Türkçe Kural Tabanlı Varlık İsmi Tanıma
Named Entity Recognition From Turkish Texts

Faik Erdem Dalkılıç[1], Semih Gelişli[1], Banu Diri[1]

1. Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
faikerdem@gmail.com, sgelisli@gmail.com, banu@ce.yildiz.edu.tr

Özetçe

Varlık İsmi Tanıma, Doğal Dil İşleme biliminin önemli alanlarından biri olup, metinlerde geçen isimleri kişi, yer ve organizasyon ismi olarak ayırmanın yanısıra formül, tarih ve parasal ifadeleri de bulabilmeyi hedeflemektedir. Kural tabanlı varlık ismi tanıma ise birtakım sözlüksel kaynaklar ile bazı kalıplar oluşturup, hazırlanan kurallar ile varlık ismi tanıma işleminin gerçeklenmesidir. Bu çalışmada, konudan bağımsız olarak Türkçe metinlerdeki özel isimlerin kişi, yer ve organizasyon isimleri olarak etiketlenmesi gerçekleştirilmiştir.

Abstract

Named Entity Recognition is an important subject of Natural Language Processing and is used to classify proper nouns into different types such as person, location and organization names in addition to formula, date and money definitions. Rule Based Named Entity Recognition means defining rules to classify named entities in a text through using lexical resources and creating patterns. This study focuses on classification of proper nouns into three types including person, location and organization names regardless of the subject of text.

1. Giriş

Varlık İsmi Tanıma, bilgi çıkarımının bir alt dalı olup, kişi ve kurum ismi, yer, zaman, saat, kısaltma ve para gibi önceden belirlenmiş olan sınıfları dile bağımlı veya dilden bağımsız olarak bir metin içerisinde arayıp bulan sistemlerdir [1].

Bu alandaki ilk çalışmalardan olan [2]'deki yazarlar dilden bağımsız varlık ismi tanıma sistemi geliştirmişler ve Türkçe, İngilizce, Romence, Yunanca ve Hintçe dillerine uygulayıp, değerlendirmesini yapmışlardır. [3]'te yer alan çalışmada ise herhangibir sözlük kullanmadan varlık ismi tanıma sistemi geliştirilmiştir. [4] ise elektronik mektup ve yazılı metin halinde tutulan telefon konuşmalarından özel isimlerin çıkarılması üzerine yapılmış bir çalışmadır. İngilizce için geliştirilmiş varlık ismi tanıma sisteminin bir benzeri [5] tarafından Türkçe için geliştirilmiş ve sonuçlar karşılaştırmalı olarak verilmiştir. [6]'da ise, Türkçe finans haber dokümanları içerisinden kişi isimlerinin çıkarılması için bir çalışma gerçekleştirilmiştir. Arapça dokümanlar ile varlık ismi üzerine çalışan [7] sonuçlarını farklı tipteki dokümanlar üzerinde göstermiştir. Türkçe için ilk kural tabanlı varlık ismi tanıma çalışması yapan [8] farklı türlerdeki dokümanlarda sistemin başarısını ölçümlemiştir.

Bu çalışmada Türkçe için konudan bağımsız kural tabanlı, yer, kurum-kuruluş ve özel isimlerin doküman içerisinden etiketlenerek çıkarılması için bir sistem geliştirilmiştir. Makalenin ikinci bölümünde geliştirilen varlık ismi tanıma sistemi üç aşamada anlatılmaktadır. Bölüm üçte ise mevcut kısıtlar hakkında bilgi verilmiş, bölüm dörte de deneysel sonuçlardan bahsedilmiştir.

2. Varlık Tanıma için Geliştirilen Sistem

İşlenecek olan doküman içerisinde öncelikli olarak sözcüğün büyük harfle başlayıp başlamadığının denetimi yapılır. Ardından büyük harfle başlayan sözcüklerin Zemberek API [9] kullanılarak özel isim olup olmadıkları belirlenir. Daha sonra her bir sınıfa ait kurallar teker teker denetlenir. Bunlardan farklı, tercihe bağlı olarak geçici isim havuzları oluşturularak doküman içerisinde ilk kez etiketlenmiş varlıklara tekrar rastlanıldığında bu isim havuzları kullanılarak varlıkların tanınması sağlanmaktadır. Bu isim havuzları dokümanın işlenmesi tamamlandığında boşaltılır. Bu çalışmada varlık tanıma üç kısımda ele alınarak incelenmiştir.

2.1. Yer İsimleri

Yer isimlerinin bulunmasında sözlüksel kaynak olarak adlandırılamayacak kadar küçük boyuttaki veritabanlarından yararlanılmıştır. Bu veritabanı coğrafi isimleri belirtmekte sıklıkla kullanılan tamlamalardaki tamlanan ve tamlayan sözcükler seçilerek oluşturulmuştur. Böylece bu tamlananlar kullanılarak Ağrı Dağı, İstiklal Caddesi gibi yer isimlerinin tespit edilebilmesi sağlanmıştır. Ayrıca yön bildirimlerinde kullanılan kuzey, güney, doğu, batı sözcükleri de tamlayan olarak kullanılmıştır ve bu sayede Doğu Anadolu, Güney Avrupa gibi yer isimleri de etiketlenebilmiştir.

Bir başka yöntem olarak Türkçe 'de özellikle ülke isimlerinde kullanılan (~%28) -ye, -ya ve -istan ekleri ile biten Türkiye, Almanya, Bulgaristan gibi sözcükler yer ismi olarak ayrıştırılabilmişlerdir. Bunlara benzer olarak, Türkçe'de coğrafya isimleri türetmekte kullanılan ve yer isimlerinde sıklıkla geçen -köy, -deniz gibi sözcüklerle türetilmiş Kadıköy, Karadeniz gibi bileşik kelimelerde yer isimleri olarak tespit edilmiştir.

Ayrıca yer isimlerinin bulunmasında ismin hal eklerinden olan ve konum bildiriminde kullanılan -de ve ayrılma bildiriminde kullanılan -den durum eklerinden de yararlanılmıştır. Bu sayede İstanbul'da kelimesindeki İstanbul, yer ismi olarak etiketlenebilmiştir.

2.2. Kurum Kuruluş İsimleri

Organizasyon isimlerinin tespit edilmesinde de yer isimlerinin bulunmasına benzer şekilde az sayıda sözcük içeren

978-1-4244-9671-6/10/$26.00 ©2010 IEEE

veritabanlarından yararlanılmıştır. Bu sözlük içerisinde şirketlerin resmi isimlendirilmelerinde kullanılan A.Ş ve Ltd. Şti. gibi ünvanlarla, kurumların niteliklerini belirten Vakfı, Üniversitesi gibi ayırt edici sözcüklerden yararlanılarak, Yıldız Teknik Üniversitesi, Galatasaray A.Ş gibi kurum-kuruluş isimleri etiketlenebilmiştir.

Bunlardan farklı olarak kurum-kuruluş isimlerinin kısaltma olarak kullanımlarındaki yazım kurallarından faydalanılarak THY, TÜSİAD gibi şirket ve oluşum isimleri de sınıflandırılmıştır.

2.3. Kişi İsimleri

Bu çalışmada kişi isminin belirlenmesinde herhangibir isim sözlüğü kullanılmamıştır. Kişi isimlerinin etiketlenmesinde iki yöntem önerilmiştir.

Bunlardan birincisi, Türkçe'de ünvanlar kişi isimlerinden önce gelebildikleri gibi (Dr. Ahmet Mertoğlu), kişi isminden sonra gelen ünvanlar da (Ahmet Bey) bulunmaktadır. Bu çalışmada da iki farklı ünvan türüne ait sık kullanılan öğeler küçük boyutta bir veritabanında tutularak kişi isimlerinin etiketlenmesi sağlanmıştır.

İkinci yöntem ise, insana özgü olan fiillerin kullanılmasıdır. Özlemek, sevmek, okumak, söylemek gibi fiillerin olduğu cümlelerdeki isimler bu yöntem sayesinde kişi ismi olarak etiketlenebilmiştir.

3. Kısıtlar

Kullanılmış olan kural tabanlı yöntemlerin varlık isimlerini etiketlemesinde yüzde yüz başarılı sonuçların alınamaması bazı kısıtlar yüzündendir. Sistemin başarısını düşüren kısıtlar üç varlık ismi üzerinde ayrı ayrı ele alınarak incelenmiştir.

3.1. Yer İsimleri-kısıt1

Yer isimlerini belirlemekte kullandığımız -ye, -ya ekleri gerçekte yer ismi olmayan Açelya, Aliye gibi özel isimlerde de bulunmaktadır. Benzer şekilde ismin durum eklerini kullanarak yer isimlerini etiketlerken kullandığımız -de, -den gibi ekler, "Ali'den aldığım gömlek bedenime uymadı." cümlesindeki Ali'yi yer ismi olarak etiketlemektedir.

Ayrıca kişilere verilen bazı coğrafi isimler de doğruluk oranını etkileyen bir diğer etken olarak karşımıza çıkmaktadır. Bir yön ismi olan Güney, Kuzey Türkçe'de aynı zamanda kişi ismi olarak da kullanılmaktadır.

Ek olarak, Türkçe'deki tüm metinler resmi dil kullanılarak yazılmadığı için hitap edilen kitle tarafından bilinmesine istinaden herhangi bir yer ifadesi onu tamlayacak sözcük kullanılmadan yazılmaktadır. Örneğin, İstiklal Caddesi büyük bir okuyucu kitlesi tarafından bilindiği için "İstiklal Caddesi'ne doğru yürüyordum." yerine "İstiklal'e doğru yürüyordum." şeklinde kullanılabilir. Bunun gibi durumlar başarımızı düşüren önemli kısıtlardandır.

3.2. Kurum Kuruluş İsimleri-kısıt2

Kurum-kuruluş isimlerinde de yer isimlerinin etiketlenmesinde karşılaşılan ve resmi dille yazılmamış metinlerdeki kullanım farklılıklarından ileri gelen durumlar söz konusu olabilir. Örneğin, Galatasaray A.Ş. yerine çoğunluk tarafından bilindiği için Galatasaray ifadesinin kullanımı gibi durumlar başarımı düşürmüştür.

3.3. Kişi İsimleri-kısıt3

Kişi isimlerinin herhangibir sınırlandırma olmadan farklı şekillerde kullanılmaları kişi isimlerinin etiketlenmesinde önemli bir sorundur. Örneğin, "Ayşe Öztürk" gibi isim ve soyisimden oluşan bir kullanım olduğu gibi, sadece Ayşe veya sadece soyismi kullanarak "Öztürk" ile de aynı kişi ifade edilebilmektedir. Bu da sistemin başarımını ciddi oranda etkilemektedir. Ancak metin içerisindeki isimlerin etiketlenmesinde geçici sözcük havuzları kullanılarak başarı oranı arttırılabilmektedir.

Kullandığımız kurallara bağlı olarak insana özgü fiillerden yararlandığımızda ise aynı cümle içerisinde bulunan birden fazla özel ismi etiketlememizde hata miktarımız artmaktadır. "Mustafa Sarıgül, Alevilik ile ilgili görüşlerini bildirdi." şeklinde bir cümlede "Mustafa Sarıgül" kişi ismi olarak belirlendiği gibi "Alevilik" de kişi ismi olarak bulunmaktadır.

4. Deneysel Sonuçlar

Geliştirilen sistemin test edilmesinde seçilen metinler yapılan çalışmanın konudan bağımsız olduğunu göstermek amacıyla siyaset, ekonomi ve sağlık alanından seçilmiştir. Her sınıftan 10 adet doküman alınarak toplam 30 farklı metin dosyası üzerinde denemeler gerçekleştirilmiştir. Başarı oranının ölçülmesinde de eşitlik-1'deki kesinlik (precision), (2)'deki çağrı (recall) ve (3)'deki kesinlik ve ölçüm değerlerinin harmonik ortalaması olan F-ölçüm (F-measure) kullanılmıştır.

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

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

F-Ölçüm=(2 * P * R) / (P + R) (3)

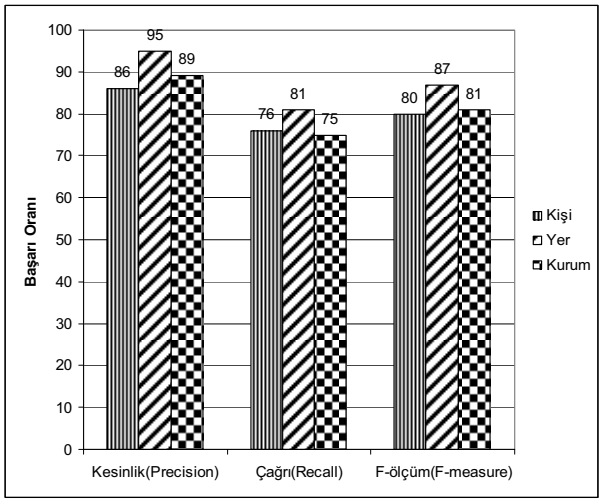
Tablo-1'de üç farklı sınıfta yer alan toplam sözcük sayıları, her sınıf içerisindeki yer, kurum ve isim varlık sayıları, sistem tarafından tespit edilen ve elle yapılmış kontroller sonrasında doğru olarak tespit edilen varlık sayıları verilmektedir.

Tablo 1. Herbir sınıftaki varlık sayısı, tespit edilen ve doğru olarak kabul edilen varlık isim sayıları

Varlık Sayısı	Ekonomi+Siyaset+Sağlık **(3 x 10 doküman)**	**Tespit edilen**	**Doğru kabul edilen**
Kişi	49+113+59=221	195	168
Yer	34+74+71=179	153	145
Kurum	54+33+61=148	125	111
sözcük sayısı	2883+1505+3439=7827		

Şekil-1'de ise üç varlık isminin başarı ölçüm kriterleri verilmiştir. Geliştirilen kurallar ile en başarılı şekilde bulunan varlık *yer* olmuştur. İkinci sırada ise *kurum* gelmektedir. En

düşük performansı *kişi* isimleri göstermiş olup, sebepleri kısıt3'te anlatılmıştır.



*Şekil 1.*Üç farklı varlık isminin başarı ölçüm değerleri

Tablo 2'de ise, kişi, yer ve kurum varlık isimlerinin dokümanın türüne bağlı olarak F-ölçüm değerleri verilmiştir. Buna göre, varlık isimlerinin tespit edilmesinde siyaset türündeki yazılar daha başarısız sonuçlar vermiştir. Sağlık türündeki yazılarda yer ve kurum isimlerinin doğru olarak tespiti daha yüksek bir başarıya sahip iken, kişi isimlerinin bulunmasında ise ekonomi türündeki yazılar daha başarılı olmuştur.

Tablo 2. Dokümanın türüne göre varlık isimlerinin F-ölçüm değerleri

	Kişi isimleri	**Yer isimleri**	**Kurum isimleri**
Ekonomi	0.88	0.90	0.82
Siyaset	0.78	0.84	0.73
Sağlık	0.79	0.90	0.85

Çalışmanın daha sonraki bölümünde kişi isimlerinin bulunmasında başarıyı attıracak yeni kuralların eklenmesi ve yer, kurum-kuruluş ve kişi isimlerinin dışında tarih, para, formül gibi diğer varlıklarında bulunması planlanmıştır. Ayrıca özel isimlerin bulunmasında ortaya atılmış olan kişilere özgü eylemlerin daha da genişletilerek sisteme verilmesi düşünülmektedir.

5. Sonuç

Varlık İsmi Tanıma, bilgi çıkarımı içerisinde kendisine yer edinsede, dile bağlı olarak geliştirdiğimiz kurallar ile doğal dil işleme altında da kendisine yer bulmuştur. İngilizce'de Named-Entity Recognition-NER olarak adlandırılan Varlık İsmi Tanıma üzerine yapılan çalışmaların çoğu İngilizce, Çince ve İspanyolca üzerine olmuştur. Bu konuda Türkçe için çalışmalara yeni yeni başlanmıştır. Bu çalışmada Türkçe için dilin bazı gramatik kurallarıda kullanılarak kural tabanlı bir yöntem geliştirilmiş ve farklı türlerdeki dokümanlarda denenip, %87 ile *yer* isimlerinde en yüksek başarı elde edilmiştir. En düşük başarı *kişi* isimlerinin tanınması olsa bile burada da %80'lik başarı elde edilmiştir.

Bu çalışmanın devamında sistemi farklı alanlardaki dokümanlarda test etmek, kişi isimlerinin bulunmasında farklı kuralların geliştirilmesi ve bu üç varlığın dışında tarih, saat, formül, vs. gibi diğer isim varlıkların tanınmasınında sisteme eklenmesi için çalışmalar planlanmaktadır.

6. Kaynakça

[1] Grishman, R., "Information Extraction", *In The Oxford Handbook of Computational Linguistics*, R. Mitkov (Eds.), Oxford University Press, 2003.

[2] Cucerzan, S., and Yarowsky, D., "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence", *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[3] Mikheev, A., Moens, M. and Grover, C., "Named Entity Recognition without Gazetteers", *Proceedings of the 9th Conference of the European ACL* (EACL 99), 1999.

[4] Poibeau, T., *"Proper Name Extraction from Non-Journalistic Texts", Proc. Computational Linguistics in the Netherlands* May 30, 2001.

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

[6] Bayraktar, Ö. and Taşkaya-Temizel, T., "Person Name Extraction From Turkish Financial News Text Using Local Grammar- Based Approach", *In P roceedings of t he International Symposium on Computer and Information Sciences (ISCIS)*, 2008.

[7] Shaalan, K. and Raza, H., "Arabic Named Entity Recognition from Diverse Text Types", *In Proceedings of the Intern ational Conference on Natural Language Processing (GoTAL)*, 2008.

[8] Küçük, D., Yazıcı, A., "Rule-based Named Entity Recognition from Turkish Texts", *International Symposium on INnovations in Intelligent SysTems and Applications*, Trabzon, Turkey, June 29-July 1, 2009.

[9] https://zemberek.dev.java.net/