinden %86, yer isimlerinden %83, kişi isimlerinden ise %84 başarı elde edilmiştir. Sayısal varlık türlerinden ise tarih varlıklarından %92, saat varlıklarından %94 ve para varlıklarından %96 başarı elde edilerek tatmin edici sonuçlar alınmıştır.

Anahtar Kelimeler: Varlık İsmi Tanıma, Doğal Dil İşleme, Kural Tabanlı, Türkçe

Abstract

Named Entity Recognition, which is a Natural Language Processing workspace, aims to recognize the names and numeric expressions such as person, organization, location, date, money and time. The Rule Based Named Entity Recognition, that aims to recognize some rules with some lexical resources, performs the Named Entity Recognition process with high accuracy.

In this article, we mentioned about a Rule Based Named Entity Recognition for Turkish system. This system is designed for different types of documents. The system's classification process includes person, location, organization names and time, date, money entities. We have defined some rules with using small- sized lexical resources to perform the classification task. As a result of the study, the system's f-measure values are; 86% for organization names, 83% for location names, 84% for person names, 92% for date entities, 94% for time entities and 96% for money entities.

Keywords: Named Entity Recognition, Natural Language Processing, Rule Based, Turkish.

1. Giriş

Doğal Dil İşleme alanında yaygın olarak ihtiyaç duyulan Varlık İsmi Tanıma ilk defa 1995 yılında 6. Mesaj Anlama Konferansı'nda (Sixth Message Understanding Conference-MUC 6) tanıtılmıştır. Bilgi Çıkarımının bir alt dalı olan Varlık İsmi Tanıma (VİT) işleminin amacı dokümanlardaki isimleri (kurum, yer, kişi, ...) ve sayısal varlıkları (saat, para, tarih, yüzdelik ifadeler, ...) tanımaktır. Varlık İsmi Tanıma, kullanıcılara büyük doküman koleksiyonlarını çok daha çabuk ve verimli bir şekilde taramak için yardımcı olmaktadır.

Varlık İsmi Tanıma alanında geliştirilen sistemler, elle çıkarılan kurala dayalı algoritmaların yanı sıra denetimli (supervised), denetimsiz (unsupervised) ve yarı denetimli (semi supervised) makine öğrenme teknikleri kullanılarak da geliştirilmiştir. 1997 yılında Bikel [1] Gizli Markov Model'i (Hidden Markov Models-HMM), 1998 yılında Sekine [2] Karar Ağaçları'nı, Borthwick [3] Maksimum Entropi Modeli (Maximum Entropy Models-ME), 2003 yılında Asahara [4] Destek Vektör Makineleri'ni (Support Vector Machines-SVM) ve McCallum [5] Koşullu Rastgele Alanlar (Conditional Random Fields-CRF) gibi denetimli makine öğrenmesi teknikleri ile farklı sistemler tasarlamışlardır.

Yarı Denetimli Öğrenme (Semi Supervised Learning) tekniğini temel alan Brin [6] 1998 yılında kitap yazarları ile eşleştirilmiş, kitap başlıkları listelerini oluşturmak için düzenli ifadeler tarafından yürütülen sözcük özelliklerini kullanan bir sistem tasarlamıştır. 2001 yılında Cucchiarelli [7] sözdizimsel ilişkilere (nesne, özne gibi) bağlı bir sistemi yarı denetimli öğrenme tekniğiyle gerçekleştirmiştir.

Denetimsiz öğrenme teknikleri, kümeleme yöntemine ve kelime ağı (word net) gibi çeşitli sözlüksel kaynaklara dayanmaktadır. Bu tekniğe dayalı sistemlerde sadece giriş örnekleri verilerek kelimelerin aralarındaki ilişkiler bulunmaya çalışılmaktadır. Herhangi bir dış kaynaktan doğru veya yanlış bilgisi verilmemekte ve geri bildirimde bulunulmamaktadır. Etiketlenmemiş dokümanlardan hiçbir ek bilgi kullanılmadan doküman kümelenmeye çalışılmıştır. Etzioni [8] 2005 yılındaki çalışmasında Karşılıklı Bilgi ve Bilgi Erişimini (Pointwise Mutual Information and Information Retrieval-PMI & R) varlık isimlerini belirlemek için bir özellik olarak kullanmıştır. Bu yöntemle web sorguları kullanılarak iki ifade arasındaki bağımlılık ölçülmekte ve bağımlılık ölçümünün yüksek çıkması bu ifadelerin birlikte meydana gelme eğiliminde olduğunu göstermektedir.

Varlık İsmi ile ilgili ilk çalışma Rau [9] tarafından kurum isimlerini bulma ve sınıflandırmaya yönelik olup, elle çıkarılan kurallara dayalı algoritmalarla gerçeklenmiştir. Varlık İsmi Tanıma uygulamaları bulunacak varlığın türüne, dokümanın içeriğine (domain) ve dokümanın diline bağlı olmak üzere üç farklı başlık altında gerçekleştirilmiştir. Gerçeklenen ilk VİT çalışmalarında amaç kurum, yer ve kişi ismi olmak üzere üç temel varlık türünü tanımlamaktır. Son yıllarda, bilinen bu varlık türleri alt kategorilere ayrılmıştır. Lee'nin [10] çalışmasında yer isimleri şehir, devlet, ülke gibi alt kategorilere ayrılmaktadır. Fleischman'nın [11] gerçeklediği sistemde ise, kişi isimleri politikacı, şovmen gibi alt kategorilere sahiptir. Bazı çalışmalar ise sadece belirli varlık türlerine yönelik gerçeklenmiştir. Narayanaswamy [12] biyolojik terim isimlerini, Rindfleisch [13] ilaç isimlerini bulmaya yönelik çalışmıştır. Sekine [14] ise çalışmasında 4 tane ayrıntılı alt kategori içeren varlık ismi hiyerarşisi tanımlanmıştır. Müze, nehir, havalimanı isimleri bir kategoriyi oluştururken, ürün ve etkinlik isimleri ayrı birer kategoriyi, madde, hayvan, din ve renk isimleri de diğer bir alt kategoriyi oluşturmaktadır. Sekine'nin bu çalışmasında kategori sayısı yaklaşık 200 civarındadır ve bu kategorilerin her birinden bir varlık ismi oluşturmak için özellikler tanımlanmaktadır.

VİT uygulamalarında önemli olan bir diğer faktör ise dokümanın türü olup, çalışmaların bir kısmı doküman türünden bağımsız, bir kısmı da bağımlı olacak şekilde gerçeklenmiştir. 2001 yılında Maynard [15] bilimsel ve dini dokümanlar, 2005 yılında ise Minkov [16] e-mail dokümanlarından varlık isimlerinin çıkarılmasında başarılı sonuçlar almıştır.

Dil faktörü de VİT uygulamalarında önemli bir etken olup, bu alandaki çalışmaların çoğunluğu İngilizce için gerçeklenmiş olsa da son yıllarda diğer diller için geliştirilen çalışmalar dışında dilden bağımsız çalışmalar da bulunmaktadır. Çince için Yu [17], Fransızca için Poibeau [18], İtalyanca için Cucchiarelli [19], Bulgarca için Silva [20], Kore dili için Whitelaw [21], Lehçe için Piskorski [22] ve Rusça için Popov [23] tarafından yayınlanmış çalışmalar mevcuttur. Ayrıca, Almanca, İspanyolca, Hollandaca, Japonca, Yunanca, Romanca, İsveççe, Türkçe, Portekizce, Arapça ile Katalan, Danimarka, Hindi dilleri için yapılan çalışmalar da mevcuttur.

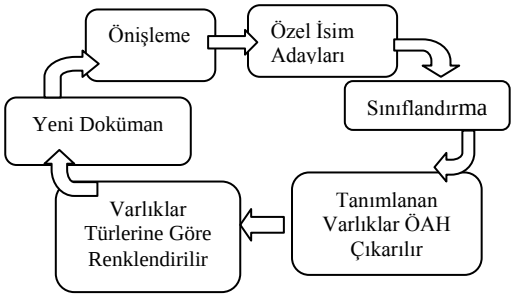
Türkçe dokümanlar için geliştirilen VİT çalışmalarının ilki Cucerzan [24] tarafından dilden bağımsız olarak tasarlanan bir sistem olup, Türkçe dokümanların yanı sıra İngilizce, Yunanca, Romanca ve Hintçe dillerinde yazılmış dokümanlar

için de değerlendirme sonuçlarını içermektedir. Bunun dışında, Türkçe için gerçeklenen diğer VİT çalışması Tür [25] tarafından gerçeklenmiş olup, İngilizce için tasarlanmış benzer bir sistemle karşılaştırmalı olarak verilmiştir. Bayraktar'ın [26] 2008 yılında tasarladığı sistem Yerel Dilbilgisi (Local-Grammar) tabanlı bir yaklaşımla gerçeklenmiş ve Türkçe finansal metinlerde ki kişi isimlerini bulmaya yöneliktir. Yine 2008 yılında, Küçük [27] tarafından yayınlanan çalışmada gazete makaleleri için tasarlanmış kural tabanlı bir kişi ismi tanıma sisteminden bahsedilmektedir. Küçük [28] tarafından gerçeklenmiş bir başka çalışmada yazarların bir önceki çalışmalarından da yararlanarak Türkçe için açıklamalı bir video arşiv aracı kullanarak daha genel kural tabanlı bir VİT sistemi tasarlanmıştır. Küçük'ün, [29] 2009 yılındaki çalışmasında ise, farklı türlerdeki dokümanlarda sistemin başarısı ölçümlenmektedir.

Bu makalede bahsi geçen sistem, bilgi çıkarımında yüksek doğruluk gösterdiği için kural tabanlı, Türkçe'nin geçerli dilbilgisi kurallarına ve karakteristik özelliklerine göre geliştirilmiş, genişletmeye açık esnek bir şekilde tasarlanmıştır. Çalışmanın ikinci bölümde tasarlanan sistemin genel yapısından, üçüncü bölümde varlık isimlerinin etiketlenebilmesi için tanımlanan kurallardan, dördüncü bölümde kullanılan veri setinden ve değerlendirme kriterlerinden bahsedilmektedir. Beşinci bölümde, sistemin başarımı ve altıncı bölümde de sonuç ve gelecek çalışmalardan bahsedilmektedir.

2. Geliştirilen Sistemin Yapısı

Kural tabanlı tanıma için oluşturulan sistemin amacı, Türkçe dokümanlarda geçen kişi, yer ve kurum isimleri ile para, tarih ve saat varlıklarının bulunarak etiketlenmesidir. Sistemin gerçeklenmesi sırasında varlık isimlerini içerisinde tutan herhangi bir sözlüksel kaynak kullanılmamıştır. Şekil 1'de gösterildiği gibi oluşturulan sistem, içerikten bağımsız olacak şekilde tasarlanmıştır. Sistemin kelimeleri birbirinden ayırt edebilmesi için kelimelerden ve noktalama işaretlerinden sonra gelen boşluklardan yararlanılmıştır. Bu nedenle dokümanı işlemeye başlamadan önce, sistem dokümandaki noktalama işaretlerinden sonra boşluk olup olmadığını kontrol eder, eğer bir sonraki karakteri boşluk olmayan bir noktalama işareti ile karşılaşırsa boşluğu kendisi ekler. Bu işlem *önişleme* olarak adlandırılmıştır. Önişleme adımı tamamlandıktan sonra, sistem büyük harf ile başlayan tüm kelimeleri "*Özel İsim Adayı Havuzunda*-ÖAH" saklar. Ardından tüm varlık isimleri için tanımlanmış olan kurallar tek tek çalıştırılarak etiketleme işlemi gerçekleştirilir. Her bir varlık türü için tanımlanan kuralların uygulanması tamamlandığında, sistem sınıflandırdığı varlık isimlerini ÖAH'dan silmektedir. Böylece tüm varlık türlerinin sınıflandırılması işlemi bittiğinde, özel isim adayı havuzunda özel isim olmayan veya sistem tarafından etiketlenemeyen varlık isimleri kalmaktadır. Kullanıcı isterse tanımlanamayan varlık isimlerini el ile etiketleyebilir.



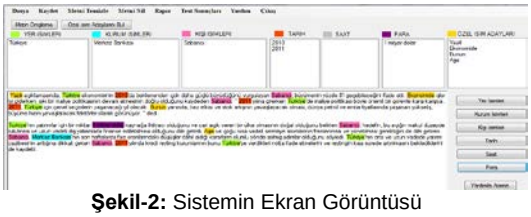
Şekil-1: Sistemin Yapısı

Etiketleme işlemi bittikten sonra bulunan varlıklar eğer ek aldıysa ekler silinmekte ve her bir varlık türü için yapılan farklı renklendirme ile (yer isimleri *yeşil,* kurum isimleri *mavi*, kişi isimleri *pembe,* tarih *kırmızı*, saat *gri*, para *mor*, özel isim adayı *sarı*) kullanıcının kolay ayırt etmesi sağlandıktan sonra geçici isim havuzları boşaltılmaktadır. Şekil 2'de C# ile geliştirilen uygulamanın bir ekonomi dokümanı üzerindeki ekran çıktısı görülmektedir.

3. Varlık İsimleri İçin Tanımlanan Kurallar

Gerçekleştirilen sistem kural tabanlı bir yapı üzerine kurulmuş olup, her bir varlık için o türe ait belirleyici durumlar araştırılıp, kural haline getirilmiştir. Her bir varlık türü için belirleyici olan anahtar kelimeler tanımlanmıştır. Anahtar kelimelerin tanımlanmasında [29]'daki çalışmadan yararlanılmış olup sistem eğitim verisi oluşturulurken gözlemlenen durumlara göre yeni kurallar ve yeni anahtar kelimeler eklenmiştir. Son durumda sistem üzerinde tanımlı toplam 33 adet

kural 335 adet anahtar kelime kullanılmıştır. Varlık isimlerinin çıkarılmasında her bir varlık ismi bağımsız olarak değerlendirilmiştir.



Şekil-2: Sistemin Ekran Görüntüsü

3.1 Kişi İsimleri

Kişi isimlerinin tespiti için tanımlanan kurallar aşağıda sıralanmaktadır:

1- Kişi isimlerinden önce gelen kelimeler araştırılarak (bay, başbakan, ağabey, tuğgeneral, diyen, eşi, annesi, diyetisyen, veren, üyesi, inanan, doğan, yaşındaki, babası, sayın, muhtar, temsilcisi, konuşan, vurgulayan, vd.) gibi 116 adet kişiyi tanımlayan kelime bulunmuştur. Her bir niteleyici kelime doküman içerisinde sistem tarafından aranır. Sistem bulduğu anahtar kelimeden sonra gelen ilk dört kelimeyi inceler ve ileriye doğru ardışık olarak büyük harfle başlayan kelimeler birleştirilerek kişi ismi olarak etiketlenir. Dört kelime sınırının kullanılmasının nedeni ilgili bölümün kısıtlarında anlatılmaktadır. Ör: "... ağabeyi Ali Murat ....", "... gazeteye konuşan Ali Murat ...", "... Vali Ali Murat Ege...".

2- Kişi isimlerinden sonra gelen (Hanım, Ağa, Teyze, Sultan, Paşa, Hoca, Efendi, vd.) anahtar kelimelerinden önce gelen ardışık büyük harfle başlayan dört kelimeyi birleştirerek kişi ismi olarak atar. Ör: "... Ayşe Hanım ...", "... Mehmet Ali Ağa ...".

3- Sistem, unvan kelimelerinden önce ve sonra gelen kelimeleri incelemektedir. "Prof. Dr." gibi bir anahtar kelimeden önce veya sonra gelen isim birden fazla kelimeden oluşuyorsa önce veya sonrasında gelen dört kelime büyük harfle başlamak koşulu ile kişi ismi olarak etiketlenmektedir.

4- Anahtar kelimelerin sonrası veya öncesi incelenen dört kelimenin her birinin büyük harfle başlaması şarttır. Döngü sırasında sıradaki kelime büyük harfle başlamıyorsa döngü sonlandırılmaktadır. "Prof. Dr. Ali Murat yarın Amerika'ya..." cümlesinde anahtar kelimeden sonra gelen dört kelime incelemekte ve "Ali Murat" kelimeleri kişi ismi olarak etiketledikten sonra "yarın" kelimesine geçmektedir. İlgili kelime büyük harfle başlamadığından inceleme sonlanmaktadır, böylece unvandan sonra gelen "Amerika" kelimesi büyük harfle başladığı ve dördüncü kelime olduğu halde kişi ismi olarak etiketlenmesi önlenmektedir.

5- Bazı durumlarda bir anahtar kelimeden sonra gelen kelimeler büyük harfle başladığı halde hepsi kişi ismi olmamaktadır. Eğer bu kelimeler arasında bir noktalama işareti varsa noktalama işaretine kadar olan kısım kişi ismi olarak etiketlenmektedir. "Prof. Dr. Ali Murat, Amerika gezisinde..." cümlesinde sadece Ali Murat kişi ismi olarak etiketlenmektedir.

6- Bulunan bir kişi ismi birden fazla kelimeden oluşuyorsa bu kelimeler geçici isim havuzuna tek tek eklenirken ekrandaki listeye bütün halinde eklenmektedir. Sistem "Ali Murat" ismini bulduysa geçici isim havuzunda "Ali" ve "Murat" ayrı ayrı tutulurken, ekrandaki listede "Ali Murat" tek bir kayıt olarak yer almaktadır. Böylece doküman içerisinde ad veya soyad tek başına geçerse her birinin bulunması sağlanmaktadır.

7- Bazı durumlarda anahtar kelimelerinde birer kişi ismi olarak etiketlendiği tespit edilmiştir "...Genel Başkan Yardımcısı Ali Murat ..." gibi cümlelerde "başkan" anahtar kelimesinden sonra gelen 3 kelime "Yardımcısı Ali Murat" kişi ismi olarak etiketlenmektedir. Etiketleme işlemi bittikten sonra çalıştırılan fonksiyon ile "Yardımcısı" kelimesinin anahtar kelime olduğu tespit edilerek silinir. Böylece kişi ismi olarak etiketlenen kelime sadece "Ali Murat" olmaktadır.

8- Etiketlenmiş bir kişi isminden önce veya sonra "ve" bağlacı geldiyse, bağlaçtan önceki veya sonraki kelimede kişi ismi olarak etiketlenmektedir. "Sayın Ali....., Ali ve Ahmet,...." cümlesinde "Ali" kelimesi, "Sayın" anahtar kelimesi ile tanımlanmış olduğundan kişi ismi olarak etiketlenmekte, "Ahmet" herhangi bir anahtar kelime almamasına rağmen daha önce etiketlenen bir kişi ismine "ve" bağlacıyla bağlandığından kişi ismi olarak etiketlenebilmektedir.

9- Etiketlenmiş olan bir kişi isminden önce veya sonra virgül (,) gelmişse ve bu sembolden önceki veya sonraki kelimeler büyük harfle başlıyorsa kişi ismi olarak etiketlenmektedir. "...Edip Gümüş, Cemal Bey...".cümlesinde, "Cemal" kelimesi, "Bey" unvanı ile nitelenmiş olduğundan sistem tarafından tanınmaktadır. "Edip Gümüş" ise etiketlenmiş bir kişi ismine virgül ile bağlandığı için kişi ismi olarak etiketlenmektedir.

3.1.1 Kişi İsimlerindeki Kısıtlar

1- Sistem, anahtar kelimeden önce veya sonra gelen dört kelimeyi dikkate aldığından bir kişi ismi dörtten fazla kelimeden oluşuyorsa, sadece ilk veya son dördü bulunabilmektedir (4'ün üzeri kelime seçildiğinde hata oranının arttığı tespit edilmiştir).

2- Bölüm 3.1'deki 7 ve 8 numaralı kurallar zaman zaman kişi ismi olmayan kelimelerinde kişi ismi gibi etiketlenmesine yol açmaktadır. **"**Prof. Dr. Ali Murat ve Amerika'daki arkadaşı...." cümlesinde "Ali Murat", "Prof. Dr." unvanına sahip olduğu için kişi ismi olarak etiketlenmektedir. "Amerika" ise kişi ismi olmamasına karşın tanımlanmış olan bir kişi ismine "ve" bağlacı ile bağlanmış olması ve büyük harf ile başlaması nedeniyle kişi ismi olarak etiketlenmektedir.

3- Bazı kurum isimleri, kurum sahibinin adı veya soyadı bilgisini de içerisinde bulundurur. Böyle durumlarda aynı kelime hem kurum hem de kişi ismi olarak etiketlenmekte fakat, renklendirirken en son bulunan varlık ismi olarak renk verilmektedir. Örneğin, "Sayın Ali Kara", "Kara Şirketler Grubu" cümlelerinde "Kara" hem kurum hem de kişi ismi olarak etiketlenmektedir.

3.2 Kurum İsimleri

Doküman içerisinde kurum isimlerini etiketleyebilmek için aşağıdaki kurallar tanımlanmıştır:

1-Kurum isimlerini niteleyen veya kurum isimlerinden sonra gelen kelimeler araştırılmış ve (gazetesi, kulübü, fabrikası, enstitüsü, bakanlığı, vakfı, başkanlığı, üniversitesi, lisesi, kurumu, müdürlüğü, komutanlığı, vd.) 79 tane anahtar kelime tespit edilmiştir. Her bir kelime doküman içerisinde aranır ve bulunan anahtar kelime kurum isimlerini niteleyen kelimelerden biriyse bu kelimeden önce gelen ve büyük harfle başlayan ilk dört kelime aranan anahtar kelime ile birleştirilerek kurum ismi olarak etiketlenmektedir. "...İstanbul Lisesi..." cümlesinde, "İstanbul Lisesi", "lise" anahtar kelimesini içerdiğinden sistem tarafından tanınabilmektedir.

Eğer aranan anahtar kelime kurum isimlerinden sonra gelen anahtar kelimelerdense, bu kelimeden önce gelen ve büyük harfle başlayan ilk dört kelime anahtar kelime eklenmeden birleştirilerek kurum ismi olarak etiketlenmektedir. "...Galatasaray Genel Başkanı..." cümlesinde, kurum ismi olan "Galatasaray Genel Başkanı" değil, "Galatasaray" dır ve aldığı anahtar kelimeden bağımsız olarak kurum isimleri listesine eklenmektedir. Sıklıkla kurum isimlerinden sonra geldiği tespit edilen anahtar kelimeler, "Yönetim Kurulu, Genel Başkanı, Muhabiri, Milletvekili, Üyesi, kulüp, takım, mağaza, gazete, marka, vd." olmuştur.

2- Kurum isimlerinin içerisinde Spor, TV, Bank gibi sıklıkla geçen kelimeler tespit edilmiş ve doküman içerisinde aranmıştır. Eğer kelimelerden biri, başka bir kelimenin içerisinde veya bir kısaltmada geçiyorsa ve büyük harfle başlıyorsa kurum ismi olarak etiketlenmektedir. Ör: Adabank, Kayserispor, NTV.

3- Kurum isimlerinin çoğunlukla kısaltma ile ifade edildiği görülmektedir. Bu nedenle sistem tamamı büyük harften oluşan kelimeleri kısaltma olarak kabul etmektedir. Kurum ve yer isimlerinin bulunması işlemi bittikten sonra kısaltmaların açılımları önce yer isimleri içerisinde aranmakta, bulunmadığı taktirde de kurum ismi olarak etiketlenmektedir."ABD" bir yer ismi olup, doküman içerisinde "Amerika Birleşik Devletleri" ifadesi varsa "Devlet" anahtar kelimesinden dolayı yer ismi, yoksa kurum ismi olarak etiketlenmektedir.

4- Bazı kurum ismi kısaltmalarından sonra parantez içerisinde kısaltmanın açılımının verildiği gözlemlenmiştir. Ör: KKDF (Kaynak Kullanımı Destekleme Fonu).

5- Bazı durumlarda bir kurum isminin kısaltmasının, ilgili kurumun açık isminin ardından parantez içerisinde verildiği görülmüştür. Ör: Kaynak Kullanımı Destekleme Fonu (KKDF).

6- Bir anahtar kelimeden önce gelen kelimeler içerisinde "ve" bağlacı varsa bu kelime anahtar kelimeden önce gelen dört kelimeye dahil edilmekte ve büyük harfle başlama koşulu aranmayarak etiketlenmektedir. Ör: Bilgi ve Teknolojileri Kurumu.

7- Bir anahtar kelimeden önce gelen dört kelimenin herhangi birinden sonra virgül (,) varsa dört kelime döngüsü sonlandırılmaktadır. Böylece kurum ismine dahil olmayan, ilgili kurum isminden virgül ile ayrılmış olan ve büyük harfle başlayan özel isimlerin yanlış etiketlenmesi önlenmektedir. "...söyleyen Ali, Bilgi ve Teknolojileri Kurumu'na..." cümlesinde, "Ali" isminin kurum ismine dahil edilmesi önlenmiştir.

3.2.1 Kurum İsimlerindeki Kısıtlar

1- Bazı durumlarda kısaltmalar kurum ismi olmamasına karşın kurum ismi olarak etiketlenmektedir. Ör: "TL".

2- Bir kurum ismi kendini tanımlayıcı bir anahtar kelimeye sahip değilse ve kısaltma değilse bulunamamaktadır.

3- Sistem bir anahtar kelimeden önce gelen dört kelimeyi dikkate aldığından bir kurum ismi dörtten fazla kelimeden oluşuyorsa bunlardan sadece son dördü bulunabilmektedir. Ör: Bedensel ve *Zihinsel Engelli Çocukları Koruma Derneği*.

4- "Galatasaray Genel Başkanı..." cümlesinde, "Başkan" anahtar kelimesinden önce geldiği için, "Galatasaray Genel" ifadesi kurum ismi olarak atanmaktadır.

3.3- Yer İsimleri (Location Names)

1- Yer isimlerini etiketleyebilmek için aşağıdaki kurallar tanımlanmıştır:

1- Yer isimlerinden önce gelen "kuzey, güney, doğu, batı, başkent, vd." gibi kelimeler anahtar kelime olarak tespit edilmiş, öncesinde gelen ve ardışık olarak büyük harfle başlayan en fazla dört kelime yer ismi olarak etiketlenmiştir. Ör: Kuzey Amerika, başkent Ankara.

2- Yer isimlerinden sonra geldiği tespit edilen "durak, cadde, mahalle, tiyatro, nehri, cezaevi, hastane, bulvar, doğusu, uyruklu, devleti, vd." gibi 84 anahtar kelime çıkarılmış ve öncesinde ardışık olarak büyük harfle başlayan kelimeler yer ismi olarak işaretlenmiştir.

3- Bazı yer isimleri bileşik isim olup, içerisinde köy, deniz, doğu gibi sisteme anahtar kelime olarak verilen kelimelerden birini içeren ve büyük harfle başlayan kelimenin tamamı yer ismi olarak etiketlenmektedir. Ayrıca, Türkçe' de sıklıkla ülke isimlerini belirtmekte kullanılan "istan" ve "ye" heceleri tek başına anlamı olan bir kelime olmasalarda bu gruba dahil edilmektedir. Ör: Kadıköy, Karadeniz, Ortadoğu, Kazakistan, Almanya, vd.

4- Yer isimlerine eklendiği tespit edilen -de, -da, -den, -dan, -te, -ta gibi hal eklerini alan kelime büyük harfle başlıyorsa yer ismi olarak etiketlemektedir. Ör: Almanya'da.

5- Bir yer isminin içerisinde "ve" bağlacı varsa, büyük harfle başlama koşulu aranmamakta ve bu kelime anahtar kelimelerden önce veya sonra gelen dört kelime içerisine dahil edilmektedir. Ör: "Moda ve Tasarım Müzesi".

6- Etiketlenmiş bir yer isminden önce veya sonra "-" işareti varsa ve bu işaretten önceki veya sonraki kelime büyük harfle başlıyorsa bu kelimede yer ismi olarak etiketlenir. Ör: "...İspanya'da..., İspanya-Fransa ilişkileri...".

7- Ardışık olarak aranan dört kelimelik döngüde kelimeler arasında bir noktalama işareti ("-" hariç) varsa döngü sonlandırılmaktadır. Ör: "**...** söyledi Ali. Amerika'da oluşan....".

3.3.1 Yer İsimlerindeki Kısıtlar

1- Bir yer ismi dört kelimeden uzunsa sistem sadece ilk veya son dördünü bulabilmekte ve yer ismi herhangi bir anahtar kelime içermiyorsa bulunamamaktadır.

2- Sistem, -de ve -den hal eki alan ve büyük harfle başlayan kelimeleri yer ismi saydığından bu eklerden birini alan ancak, yer ismi olmayan bazı özel isimlerde yer ismi olarak etiketlenmektedir. Ör: "Eşyalar Ali' de kaldı

3- -ye, -ya ekleri, yer ismi olmayan bazı özel isimleri hatalı etiketlemektedir. Ör: "Aliye".

3.4 Tarih Varlıkları

Türkçede tarihlerin gösteriminde gün ay yıl sırası izlenmekte olup, 29 Temmuz 2011, 29.07.2011, 29-07-2011, 29/07/2011, 29.07.11, 29-07-11, 29/07/11, 2011 yılında, Temmuz ayında..., 2011'de...,vd. şeklinde örnekler verebiliriz. Tarih varlık isimlerini bulmak için aşağıdaki kurallar çıkarılmıştır:

1- 29.07.2011, 29-07-2011, 29/07/2011, 29.07.11, 29-07-11 ve 29/07/11 kullanımları için sonlu durum algoritmaları oluşturulmuştur.

2- “29 Temmuz 2011” şeklinde ay ismini içeren gösterimleri tespit edebilmek için, ay isimlerinden biri dokümanda geçtiğinde önce ve/veya sonrasında rakam gelip gelmediği kontrol edilmektedir.

3- -de, -den eki almış bir kelime, rakamlardan oluşuyorsa tarih varlığı olarak etiketlenmektedir.

4- Tarihlerden sonra geldiği tespit edilen “yılı, senesinde, tarihinde, kuşağı, seçimleri, tarihli, 'li yıllarda, yılları, vd.” gibi anahtar kelimelerden önce gelen kelime, rakamlardan oluşuyorsa tarih olarak etiketlenmektedir. Ör: “1990'lı yıllarda”, “1990 tarihli”.

5- Etiketlenen tarih varlığından önce veya sonra "ve" bağlacı gelmiş ise, bağlaçtan önce veya sonraki kelime tarih varlığı kurallarından birine uyuyor ve bağlaçtan önce gelen kelimenin rakam olması durumunda tarih olarak etiketleme yapılmaktadır. Ör: “2 ve 13 Mayıs tarihlerinde...”.

3.4.1 Tarihler İçin Kısıtlar

1- -de ve -da ekleri hem yer hem de tarih varlıklarına eklenebilmektedir. Rakamlardan oluşan tarihlerin yer ismi gibi algılanmasını önlemek amacıyla hal ekini alan kelimenin rakam veya harflerden oluşmasına göre etiketleme gerçekleştirilir. Ör: "Temmuz' da”.

2- "Ali ayın 3'ünde gelecek." cümlesi içerisinde “ayın 3” ifadesini bulabilecek bir kural tanımlanmamış olduğundan ilgili varlık ismi bulunamamaktadır.

3.5 Saat Varlıkları

Türkçe'de saatlerin gösteriminde saat, dakika ve saniye sırası izlenmekte olup, “15:20, 15.20, 5:20, 5.20, saat 5, 15:20:30, 15.20.30, vd.” şeklinde örnekler verebiliriz. Saat varlıklarının bulunması için çıkarılan kurallar:

1- 15:20, 15.20; 5:20, 5.20 ve 15:20:10, 15.20.10 şeklindeki gösterimler için üç farklı sonlu durum algoritması tanımlanmış olup, bu formatlara sahip tüm saat varlıkları etiketlenmektedir. Ayrıca, saat 7 gibi gösterimler için de düzenleme yapılmıştır.

3.6.1 Saat Varlıkları İçin Kısıtlar

1- Yüzdelik veya reel sayı içeren dokümanlarda yanlış saat etiketlemeleri olabilmektedir. Ör: "...borsadaki 3.10'luk düşüş...".

3.6 Para Varlıkları

Türkçe dokümanlarda para ifade eden sayıların gösteriminde miktarın tam ve ondalıklı kısmı virgül (,) ile ayrılmakta ve miktar bir tamsayı ise virgülden sonra sıfır/sıfırlar eklenmekte veya virgülsüz olarak gösterilmekte ve genelde miktarı belirten sayının ardından para birimi yazılmaktadır. Para varlıkları için tanımlanan kurallar:

1- Bir para birimi dokümanda geçiyorsa, para biriminden önceki karakterler bir sonlu otomat algoritmasıyla incelenmektedir. Virgülden önce ve sonra gelen rakamların adedinde bir kısıtlama yapılmamaktadır.Ör**:** “2,300 TL”.

2- “4 milyon TL” örneğinde olduğu gibi, para varlığının tamamı dokümanda rakamla gösterilmez. “TL, Lira, Dolar, Euro, Bin TL, milyon TL, vd." gibi ifadeler anahtar kelime olarak sisteme eklenerek doğru etiketleme yapılmaktadır.

3- "300-400 dolar" şeklindeki bir gösterimde "dolar" anahtar kelimesi sayesinde "400 dolar" para varlığı olarak seçilmekte ve her para varlığından önce (-) işaretinin gelip gelmediği kontrol edilerek, (-) 'den önceki karakterlerin sayı olması durumunda bir önceki sayısal değer de para varlığı olarak etiketlenmektedir.

4- Bazen, para varlıkları “2 bin 500 lira” gibi iki ayrı ifade ile gösterilmektedir. Burada 500 lira "lira" anahtar kelimesinden dolayı bulunabilmektedir. Sistem "500 lira" dan önce gelen kelimenin anahtar kelime ve öncesindeki kelimenin de rakamlardan oluşup oluşmadığını kontrol eder. Şartlar sağlanırsa tüm kelimeler birleştirilerek etiketleme tamamlanır.

3.6.1 Para Varlıkları İçin Kısıtlar

1- Türkçe dokümanlar içerisinde para birimi içermeyen varlıklar yer alabilir. Ör: 60 milyar.

2- Türkçe'de para varlıklarının ondalıklı kısmının tam kısmından virgül ile ayrılması kuralı olmasına karşın, bazı dokümanlarda virgül yerine nokta kullanıldığı görülmüştür.

4- Veri Seti

Geliştirilen Varlık İsmi Tanıma sisteminin konudan bağımsız olarak tasarlandığını göstermek amacıyla, güncel gazetelerin farklı tarihlerde yayınlanmış sağlık, siyaset, güncel haberler, spor, ekonomi ve magazin gibi her türden 20 adet doküman olmak üzere 6 farklı türde toplam 120 doküman üzerinde sistem test edilmiştir. Çalışmada kullanılan veri setine [30]'dan erişilebilinir. Test dokümanlarının toplam boyutu 158 Kb olup, Çizelge-1'de veri setine dair bilgiler bulunmaktadır. Sistemin başarısını ölçmek için dokümanlardaki varlık isimlerinin tespiti için elle etiketleme yapılmıştır. Başarı oranının ölçülmesinde de sırasıyla eşitlik-1, 2 ve 3'teki Tutturma (Precision-P), Bulma (Recall-R) ve ikisinin harmonik ortalaması olan F-ölçüm (F-measure) kullanılmıştır.

Çizelge-1: Veri Seti (#toplam sayı)

Metin Türü	# Kelime	# Kişi İsmi	#Kurum İsmi	# Yer İsmi	#Tarih, Saat, Para
Ekonomi	3673	94	111	110	89
Spor	2778	99	132	41	25
Siyaset	4019	144	91	123	30
Sağlık	3779	75	60	58	23
Magazin	2449	149	25	49	19
Güncel	2814	103	64	99	18
Toplam	19512	664	483	480	204

P=Doğru tespit edilen isim-sayısı/Tespit edilen isim-sayısı (1)

R=Doğru tespit edilen isim-sayısı/Test kümesindeki toplam isim-sayısı (2)

F-Ölçüm= (2 * P * R) / (P + R) (3)

5- Deneysel Sonuçlar

Ekonomi, sağlık, güncel, magazin siyaset ve spor alanındaki dokümanlar içerisinden çıkarılan varlık isimlerinin başarıları kişi, kurum, yer, tarih, saat ve para için ayrı ayrı incelenmiştir.

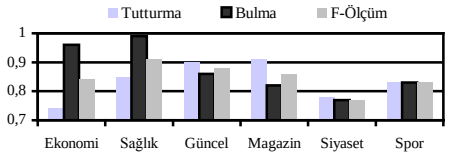
5.1 Kişi İsimleri

Sistemin kişi isimlerindeki başarısı incelendiğinde, F-ölçüm kriterine göre en yüksek ve düşük başarı sırasıyla %91 ile sağlık, %77 ile siyaset alanındaki dokümanlardan alınmıştır (Şekil-3). Siyaset alanındaki hataların nedenleri:

1- Bazı kurum isimlerinin, kişi isimlerini niteleyen anahtar kelimelerle ifade edilmiş olması. "...diyen X Parti..." örneğinde, "diyen" kelimesi kişi ismi bulmada kullanılan bir anahtar kelime olduğundan, "X Parti" aynı zamanda kişi ismi olarak da etiketlenmektedir.

2- Bazen bir kişi ismi birden fazla anahtar kelimeyle nitelendirilmektedir. "...konuşan Bakan Ali Murat..." örneğinde, "konuşan" ve "Bakan" kelimeleri kişi ismi bulmada kullanıldığından, "Bakan Ali Murat" ve "Ali Murat" ayrı ayrı kişi ismi olarak etiketlenmektedir.

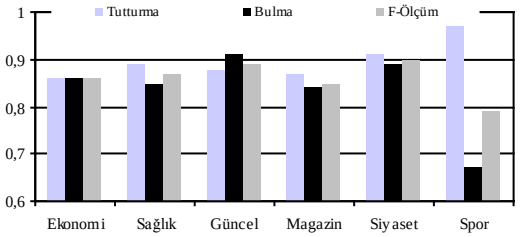
3- Bulunamayan bir kişi ismi dokümanda birden fazla yerde geçiyorsa başarım düşmektedir.



Şekil-3: Kişi İsimleri

5.2 Kurum İsimleri

Kurum isimlerinde en yüksek başarı F-ölçüm kriterine göre %90 ile siyaset, en düşük başarı %79 ile spor alanındaki dokümanlardan alınmıştır (Şekil-4). Spor alanındaki dokümanlarda takım isimleri herhangi bir anahtar kelime ile nitelenmeden Galatasaray, Fenerbahçe gibi yalın halde bulunduğundan ve fazla sayıda geçtiğinden başarım düşmektedir.

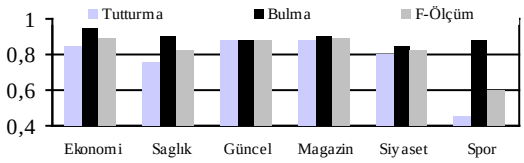


Şekil 4: Kurum İsimleri

5.3 Yer İsimleri

Yer isimleri, F-ölçüm kriterine göre değerlendirildiğinde Şekil-5'te görüldüğü gibi, en yüksek başarı %89 ile ekonomi ve magazin alanında elde edilirken, en düşük başarı %60 ile spor

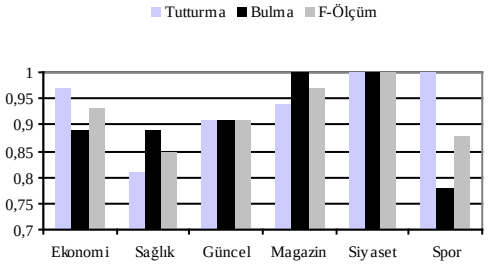
alanındaki dokümanlardan elde edilmiştir. Takım isimlerinin -de hal ekini almasından kaynaklanan hatalar spor alanındaki dokümanlarda başarının düşmesine neden olmaktadır.



Şekil-5: Yer İsimleri

5.4 Tarih Varlıkları

Tarih varlıkları için en yüksek başarı %100 ile siyaset alanında alınırken, en düşük başarı %85 ile sağlık dokümanlarından elde edilmiştir (Şekil-6). Tarihlerin, "2011" ifadesinde olduğu gibi belirlenen formatlarda yazılmaması, ek, anahtar kelime veya ay isimlerini içermemesi durumunda hatalı çalışmaktadır.



Şekil-6: Tarih Varlıkları

5.5 Saat Varlıkları

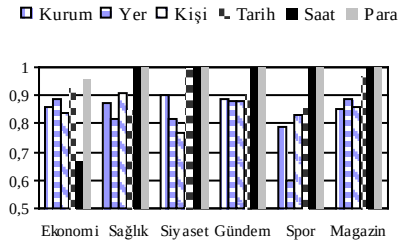
Saat varlıklarında, ekonomi türündeki dokümanlar hariç %100 başarı alınmış iken, Ekonomi dokümanlarında yüzdelik ifadelerin saat varlığı formatında verilmiş olması nedeniyle başarı %67 olmuştur."....borsadaki 3.57'lik düşüş..." cümlesinde, "3.57" ifadesi saat varlığı olarak etiketlenmektedir.

5.6 Para Varlıkları

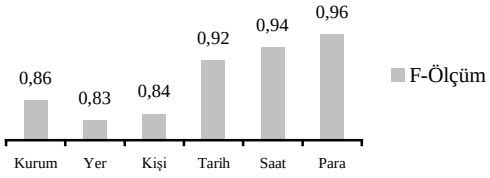
Para varlıklarında, ekonomi alanındaki dokümanlarda %96 ile en düşük başarı alınırken, diğer alanların hepsinde %100'lük başarı elde edilmiştir. Ekonomi alanındaki dokümanlarda, "60 milyon" gibi para varlıkları, para birimi içermediği için sistem tarafından bulunamamaktadır.

5.7 Varlık İsmi Türlerinin Karşılaştırma

Varlık türlerinin doküman tiplerine göre karşılaştırmalı değerlendirmesi Şekil-7'de verilmektedir. Kurum isimlerinde en iyi sonuç %90 ile siyaset, yer isimlerinde %89 ile ekonomi ve magazin, kişi isimlerinde %91 ile sağlık, tarih varlıklarında %100 ile siyaset, para ve saat varlıklarında ise ekonomi dışındaki tüm doküman türlerinden %100 başarı alınmıştır. Şekil-8'de sistemin genel başarı grafiği doküman türü ayrımı yapılmadan verilmiştir. Veri setimizde 664 adet kişi, 483 adet kurum, 480 adet yer ismi ve 204 adet tarih, saat ve para varlık ismi mevcut olup, geliştirdiğimiz sistem sırasıyla bu varlık isimlerinin 569, 395, 425 ve 190 tanesini doğru olarak tespit etmiştir. Sistem para ve saat varlıklarından en yüksek başarıyı sırasıyla %96 ve %94 olarak elde etmiştir. Sayısal varlıklar için en düşük başarı ise %92 ile tarih varlıklarından elde edilmiştir. Özel isimlerde, yer isimlerinden %83, kişi isimlerinden %84 ve kurum isimlerinden %86 başarı elde edilmiştir.



Şekil-7: Varlık Türlerine Göre Karşılaştırma



Şekil 8: Genel Değerlendirme

6- Sonuç

Varlık İsmi Tanıma son yıllarda gelişmekte olan önemli bir bilgi çıkarımı alanıdır. Dokümanlardan varlık isimlerinin çıkarılması için 20'den fazla dilde çalışma yapılmıştır. Bu makalede Türkçe dokümanlar için konudan bağımsız, kural tabanlı, kurum, yer, kişi isimleri ile tarih, para, saat varlık isimlerinin bulunması ve etiketlenmesi için

gerçekleştirilmiş bir uygulamadan ve çıkarılan kurallardan bahsedilmektedir. Sistem, 6 farklı türde, 120 doküman üzerinde test edilmiştir. Tasarlanan sistem üzerinde yapılan testler sonucunda sırasıyla kurum, yer ve kişi isimlerinden %86, %83 ve %84 ortalama başarı elde edilmiştir. Sayısal varlık türlerinden de tarih, saat ve para varlıkları için sırasıyla %92, %94 ve %96 başarı alınmıştır. Türkçe için tasarlanmış Varlık İsmi Tanıma çalışmalarında farklı test dokümanları kullanıldığı için, bu alandaki diğer çalışmalarla kıyaslama yapılamamıştır. Bu çalışmanın devamında sistemi farklı alanlardaki dokümanlarla test etmek ve makine öğrenmesi yöntemlerini kullanarak başarımın artırılması ve farklı varlık tiplerinin de bulunup, etiketlenmesini sağlayacak kuralların geliştirilmesi için çalışmalar planlanmaktadır.

Kaynakça

[1] **Daniel, B**., **Miller, M**., **Schwartz, S**. ve **Weischedel**, R., 1997. Nymble: a High-Performance Learning Name-finder, The 5th Conference on Applied Natural Language Processing, Washington, DC, USA, 194-201.

[2] **Satoshi, S**., 1998. Nyu: Description of the Japanese NE System Used For Met-2, The 7th Message Understanding Conference,Virginia, USA.

[3] **Borthwick, A**., **Sterling, J**., **Agichtein, E**. ve **Grishman, R**., 1998. NYU: Description of the MENE Named Entity System as used in MUC-7, Message Understanding Conference, Virginia, USA.

[4] **Masayuki, A**. ve **Matsumoto, Y.**, 2003. Japanese Named Entity Extraction with Redundant Morphological Analysis, The North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL.

[5] **McCallum, A**. ve **Li, W**., 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons, Computational Natural Language Learning-CoNLL, Edmonton, Canada.

[6] **Brin, S**., 1999. Extracting Patterns and Relations from the World Wide Web, The World Wide Web and Databases - LNCS, Vol. 1590, 172-183.

[7] **Cucchiarelli, A**. ve **Velardi, P.,** 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence, Computational Linguistics, Vol. 27:1, 123-131, Cambridge: MIT Press.

[8] **Etzioni, O**., **Cafarella, M. J**., **Downey, D.**, **Popescu, A. M**., **Shaked, T**., **Soderland, S**., **Weld, D. S**. ve **Yates, A**., 2006. Unsupervised Named-Entity Extraction from the Web: An Experimental Study, Artificial Intelligence, Vol. 165, 91-134, Essex: Elsevier.

[9] **Rau, L**., 1991. Extracting Company Names from Text, The 7th IEEE Conference on Artificial Intelligence Applications, 29-32.

[10] **Lee, S**. ve **Lee, G. G**., 2005. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping, Natural Language Processing, Korea, 658-669.

[11] **Fleischman, M**. ve **Hovy, E.**, 2002. Fine Grained Classification of Named Entities, The 19th International Conference on Computational Linguistics, Taiwan.

[12] **Narayanaswamy, M**., **Ravikumar, K. E**. ve **Vijay-Shanker, K**., 2003. A Biological Named Entity Recognizer, The 8th Pacific Symposium on Biocomputing, Hawaii, USA, 427-438.

[13] **Rindfleisch, T. C., Tanabe, L.** ve **Weinstein, J. N.,** 2000. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature, The 6th Pacific Symposium on Biocomputing, Hawaii, USA, 517-528.

[14] **Sekine, S**. ve **Nobata, C**., 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy, The 4th International Conference on Language Resources and Evaluation, Lisbon, Portugual.

[15] **Maynard, D**., **Tablan, V**., **Ursu, C**., **Cunningham, H**. ve **Wilks, Y**., 2001. Named Entity Recognition from Diverse Text Types, Recent Advances in Natural Language Processing, Bulgaria.

[16] **Minkov, E**., **Wang, R**. ve **Cohen, W.**, 2005. Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text, Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, Canada.

[17] **Yu, S**., **Bai S**. ve **Wu, P**., 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7, The 7th Message Understanding Conference, Virginia, USA.

[18] **Poibeau, T**., 2003. The Multilingual Named Entity Recognition Framework, The 10th Conference on European Chapter of the Association for Computational Linguistics-EACL, Budapest, Hungary.

[19] **Cucchiarelli, A**. ve **Velardi, P**., 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence, Journal of Computational Linguistics, Cambridge: MIT Press, Vol. 27:1, 123-131.

[20] **Silva, D**., **Ferreira, J**., **Kozareva, Z**., **Gabriel, J**. ve **Lopes, P**., 2004. Cluster Analysis and Classification of Named Entities, The 4th International Conference on Language Resources and Evaluation, Lisbon, Portugual.

[24] **Cucerzan, S**. ve **Yarowsky, D.**, 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence, The Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, USA 90-99.

[25] **Tür, G**., **Hakkani-Tür D**. ve **Oflazer, K**., 2003. A Statistical Information Extraction System for Turkish, Natural Language Engineering, Vol. 9, No. 2, 181-210.

[26] **Bayraktar, Ö**. ve **Taşkaya-Temizel**, **T**., 2008. Person Name Extraction From Turkish Financial News Text Using Local Grammar-Based Approach, The 23rd International Symposium on Computer and Information Sciences-ISCIS, Istanbul, Turkey.

[21] **Whitelaw, C**. ve **Patrick, J**., 2003. Evaluating Corpora for Named Entity Recognition Using Character-Level Features, The 16th Australian Conference on Artificial Intelligence-AI, 910-921, Perth, Australia.

[22] **Piskorski, J.**, 2004. Extraction of Polish Named-Entities, The 4th International Conference on Language Resources and Evaluation, Lisbon, Portugual.

[23] **Popov, B**., **Kirilov, A**., **Maynard, D**. ve **Manov, D**., 2004. Creation of Reusable Components and Language Resources for Named Entity Recognition in Russian, The 4th International The Conference on Language Resources and Evaluation, Lisbon, Portugual.

[27] **Küçük, D**. ve **Yazıcı, A**., 2008. Identification of Coreferential Chains in Video Texts for Semantic Annotation of News Videos, The 23rd International Symposium on Computer and Information Sciences-ISCIS, Istanbul, Turkey.

[28] **Küçük, D.** ve **Yazıcı, A**., 2009. Exploiting Information Extraction Techniques for Semantic Annotation of Videos in Turkish, The 14th International Conference on Applications of Natural Language to Information Systems-NLDB, Saarland, Germany.

[29] **Küçük, D.** ve **Yazıcı, A**., 2009. Rule-based Named Entity Recognition from Turkish Texts, INnovations in Intelligent SysTems and Applications, Trabzon, Turkey.

[30] http://www.ce.yildiz.edu.tr/personal/zbanu/file