Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 32, 23-30, 2014

Research Article / Araştırma Makalesi

PREDICTION OF FUNCTION TAGS OF THE SIMPLE TURKISH SENTENCES BY CONDITIONAL RANDOM FIELDS

Mustafa AYGÜL, Gürkan KARAALİOĞLU, M. Fatih AMASYALI*

Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Müh. Bölümü, Davutpaşa-İSTANBUL

Received/Geliş: 19.06.2013 Revised/Düzeltme: 01.10.2013 Accepted/Kabul: 24.10.2013

ABSTRACT

The prediction of function tags is a key component of several natural language tasks. In this study, Conditional Random Fields are employed for Turkish sentences. The affects of the size of training set, the usage of morphological features of the words are investigated. As a result, we achieved 75% success ratio on our datasets having 2000 simple sentences.
Keywords: Natural language processing, dependency parsing, artificial intelligence, sequence labeling, conditional random fields.

KOŞULLU RASTGELE ALANLARLA BASİT TÜRKÇE CÜMLELERİN ÖĞELERİNE AYRILMASI

ÖZET

Doğal dil işleme çalışmalarında, cümlelerin otomatik olarak bileşenlerine/öğelerine ayrılabilmesi birçok uygulama için gereklidir. Bu çalışmada basit Türkçe cümleler için bu işlemi Koşullu Rastgele Alanlar'ı kullanarak gerçekleştiren bir araç geliştirilmiştir. Eğitim setinin büyüklüğünün ve kelimelerin morfolojik özelliklerinin kullanımının etkileri araştırılmıştır. Sonuç olarak 2000 basit cümleden oluşan veri kümemizde %75'lik doğruluk oranına erişilmiştir.
Anahtar Sözcükler: Doğal dil işleme, öğelerine ayırma, yapay zeka, dizi etiketleme, koşullu rastgele alanlar.

1. GİRİŞ

Cümleler anlamlı kelime birlikteliklerdir. Cümleler öğelerden, öğeler ise kelime veya kelime gruplarından oluşur. Türkçede cümleler içerdikleri fiil türündeki kelime sayısına bağlı olarak basit ya da bileşik cümleler olmak üzere ikiye ayrılırlar. Birden fazla fiil içeren bileşik cümlelerde yan cümleler ana cümlenin bir öğesidir. Örneğin "Okula giderken bir fare göndüm" cümlesi "giderken" ve "gördüm" olmak üzere iki adet fiil türü kelime içerdiğinden bileşik cümledir. "okula giderken" yan cümlesi ana cümlenin "zarf tümleci" öğesidir.

Türkçede özne, belirtili nesne, belirtisiz nesne, dolaylı tümleç, zarf tümleci ve yüklem olmak üzere 6 farklı öğe türü bulunmaktadır [1].

Türkçede yüklem cümledeki eylemi belirtirken diğer öğeler yükleme sorulan sorularla belirlenir. Özne ve belirtisiz nesne kim/ne sorularına, belirtili nesne kimi/neyi sorularına, dolaylı

*Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: mfatih@ce.yildiz.edu.tr, tel: (212) 383 57 30

tümleç nerede, nereden sorularına, zarf tümleci ise zaman, nasıl, ne için, neyle, kimle sorularına cevap verir. Çalışmamızda Zarf tümleci "zaman, nasıl, ile, için, kadar" öğe türlerine bölünmüştür. Bunun sebebi daha belirgin öğe türlerini bulabilmektir.

Cümlelerin öğelerinin bulunmasında kelimelerin türleri, aldıkları ekler, kelimelerin anlamları ve metnin bağlamı kullanılmalıdır. "Hastalıktan öldü." cümlesi ile "Denizden çıktı." cümlelerinin morfolojik çözümleri birbirinin aynı olmasına rağmen "hastalıktan" kelimesi sebep belirttiğinden zarf tümleci, "denizden" kelimesi ise dolaylı tümleçtir. Aynı morfolojilere sahip bu iki kelimenin öğe türlerinin doğru belirlenmesi için kelimelerin anlamlarına da ihtiyaç duyulduğu görülmüştür.

Ortaokul ve liseden bildiğimiz cümleleri öğelerine ayırma işlemini bilgisayarlara yaptırabilmek (dolayısıyla otomatikleştirebilmek) bilgi çıkarımı, diyalog sistemleri, metin sınıflandırma, metin anlama gibi çeşitli doğal dil işleme problemlerinin daha iyi/doğru çözülebilmesini sağlamaktadır.

Cümleleri öğelerine ayırma problemi (X1, X2, X3) gibi bir dizilimden (Y1, Y2, Y3) gibi etiket diziliminin üretilmesi olarak formelleştirilebilir. Xi'ler kelimeleri ve aldıkları ekleri ifade ederse, Yi'ler öğe türlerini ifade edecektir. Bu öğrenme probleminin çözümü için literatürde, Saklı Markov Modelleri, Maksimum Entropili Markov Modelleri ve Koşullu Rastgele Alanlar-CRF [2] olmak üzere çeşitli yöntemler önerilmiştir. Literatürdeki çeşitli çalışmalar göstermiştir ki bu yöntemlerden en iyisi CRF'dir [3]. Bu nedenle biz de çalışmamızda CRF'yi kullandık.

Bu çalışmada basit Türkçe cümlelerin öğelerine otomatik olarak ayıran bir sistem geliştirilmiştir. Makalenin 2. bölümde bu konudaki benzeri çalışmalar özetlenmiştir. 3. bölümde önerilen sistemin bileşenleri, 4. bölümde kullanılan cümleler kümesi tanıtılmıştır. 5. bölümde elde edilen sonuçlar verilmiştir. Son bölümde ise sonuçlar yorumlanmış, sistemin eksiklikleri ve olası gelecek çalışmalar anlatılmıştır.

2. BENZER ÇALIŞMALAR

Çalışmamızın konusuna en yakın çalışma Özköse ve Amasyalı tarafından yapılmıştır [4]. Çalışmalarında basit Türkçe cümlelerin öğelerini bulmuş ve öğe ikililerinden hayat bilgisi çıkarımı yapmışlardır. Çalışmanın asıl amacı bilgi çıkarımı olduğundan öğelere ayırma işleminin doğruluk oranı ölçümlenmemiştir. Öğeleri bulmak için elle üretilmiş kural tabanlı bir yöntem kullanmıştır. Kuralların elle üretilmesi oldukça fazla zaman alıcı bir işlem olduğundan bu çalışmada bu kuralları eğitim kümesinden otomatik öğrenen bir araç kullanılmıştır.

Türkçe cümlelerin öğelerinin bulunması için yapılan bir başka çalışma Coşkun tarafından yapılmıştır [5]. Bu çalışmada elle hazırlanan kural tabanlı bir yapı kullanılmıştır. Bunun haricinde İngilizce için birçok örneği olmasına rağmen [6] CRF'yi Türkçe cümle öğelerini bulmada kullanan bir başka çalışma bulunmamaktadır. Bununla birlikte CRF'yi Türkçe Varlık İsmi Tanıma (Name Entity Recognition) için kullanan birkaç çalışma vardır. Bunlardan biri Şeker ve Eryiğit [7] tarafından haber metinleri üzerinde yapılmış bir çalışmadır. Bir diğer çalışma ise Özkaya ve Diri [8] tarafından yapılmış ve email metinleri üzerindeki çalışmadır. Her iki çalışmada da 3-4 farklı varlık isim türü (kişi ismi, yer ismi, kurum ismi vb.) bulunmaya çalışılmış ve %90'a yakın başarılar elde edilmiştir.

Ayrıca Eryiğit'in Türkçe için bağımlılık çözümlemesi (dependency parsing) çalışması bulunmaktadır [9]. Bu çalışma ile kelimelerin birbirleriyle bağımlılıkları bulunmuştur.

3. KOŞULLU RASTGELE ALANLARLA CÜMLELERİ ÖĞELERİNE AYIRMA

Bir dizi etiket için yine bir dizi etiket üretmeyi öğrenmek için en çok kullanılan algoritmalardan biri CRF'dir. Bu çalışmada en popüler CRF uygulaması olan CRF++ [10] kullanılmıştır.

Eğiticili tüm yöntemlerde olduğu gibi CRF'de de önce bir eğitim setinden etiket üretiminin kuralları/modeli öğrenilmektedir. Daha sonra da test verileri üzerinde öğrenilen

kurallara/modele göre etiketleme yapılmaktadır. CRF'lerin en yaygın türü olan Doğrusal Zincir (Linear Chain) tabanlı CRF ile *x* giriş dizisine ait *y* etiketlerinin bulunması koşullu olasılığı Eşitlik 1'de verilmiştir.

$$P_\lambda(y \mid x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

Eşitlik 1'deki Z_x normalizasyon faktörüdür. $f_k(y_{t-1}, y_t, x, t)$ ise bir fonksiyondur. Bu fonksiyon öğe türü belirleme için örneğin y_{t-1} = "Özne" ise ve y_t ="Belirtili Nesne" ise ve t anındaki x (kelimemiz) = "top" ise 1 değilse 0 değerini alır. K, fonksiyon sayısını, T pencere uzunluğunu göstermektedir. λ_k ise, eğitim kümesindeki etiketli cümlelerden öğrenilen ağırlık değerleridir. Yapay sinir ağlarının eğitimi, etiketli örneklerden katmanlar arası ağırlıkların öğrenilmesi olduğu gibi, CRF'nin eğitimi de etiketli cümlelerden bu ağırlıkların (λ) iteratif olarak öğrenilmesidir.

Fonksiyonlardaki pencere boyutu kullanıcı tarafından belirlenir. Fonksiyonlarda kelimelerin ve etiketlerin birbirlerinden sonra gelme ihtimalleri yer aldığından fonksiyon sayısının çok fazla olacağı açıktır. Öğrenilmesi gereken ağırlık sayısı da fonksiyon sayısı kadardır. Öğrenilmesi gereken değişken sayısının çok fazla olduğu bu tür optimizasyon problemlerinde gerekli iterasyon sayısını çok düşürdüğünden, CRF'lerin eğitimi için genelde sınırlı hafızalı Broydon - Fletcher - Goldfarb - Shanno (L-BFGS) algoritması kullanılmaktadır [11].

Fonksiyonların yapısına dikkat edildiğinde *y* çıktısının tahmininde sadece belirli önceki ya da sonraki x'lerin değil tüm *x*'lerin kullanılabildiği görülmektedir. Bu sayede CRF, uzak bağımlılıkların da işleme katılabilmesine imkan vermektedir. Bununla birlikte kural sayısı da artmakta ve parametre optimizasyonu güçleşmektedir. Bu nedenle bu çalışmada pencere boyutu olarak [-2,+2] kullanılmıştır.

Birinci bölümde anlatıldığı gibi öğeleri bulmak için kelimelere, kelime türüne, eklere, kelimenin anlamına ve metnin bağlamına ihtiyaç vardır. Bu çalışmada kelimelerin kendileri, türleri ve aldıkları ekler kullanılmıştır.

CRF dizi etiketlemeyi öğrenirken giriş dizisinin birden fazla boyutlu olmasına izin vermektedir. Çıkış dizisi ise tek boyutlu olmalıdır.

Çalışmamızda giriş dizisinin ilk boyutunu kelimenin kendisi, ikinci boyutunu ise kelimenin morfolojik çözümlemesi olarak kullanılmıştır. Kelimeler, karakter katarı (string) olarak, morfolojik çözümlemesi ise 1 ve 0'lardan oluşan bir karakter katarı olarak ifade edilmiştir. Kelimelerin morfolojik çözümlemesinin ifadesi için Zemberek [12] kullanılmıştır. Zemberek'in kelimenin çözümlemesi için verdiği ilk sonuç doğru kabul edilmiştir. Zemberek morfolojik çözümleme için 113 farklı etiket üretmektedir. Buna göre kelimenin morfolojik çözümlemesinde yer alanlar 1, yer almayanlar 0 olmak üzere 113 elemanlı bir dizi ile kelimenin morfolojik çözümlemesi ifade edilmiştir.

Şekil 1'de eğitim setinden bir cümlenin (Akşam çay bahçesinde arkadaşlarıyla buluşacak) CRF'ye verilen hali gösterilmektedir.

Akşam
00000000000000000000000000000000000000000000000000000000001000000000000000000000
000000000000000000000000000000000000 Zaman
çay
01000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000 DT
bahçesinde
01000000000101000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000 DT
arkadaşıyla
01000010000001000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000 İle
buluşacak
10000000000000000000000000000000000000000000000000000000000010000000000000000000
000000000000000000000000000000000000 Y
.
00000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000 O

Şekil 1. "Akşam çay bahçesinde arkadaşlarıyla buluşacak" cümlesinin işlenmiş hali

Cümlede her satırda bir kelime bulunmaktadır. Her satırda kelimenin kendisi ardından 113 elemanlı morfolojik çözümlemesi, ardından da çıkışta öğrenilmesini istediğimiz etiket (öğe türü) bulunmaktadır. Öğe türü kısaltmalarının açıklamaları Çizelge 1'de verilmiştir. CRF'ye verilen giriş dizisinin ilk boyutu kelimenin kendisi, ikinci boyutu ise o kelimenin morfolojik çözümlemesi olmaktadır.

4. KULLANILAN VERİ KÜMESİ

Sistemin eğitiminde ve test edilmesinde kullanılacak cümleler "Bilkent Üniversitesi Bilgiye Erişim Grubu" tarafından haber metinlerinden derlenmiş olan Bilcol [13] derleminden seçilmiştir. Derlemden içinde tek fiil içeren 2000 cümle rastgele seçilmiştir. Cümleler minimum 2, maksimum 10, ortalama 5 adet kelime içermektedir. Cümlelerin içerdikleri öğe türleri ve sayı/oranları Çizelge 1'de verilmiştir. Cümleler toplamda 10500 kelime içermektedir. Çizelge 1'deki oranlar kelimelerin yüzde kaçının o öğe türü olduğunu göstermektedir.

Çizelge 1. Denemelerde kullanılan cümlelerin öğe türlerinin oranları

Kısaltma	Açıklama	Sayı	Oran (%)
Ozne	Özne	1769	16,85
DT	Dolaylı Tümleç	1359	12,94
Bsiz	Belirtisiz Nesne	569	5,42
Bli	Belirtili Nesne	952	9,07
Y	Yüklem	2234	21,28
O	Noktalama işareti	2105	20,05
Kadar	Kadar	143	1,36
Zaman	Zaman	611	5,82
Nasıl	Nasıl	470	4,48
İle	Birliktelik	137	1,30
İçin	Sebep	151	1,44

Veri kümesinde yer alan cümlelere örnek olarak "Aşığa Bağdat sorulmaz", "Çin, ABD'yi kaygılandırıyor.", "Biz görevimizi yapıyoruz." cümleleri verilebilir. Veri kümesinin tamamı http://www.kemik.yildiz.edu.tr/?id=28 adresinden elde edilebilir.

5. DENEYSEL SONUÇLAR

Cümlelerin otomatik olarak öğelerine ayrılması için önerdiğimiz yöntemin performansını ölçmek için 2000 cümlenin 500 tanesi test cümlesi olarak kullanılmıştır.

Öğrenme problemlerinde eğitim setinin büyüklüğü ve verilerin temsili 2 büyük problemdir.

Verilerin temsilinde (giriş dizisinin boyutları) sadece kelimeleri kullanmak ve kelimelerle birlikte morfolojik çözümlemelerinin de kullanmanın performansı 1500 eğitim cümlesi ile 500 test cümlesi üzerinde karşılaştırılmıştır. Performans ölçümünde ilk 1 ve ilk 3 değerler bulunmuştur. İlk 1 değeri, modelin en yüksek olasılık verdiği öğe türünün doğru olma oranıdır. İlk 3 değeri ise modelin en yüksek olasılık verdiği 3 öğe türünden herhangi birinin doğru olma olasılığıdır. Buna göre elde edilen sonuçlar Çizelge 2'de verilmiştir.

Çizelge 2. Sadece kelimeler ve kelimelerle morfolojik özellikleri kullanımının karşılaştırılması

Giriş dizisi boyutları	**İlk 1 Doğruluk Oranı**	**İlk 3 Doğruluk Oranı**
Kelime	% 57	% 68
Kelime + kelime morfolojisi (113 özellik)	% 72	% 78

Çizelge 2'de görüldüğü gibi giriş dizisinde 2 boyut kullanmak (kelime ve morfoloji) başarıyı arttırmaktadır ki bu beklenen bir sonuçtur. Çünkü öğe türlerinin belirlenmesinde morfolojik özelliklerin önemi bilinmektedir.

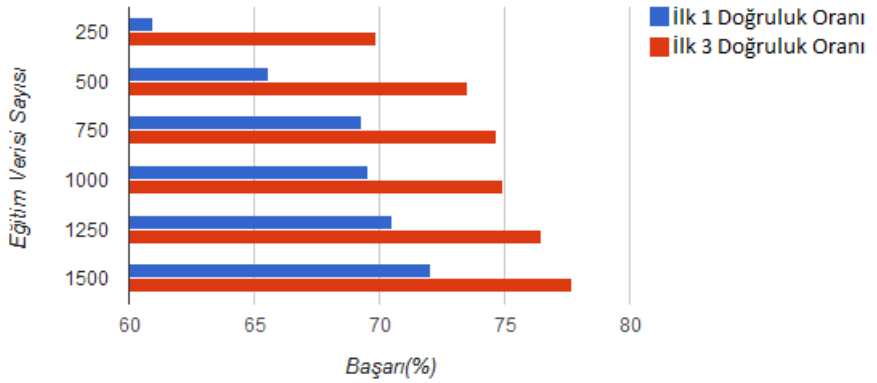
Kelime morfolojisini ifade eden 113 adet özelliğin hepsinin yerine bir kısmının kullanımının sistemin performansını nasıl etkileyeceği de bu çalışma içerisinde araştırılmıştır. Bu denemeden elde edilen sonuçlar Çizelge 3'te verilmiştir. Morfolojik özelliklerin sayısını azaltmak için 2 yöntem denenmiştir. Birincisi, özellik seçim yöntemlerinden CFS (Correlation-based Feature Selection) [14] ile seçim yapılmasıdır. İkincisi için ise önce 113 özelliğin her birinin tek başına kullanıldığındaki sistemin performansı ölçülmüş, performansı en yüksek olan 36 tanesi birlikte kullanılmıştır. Performans ölçümünde yine aynı 1500 eğitim cümlesi ve 500 test cümlesi kullanılmıştır.

Çizelge 3. Morfolojik özelliklerin azaltılmasının etkileri

Giriş dizisi boyutları	**İlk 1 Doğruluk Oranı**	**İlk 3 Doğruluk Oranı**
Kelime + kelime morfolojisi (113 özellik)	% 72	% 78
Kelime + kelime morfolojisi (CFS ile seçilmiş 11 özellik)	% 60	% 68
Kelime + kelime morfolojisi (113'ün tek başına en iyi 36 özelliği birlikte)	% 71	% 76

Çizelge 3 incelendiğinde, CFS ile özellik seçiminin başarıyı düşürdüğü ancak tekil performanslardan seçilen 36 özelliğin başarıyı çok az düşürdüğü görülmüştür.

Eğitim kümesinin boyutunun sitem üzerindeki etkisini görmek içinde denemeler yapılmış ve sonuçları Şekil 2'de verilmiştir. Denemelerde 113 özelliğin tümü kullanılmıştır. Eğitim kümesinin boyutu 250 cümleden başlayarak 250'şer artımla 1500'e kadar çıkarılmıştır. Üretilen modellerin hepsi aynı 500 test cümlesi üzerinde çalıştırılmıştır.



Şekil 2. Eğitim kümesinin büyüklüğünün performansa etkisi

Şekil 2 incelendiğinde eğitim setinin büyüklüğü ile sistemin performansı arasında doğru orantı olduğu açıkça görülmektedir. Ayrıca artışın henüz bir doyuma ulaşmadığı da gözlenmektedir. Diğer bir ifadeyle eğitim setinin boyutu daha da arttırılırsa başarının daha da artacağı söylenebilir.

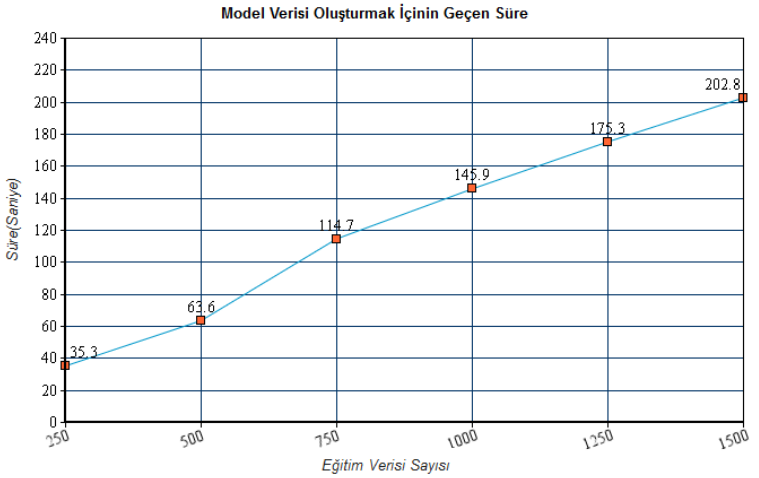
Sistemin elde ettiği en yüksek başarıya ait hata matrisi Çizelge 4'te verilmiştir. Satırlardakiler gerçek öğe türlerini sütunlardakiler ise tahmin edilen öğe türlerini göstermektedir. Örneğin, gerçek 305 öznenin 203'ü özne olarak 36'sı Dolaylı Tümleç olarak, 16'sı Belirtisiz Nesne olarak bulunmuştur.

Çizelge 4. Öğe türlerinin karışım matrisi (Satırlar gerçek, sütunlar tahmin edilen değerler)

Öğe	**Ozne**	**DT**	**Bsiz**	**Bli**	**Y**	**O**	**Kadar**	**Zaman**	**Nasıl**	**İle**	**İçin**	**Toplam**	**D. Oranı**
Ozne	203	36	16	35	2	0	0	6	7	0	0	305	67
DT	32	278	6	21	8	0	0	4	7	0	0	356	78
Bsiz	45	24	58	23	14	0	5	3	13	0	1	186	31
Bli	61	53	18	148	0	0	0	4	7	0	0	291	51
Y	4	0	10	2	503	0	1	0	9	0	0	529	95
O	0	0	0	0	0	503	0	0	0	0	0	503	100
Kadar	0	3	0	0	1	0	16	0	0	0	0	20	80
Zaman	33	54	7	20	0	0	0	51	1	0	0	166	31
Nasıl	7	6	4	1	3	0	1	0	28	0	0	50	56
İle	26	22	5	10	0	0	0	0	5	0	0	68	0
İçin	0	0	0	0	0	0	0	4	0	0	34	38	89

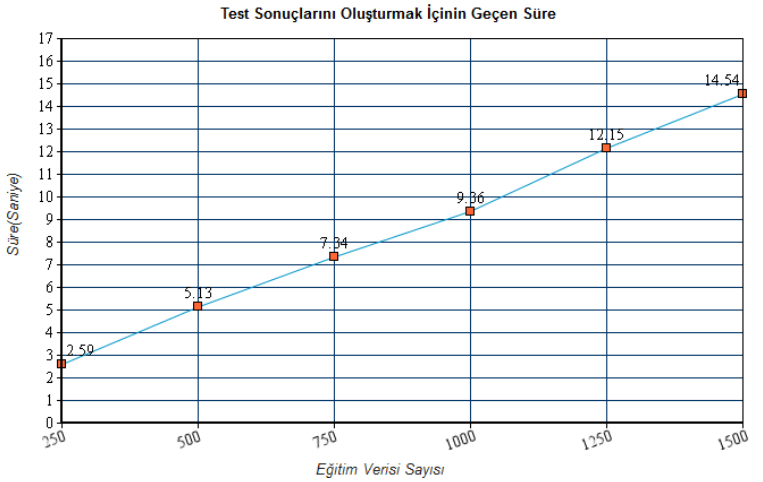
Çizelge 4 incelendiğinde en başarılı bulunan öğe türünün "Noktalama işareti" olduğu görülmektedir. Bunu takip eden öğe türlerinin de "Yüklem" ve "İçin" olduğu görülmektedir. En başarısızlar ise "ile", "Belirtisiz Nesne" ve "zaman" dır. "Özne" en çok "Dolaylı Tümleç" ve "Belirtili Nesne" ile, "Belirtisiz Nesne" ise en çok "Özne" ile karıştırılmıştır.

Sistemin ölçeklenebilirliğinin testi için model üretme ve test süreleri ölçülmüştür. Şekil 3'te eğitim cümlelerinden modelin üretilme sürelerinin eğitim kümesinin boyutuna göre değişimi verilmiştir.



Şekil 3. Model üretim sürelerinin eğitim veri boyutuyla ilişkisi

Şekil 4'te 500 cümlenin test işleminin eğitim kümesinin boyutuyla değişimi gösterilmiştir.



Şekil 4. Test süresinin eğitim veri boyutuyla ilişkisi

Şekil 3 ve 4 incelendiğinde veri boyutuyla işlem sürelerinin lineer olarak arttığı ve bu sebeple ölçeklenebilir olduğu görülmektedir.

5. SONUÇLAR VE GELECEK ÇALIŞMALAR

Türkçe basit (tek fiil içeren) cümlelerin otomatik olarak öğelerinin bulunmasını amaçlayan çalışmamızda Koşullu Rastgele Alanlar kullanılarak %75'lik bir başarı elde edilmiştir. Çalışmada elde edinilen bulgular aşağıda sıralanmıştır.

- Eğitim kümesinin boyutunun artışıyla, sistemin test kümesi üzerindeki başarısının artışı görülmüştür.

- Morfolojik özelliklerin kullanımı da sistemin başarısını arttırmıştır.
- Morfolojik özelliklerden hepsinin yerine bir kısmının kullanımı başarıyı iyileştirmemiştir.
- Uygulama öğrencilerin eğitiminde, çeşitli doğal dil işleme uygulamalarında kullanılabilir.

Gelecek çalışma olarak basit cümlelerin ötesinde fiilimsi de içeren bileşik cümlelerin öğelerine ayrılması hedeflenmiştir. Ayrıca kelimelerin morfolojik özelliklerinin yanı sıra kelimelerin büyük/küçük harf ile yazılma özelliklerinin de kullanımı da düşünülmektedir. Bununla birlikte sistemin daha da geliştirilmesi için kelime anlamları (yer ismi, eşya ismi vb. kategoriler) ve metin bağlamı (önceki ve sonraki cümleler) da giriş boyutlarına dahil edilebilir.

REFERENCES / KAYNAKLAR

[1] Milli Eğitim Bakanlığı (2014) , Eğitim Bilişim Ağı, 9. Sınıf Ders içerikleri, Cümlenin Öğeleri, [Internet] www.eba.gov.tr/video/izle/02587b6392e7b8b634f78977bd638f5cc482581ed6300 [Erişim tarihi;11.02.2014].

[2] S.V. N. Vishwanathan, Nicol N. Schraudolph, Mark W. Schmidt, Kevin P. Murphy, "Accelerated Training of Conditional Random Fields with Stochastic Gradient Methods", In Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[3] Lafferty, J. D., McCallum, A., ve Pereira, F., "Conditional random fields: Probabilistic modeling for segmenting and labeling sequence data", In Proc. Intl. Conf. Machine Learning, vol. 18. 2001.

[4] Cihan Özköse, M.Fatih Amasyalı, "Cümle Öğelerinden Hayat Bilgisi Çıkarımı", Türkiye Bilgisayar Mühendisliği Dergisi, Sayı:06, Aralık 2012.

[5] Nilay Coşkun, "Türkçe Tümcelerin Öğelerinin Bulunması", Yüksek Lisans Tezi, İTÜ Fen Bilimleri Enstitüsü, 2013.

[6] Charles Sutton, Andrew McCallum, "An Introduction to Conditional Random Fields", Foundations and Trends in Machine Learning 4 (4). 2012.

[7] Gökhan Akın Şeker, Gülşen Eryiğit. "Initial explorations on using CRFs for Turkish Named Entity Recognition", In Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 2012.

[8] Ozkaya, S., Diri, B., "Named Entity Recognition by Conditional Random Fields from Turkish informal texts", Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference.

[9] Gülşen Eryiğit, "Dependency parsing of Turkish", 2006. Ph.D. Thesis, Istanbul Technical University, Istanbul.

[10] Kudo, T. (2009) CRF++: Yet Another CRF toolkit, [Internet] https://code.google.com/p/crfpp [Erişim tarihi;11.02.2014].

[11] Han-Shen Huang, Yu-Ming Chang, Chun-Nan Hsu, "Training Conditional Random Fields by Periodic Step Size Adaptation for Large-Scale Text Mining", ICDM, 511–516, 2007.

[12] Akin, A.A., Akin, M.D. (2007) Zemberek, an open source NLP framework for Turkic Languages, [Internet] http://zemberek.googlecode.com/files/zemberek_makale.pdf [Erişim tarihi;11.02.2014].

[13] Can, F., Koçberber, S., Bağlıoğlu, O., Kardaş, S., Öcalan, H.C., Uyar, E., "Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması", Akademik Bilişim Sempozyumu, 2009.

[14] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", Ph.D. thesis, University of Waikato, 1998.

Mechanical Engineering Article

/

Makine Mühendisliği Makalesi