Davranışsal Türkçe Metin Sınıflandırıcı Tasarımı ve Kodlanması

Şadi Evren, ŞEKER[1] Banu, DİRİ[2]
[1]*http://www.sadievrenseker.com*
[2]*http://www.ce.yildiz.edu.tr/~diri*
ses@sadievrenseker.com[1], banu@ce.yildiz.edu.tr[2]
Proje e-posta&WEB: tusse@shedai.net http://www.shedai.net/tusse

Özet

Günümüzde hızla büyümekte olan bilişim teknolojilerinin en önemli bilgi kaynaklarından biri olan İnternet'in, daha verimli kullanılabilmesi için, veri ayıklama (information extraction) ve veri sınıflama (data classification) konuları hızla gelişmektedir. Bu çalışmada, veri sınıflama konularından biri olan dokümanın yazım dilinin belirlenmesi ile ilgili bir uygulama geliştirilmiş ve bir arama örümceği içerisine gömülerek kullanılmıştır. Makalede de belirtildiği gibi dil sınıflandırması, özellikle İnternet gibi değişik geçmişlerden ve çevrelerden gelen tasarımcıların olduğu ortamlarda oldukça karmaşık bir durum alabilmektedir. İnternet örümceği yardımıyla, bu çalışmada yazılmış kodların kolay bir şekilde binlerce sayfa üzerinde test edilmesine olanak sağlanmış ve ilgili sayfanın dilinin belirlenmesinde başarılı sonuçlar elde edilmiştir. Ayrıca arama motorları ve veri ayıklama uygulamalarıyla da sorunsuz çalıştığı görülmüştür.

Teşekkür

Projede kullanılmış olan şekilbilim (morphology) parçası için Prof. A. C. Cem SAY'a teşekkürü bir borç biliriz.[1]

1. Giriş

Herhangi bir dokümanın dilinin Türkçe olup olmadığını algılamak bir problemdir. Bu çalışmada hedef alan olarak, İnternet üzerindeki dokümanlar seçilmiş ve İnternet'te verilen bir hedef etki alanının (domain) ne kadar başarılı dil sınıflandırması yapabileceği ölçülmüştür.

Çalışmalarımız sırasında, İnternet üzerinde dolaşabilen bir örümcek ve İnternet sayfalarında kullanılan dil olan HTML (HyperText Markup Language) dilinin parçalayıcısı (parser) alt yapı çalışması olarak bu proje içerisinde yazılmıştır. Ayrıca projenin şekilbilim (morphology) kısmında kullanılan kodlar daha önceden yazılmış olan TUSA[2] projesinden alınmıştır.

Yapılan literatür çalışmaları sonucunda görülmüştür ki dil sınıflandırması konusunda pek çok çalışma olmasına karşın, özellikle Türkçe için yapılan çalışmaların tamamı, doğal dil özelliklerinden uzak istatistiksel çalışmalardır. Literatür taramasında farklı diller içinde benzer çalışmalara rastlanmıştır.[3][4] Bu çalışma ile bilinen insan davranışına en yakın şekilde dil sınıflandırması yapılmaya çalışılmıştır.

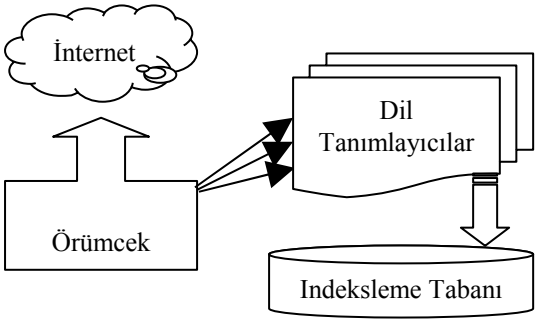
Makalenin ikinci bölümünde alt yapı çalışması olarak hazırlanan İnternet örümceğinin tasarımı, kodlanması ve projeye olan katkılarından bahsedilmiş, üçüncü bölümde insanın, dil sınıflandırmadaki davranışları analiz edilmiştir. Bu çalışmada, dili belirlerken karakter ve kelime bazında inceleme yapılmış, cümle seviyesine inilmemiştir. Dördüncü bölümde karakter seviyesindeki sınıflandırmadan bahsedilmiş, beşinci bölümde ise kelime seviyesindeki dil sınıflandırmada kullanılan iki farklı yöntem (şekilbilimsel ve istatistiksel) ayrı ayrı ele alınarak incelenmiş, tasarlanmış, kodlanmış ve sonuçlar karşılaştırılmıştır. Altıncı bölümde de bu üç yöntem hibrid bir uygulama ile birleştirilerek performans artışına gidilmiştir.

2. İnternet örümceğinin tasarımı ve kodlanması

Bu çalışmanın ana amacı dil sınıflandırma olmasına karşılık, İnternet üzerinde dolaşan ve sayfaları kategorize eden çok iplikli (multi threaded), dağıtık (distributed) bir kod yazıp, dil belirlemesi yapacak olan modüle HTML kodlarından temizlenmiş saf doküman verilebilmesidir.

Örümcek, basit bir şekilde hedef gösterilen siteden başlayarak bütün İnternet'i dolaşmakta ve karşılaştığı her sayfada, diğer sayfalara olan bağlantıları gidilecekler listesine ekleyerek ilerlemektedir. Sunmuş olduğu eşsiz İnternet ve string işleme olanakları ve ortam dinamikliliği yüzünden "şekilbilim" dışındaki bütün modüller JAVA ortamında geliştirilmiştir. Çalışma kapsamında donanım limitlerinden dolayı örümceğin gidebileceği sayfa sayısı 100 ile kısıtlanmıştır, ayrıca örümcek gidecek yeni bir bağlantı bulamadığı durumda sonlanmakta ve uğradığı her sayfayı, ilgili dil belirleme modüllerine yönlendirerek çalışmanın ilerideki adımlarına veri sağlamaktadır.

Örümcek internet üzerinde dolaşır ve farklı süreç (process) veya bilgisayarlar üzerinde beklemekte olan dil tanımlayıcılara internetten aldığı bilgileri göndererek dil tanımını ister. Dil tanımlayıcılar, dile karar verdikten sonra, bu bilgiyi beklemekte olan indeksleme sürecine veya bilgisayarına yönlendirir.



Şekil 1. İnternet, Dil Tanımlayıcısı, Örümcek, İndeksleme ilişkisi

Programlama ortamının JAVA olması sayesinde, yazılmış olan her modül (class) bir süreç veya ayrı bir server olarak çalışabilmektedir. Indekslenmesi istenen bilginin, artması durumunda Örümcek, Dil Tanımlayıcılar ve İndeksleme Tabanının ayrı bilgisayarlarda tutularak performansın arttırılması, ayrıca bütün modüllerin aynı bilgisayar üzerinde çalışması da mümkün olmaktadır.

3. Davranışsal dil sınıflandırması

İnternet ortamı için gerçekleştirilmiş bir uygulama olsa da, İnternet'in karakteristik yapısı bu çalışmanın ilerlemesinde katkı sağlayamamıştır. Örneğin, etki alanı adında (domain name) Türkiye bilgisini içeren (".tr" uzantılı) sitelerin Türkçe olması gibi bir zorunluluk yoktur ve sayfanın dilinin analiz edilmeden anlaşılması imkansızdır. Yazılan kodun geliştirilme aşamasında Turing testi [5] esas alınarak, insan davranışı[6] örnek alınmış ve bir kişinin, eline aldığı bir dokümanın diline nasıl karar verdiği taklit edilmiştir.

Temel olarak bir doğal dil işleme çalışmasının 4 ana aşaması bulunur. Bunlar şekilbilim (morphology), cümlebilim (syntax), anlambilim (semantics) ve sesbilim (phonetics)'tir. [7] Bu çalışmanın İnternet sayfaları üzerinde çalışmakta olduğu, dolayısıyla doğal dil işlemenin parçası olan sesbilim (phonetic) çalışmaları ile bir ilgisinin olmadığı unutulmamalıdır.

Dil sınıflaması yapan kişi, bir dokümanı eline aldığında önce dokümandaki yazıların karakterlerine ve alfabesine bakar, tanımadığı bir alfabe ise bir seviye üstte yer alan kelime seviyesine çıkar ve yine karar verilememesi durumunda daha üst seviyede ki cümle yapısına bakar.[8] Ancak tecrübeler ile belirlenmiştir ki, çoğu dilde dokümanın dili, kelime seviyesinde anlaşılabilmekte ancak çok benzer dillerde cümle seviyesine çıkılması gerekmektedir [9] (örneğin Azeri / Türkçe veya Urduca / Arapça gibi).

4. Karakter seviyesinde sınıflandırma

Bir dokümanın dilini belirlemede ki en basit ve hızlı yöntem dokümandaki karakterlere bakmaktır. Ancak bilgisayar ortamında, yazılı olan karakterler bulunmamakta bunun yerine harflerin karşılığı olan kodlar yer almaktadır. En çok kullanılan ASCII (American Standard Code for Information Interchange), ISO (International Organization for Standardization) ve UTF (Unicode Transformation Format) kodu olup, seçili olan dile göre farklı karakterleri göstermekte ve doküman bu tablolara göre şekillenmektedir. Örneğin çizelge 1' deki harflerin hepsi ASCII 254 kodlu harfin farklı dillerdeki karşılıklarıdır yani doküman içerinde 254 koduyla temsil edilmekte ancak farklı semboller ile kullanıcıya gösterilmektedir.

Çizelge 1. ASCII 254 kodunun farklı dillerdeki karakter kodlamaları

Þ	Latince	þ	Yunanca
ﺚ	Arapça	ş	Türkçe

Bu kodların basılma işlemi ise değişik seviyelerde yapılabilmektedir. En kesin çözümü HTML kodları ile şekil 2'deki bilgilerin yazılmasıdır.

```
<html>
<head>
<meta http-equiv="Content-Language"
content="tr">
<meta http-equiv="Content-Type"
content="text/html; charset=windows-
1254">
<meta http-equiv="Content-Type"
content="text/html; charset=ISO-8859-
9">
...
</head>
<body>
...
</body>
</html>
```

Şekil 2. HTML Dilinde karakter kodlamasının belirtilmesi

Bu sayede yazılan HTML dokümanının dilinin Türkçe olduğu meta tagleri ile anlaşılır. Şekil 2'de bulunan üç meta kodu da aynı amaç için farklı formatlarda yazılmış olup, sadece birisinin yazılması yeterlidir. Ancak uyum problemleri yüzünden aynı anda ikisinin veya hepsinin yazıldığı örnekler de mevcuttur.

Dil belirlemesi için geliştirdiğimiz uygulamanın öncelikli olarak kontrol ettiği bilgi bu olup, gerekli

olan meta bilgisinin mevcut olduğu durumlarda ilgili sayfayı Türkçe olarak kabul eder.

5. Kelime seviyesinde sınıflandırma

HTML kodu içerisinde, ilgili meta taglerinin bulunmaması durumunda sayfanın karakter kodlaması, işletim sistemi veya İnternet tarayıcısı (browser) üzerinde yapılan ayarlar ile gerçekleştirilir. Örneğin, Türkçe işletim sistemi kullanan veya Türkçe ayarları yapılmış bir tarayıcı ile İnternet sayfası tasarlayan bir tasarımcının, sayfayı Türkçe olarak görebilmesi için bu kodları yazması zorunlu değildir. Çünkü bu kodlama otomatik olarak yapılmaktadır.

Yapmış olduğumuz testler göstermiştir ki sayfaların yaklaşık %70'inde meta tag kodlaması bulunmamaktadır. Bu gerçek de bizi dil belirleme işlemini kelime seviyesinde yapmaya zorlamıştır. Bu aşamada Şekilbilimsel (morphological) ve Kelime İstatistiksel Analiz (word-gram) uygulanmıştır.

5.1. Şekilbilimsel kelime analizi

Şekilbilimsel Kelime Analizi, insan davranışına en yakın dil belirleme yöntemidir. Buna göre okunan bir kelimenin Türkçe olup olmadığı, kelimenin kökünün ve eklerinin tanınıp tanınmamasına bağlıdır. Gerçekte de bir insan, dokümandaki harfleri algıladıktan sonra, kelimeleri daha önceden bildiği kelimeler ile karşılaştırmaktadır. Ancak insan beyni analitik düşünmeye sahip olduğundan, daha önceden görmemiş olduğu bir kelimenin kökünü biliyor ve eklerini de tanıyorsa dokümanın dilini Türkçe olarak algılayabilir.

Bu çalışmada geliştirdiğimiz uygulama yukarıda anlatılan yaklaşımın bire bir aynısıdır. Buna göre bir dokümandaki kelimeler rast gele seçilerek[1] şekilbilimsel analizden geçmekte ve Türkçe gramer kurallarına göre kök ve eklerine ayrılıp ayrılamayacağı kontrol edilmektedir. Şayet başarılı olunursa, bilinen bir Türkçe kök olup olmadığı sorgulanmakta ve kök sözlüğümüzde bulunan 28000 kök ile karşılaştırılmaktadır.[10][11]

Bütün kelimelerin analizi yapıldıktan sonra ve Türkçe kelime bulunma oranı %50'nin üzerinde ise sayfa Türkçe olarak, aksi taktirde bilinmeyen bir dilde yazılmış sayfa olarak sınıflandırılmaktadır.

Bu çalışmada, %50 gibi bir eşik değer (treshold value) konulmasının sebebi ne kadar mükemmel şekilbilimsel analiz yapılırsa yapılsın, diller arasında ortak olan ve dolayısıyla bizim kök sözlüğümüzde bulunmayan kelimelerin mevcut olmasıdır. Örneğin, özel isimler, yer isimleri veya rakam ile yazılan sayılar her dilde ortaktır ve bir sözlük içerisinde yer alması imkansız[2] kelimelerdir. Kısaca bir doküman mükemmel bir Türkçe ile yazılmış olsa da, şekilbilimsel analize dahil edilmeyecek kelimeler bulunacaktır. Bu sebepten dolayı bir eşik değeri belirlenmiştir.

İnternet üzerinde aynı sayfada birden fazla dilin kullanılmış olması da mümkündür. Çalışma kapsamında bir sayfadaki kelimelerin en az %50'sinin Türkçe olması durumunda, ilgili sayfayı Türkçe olarak kabul etmekteyiz.

5.2 İstatistiksel kelime analizi

Kelime analizi seviyesinde ikinci olarak kullandığımız yaklaşım, kelime istatistiklerine göre karar verilmesidir. Sözgelimi bir doküman sınıflandırılması yapılırken, aynı doküman içerisinde birden fazla dil kullanılmış olabilir veya bir dokümanın içinde alıntı cümleler yer alabilir, dolayısıyla bu az sayıdaki yabancı kelimenin, sonucu etkilememesi ve hatta ileride farklı dilleri de belirleme amacıyla kullanılabilmesi için, internet üzerinde bu bilgilerin istatistiğini toplayan ve elde ettiği istatistiklere göre karar veren bir modül yazılarak çalışmanın kapsamına dahil edilmiştir. Buna göre ilgili modül, internet üzerinde bir dokümanın dilini belirlerken, bu yöntemlerle kendisini eğitmekte, kelime istatistiklerini güncellemekte ve daha sonra karar verilmesi istenilen sayfada bu istatistikleri kullanmaktadır.

Bu çalışma da, sistemi eğitmek için basın siteleri (com.tr ile biten gazete ve televizyondan oluşan sınırlı bir alan), eğitim siteleri (edu.tr) ve devlet sitelerinden (gov.tr ve mil.tr) oluşturulmuş 40 farklı site [3] kullanılmıştır. Analiz sırasında bazı problemler ile karşılaşabiliriz. Bir sayfanın iki dilden hangisine ait olduğunun sorulması durumunda, sınırlı bir etki alanında çalıştığı için analiz başarılı olabilir. Burada karşılaşılan en büyük problem, bilinmeyen diller arasından tek bir bilinen dilin ayrıştırılması isteniyorsa (bizim örneğimizde olduğu gibi internetteki bütün sitelerden sadece türkçe sitelerin ayrıştırılması isteniyor ve diğer diller ile ilgili bir çalışma yapılmıyorsa), bu durumda sağlama yapılacak bir dil olmaması (yani sayfanın diğer dillerdeki başarısının ölçülüp bunların kararı etkilememesi) başarıyı etkilemektedir.

6. Hibrid yaklaşım ve test sonuçları

[1] Proje kapsamında tamamen performans endişesi ile bir dokümandan azami 1000 kelime rast gele seçilmiş, bu sayede çok büyük dosyalarda bütün kelimelerin tek tek analiz edilmesi önlenmiştir. Bu kısıt kullanılan donanım ve internet bağlantı hızına göre kodlanmıştır.

[2] Örneğin, "Ali" kelimesi hem İngilizce sayfalarda hem de Türkçe sayfalarda yer alabilir ve bu kelime sayfanın dilinin belirlenmesinde etkili değildir ve kök sözlüğümüzde yer almayan bir kelimedir.

[3] Bu sitelerin tam listesi için www.shedai.net/tusse/html/testpad.html adresinden yararlanabilirsiniz. Her sitenin girişinden itibaren 1000 sayfa taranmaktadır.

Sınıflandırmadaki başarıyı arttırmak için önerilmiş olan yöntemleri birleştirip kullanmak önemlidir. Bu çalışmada, öncelikle karakter ayarlarına ve kelime istatistiklerine, daha sonra da kelimelerin Türkçe ek ve kök yapılarına uygun olup olmadığına bakılmakta ve yöntemlerin önce ayrı ayrı daha sonra da hibrid olarak karşılaştırmaları yapılmıştır.

Test için 40 site seçilmiş olup, bunların 30 tanesi özel olarak seçilirken, diğer 10 tanesi de rast gele seçilmiştir. Test edilen siteleri 4 ana grupta toplamak mümkündür. [3]

- Basın siteleri (gazeteler ve televizyon kanallarının siteleri)
- Devlet siteleri (başbakanlık, TBMM, TSK, konsolosluk gibi)
- Üniversite siteleri (Türkiyenin değişik vakıf ve devlet üniversiteleri)
- Rast gele seçilmiş siteler (Internet üzerinde rastlanabilecek belirli bir nizamnameye tabi olmayan, amatör ve profesyonel siteler)

Bu testlerde kullanılan siteler Aralık 2005 ayı içinde indekslenmiş olup bu tarihten önceki ve sonraki değişiklikler testte yer almamaktadır.

6.1 Karakter Seviyesi Sınıflandırma

Test için seçilmiş 40 siteden, 30'unun karakter kodlaması için bir uyum içerisinde olmadığı görülmüştür. Aşağıdaki örneklerde her sitede karşılaşılan farklı "<META http-equiv="Content-Type" tag bilgisi site isminin altında belirtilmiştir.

- **http://www.milliyet.com.tr/**
content="text/html; charset=iso-8859-9"
- **http://www.tercuman.com.tr/**
content="text/html; charset=windows-1254"
- **http://www.yeditepe.edu.tr/**
content="text/html; charset=windows-1252"
- **http://www.basbakanlik.gov.tr/**
content="text/html; charset=windows-1254"/
- **http://www.iem.gov.tr/**
content="text/html; charset=iso-8859-1" /

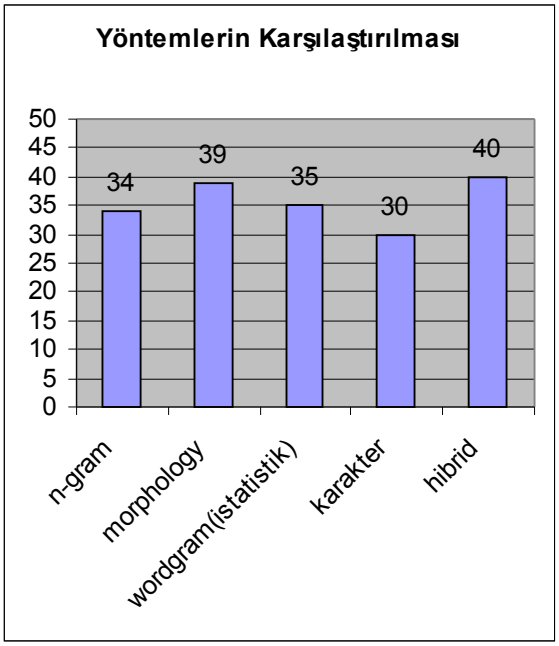
Buna göre yukarıdaki 5 sitenin hepsinde birbirinden farklı kodlamalar bulunmaktadır. Örneğin, milliyet sitesinde sadece ISO standardı kullanılırken, tercuman'nın sitesinde windows standardı, Yeditepe üniversitesinde ise farklı bir windows karakter seti kullanılmıştır. Başbakanlık ise windows standardını kullanmış ancak sonuna "/" işaretini koyarak sitenin kodunu XML uyumlu hale getirmiştir. İstanbul Emniyet Müdürlüğü'de ISO standardının farklı bir tablosunu, muhtemelen giriş sayfasında türkçe karakter olmamasından dolayı, giriş sayfasında kullanmıştır.

Kısaca, hem ISO hem de windows standartlarını aynı anda sitelerinde bulunduranlar olduğu gibi, Maltepe üniversitesinin sitesinde sadece windows standardı kullanılmış ancak aynı standart gereksiz yere 4 kere arka arkaya yazılmıştır. Karşılaşılan bu durum 4. grup olan rastgele sitelerde de sergilemektedir. Amatörce hazırlanılan sitelerin pek çoğunda da gerek dikkatsizlikten gerekse bilgisizlikten karakter kodu bulunmamakta bu da bizi kelime seviyesi çalışmalarına yönlendirmiştir.

6.2 Kelime Seviyesi Dil Analizi

Kelime analizi seviyesinde, istatistiksel ve şekilbilimsel olmak üzere iki farklı yöntem kullanılmıştır. Ayrıca çalışmamızda yer almayan n-gram[4] yöntemi de dil sınıflandırmasında kullanılmaktadır. Bu bölümde ismi geçen üç yöntemin başarıları ölçülmüş ve karşılaştırılmıştır.

Şekil 3'te, kullanılan yöntemlerin test alanı olarak seçilen 50 siteden, ki 10 tanesi İngilizcedir, kaçını Türkçe olarak sınıflandırdığı gösterilmiştir. Kelime analizine dayalı yöntemin başarılı olmasının en büyük sebebi, diğer hiçbir yöntemde bütün dili taramanın mümkün olmayışıdır.

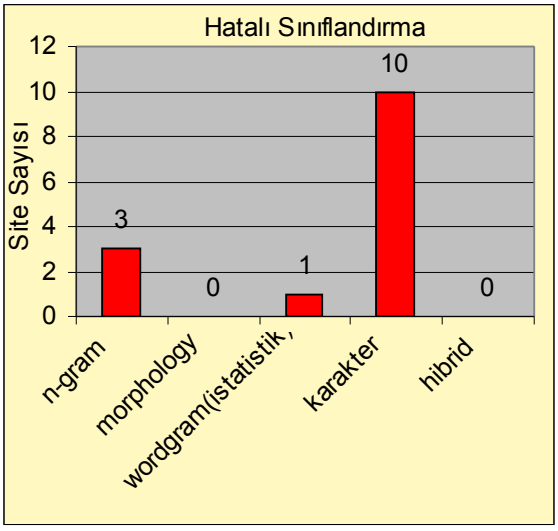


Şekil 3. Kullanılan 4 farklı yöntemin başarı grafiği

Örneğin word-gram kelime istatistik yönteminde, Türkçe'deki bütün kelimelerin eklenmesinden sonra şekilbilimsel analizdeki başarının yakalanması mümkün olabilmektedir. Ne yazık ki bu yaklaşım gerek kapladığı yer, gerekse her kelimede kaybedilecek zamanın getirdiği yük yüzünden başarılı görülmemektedir. Benzer şekilde n-gram ve karakter analizi yöntemlerinde de veri kaybı bulunmakta yani bir metnin bütün kelimeleri taranmamakta, dolayısıyla da %100 başarı yakalanamamaktadır. Uygulanan bu yeni yaklaşımla %95 civarı başarı elde

[4] N-gram sınıflandırma yöntemi, bir metindeki n sayıdaki harften oluşan karakter gruplarının frekans analizinin yapılarak dilin sınıflandırılmasına dayanır. Örneğin İngilizce dilinde "th" 2-gramı Türkçe'nin 23 misli daha fazla geçmektedir. Dolayısıyla iki dil arasında sınıflandırma yapan bir 2-gram makinesi, "th" 2-gramını fazla gördüğü zaman metni İngilizce olarak sınıflandıracaktır.

edilmektedir. Başarısız olunan tek örnek özel isim ve sayı gibi dil bağımsız kelimelerin çoğunlukta olduğu örneklerdir. Burada da başarı mümkün olmasına karşın, konulan eşik değer uygulaması başarıyı olumsuz etkilemektedir.



Şekil 4. Yöntemlerin sınıflandırmada yaptıkları hata sayısı

Hibrid olarak adlandırdığımız, karakter analizi ardından şekilbilimsel analiz ve istatistiksel analizlerin aynı anda uygulandığı analiz yöntemi %100 başarı vermiştir.

Şekil 4'te görüldüğü üzere en büyük hata payı karakter analizine aittir. Bunun sebepleri bir önceki bölümde anlatılmıştır. Ancak dokümanın uzunluğunun azaldığı örneklerde, n-gram yönteminin de başarısız olduğu anlaşılmıştır. Bunun sebebi n-gram yönteminin ancak belirli bir sayıdaki karakter grubunu karşılaştırdıktan sonra frekansı sabit hale getirmesidir. Ancak şekilbilimsel yöntem ile sınıflandırma daha başarılı olmaktadır, çünkü dili belirleyen o dildeki harflerin yanyana gelme oranı değil, harfler, kelimeler ve nihayet cümlelerdir. Benzer şekilde insan da, bir metni tasnif ederken bu unsurları göz önüne almakta, bir dokümanda geçen kelimeleri saymamakta veya metindeki harflerin ne sıklıkta yanyana geldiğine dikkat etmemektedir.

7. Sonuç

Yapay zeka çalışmaları günümüzde hızla yeni alanlarda kullanılmaktadır. İnsanlığın hayatını kolaylaştırmasının yanında, insanın kendisini anlaması ve kendi öz varlığında bulunan özellikleri yaşama kazandırmasında önemli bir rol üstlenen yapay zeka çalışmalarına bir örnek olan çalışmamızda Turing testini geçmek gibi iddialı bir hedefimiz olmasa da, daha önceki yöntemlerden daha başarılı bir sonuç elde edilmesi yine insanın düşünce yapısının bir probleme uyarlanmasıyla elde edilmiştir. Dil sınıflandırılması her ne kadar insan kaynaklı bir problem de olsa, şimdiye kadar uygulanan yöntemlerin bu konuda başarısız olabilmesinin temel nedeni insan kaynaklı analitik yöntemler olmayıp istatistik gibi doğa kaynaklı yöntemler olmasıdır.

Bu çalışmada Türkçe olarak hazırlanmış olan internet sayfaları, diğer dillerde yazılmış olan sayfalardan kolayca ayırt edilebilmektedir. Bunun için üç farklı yöntem birleştirilerek kullanılmış ve %100 başarı elde edilmiştir.

Çalışmada elde edilmiş olan bu başarı sayesinde, ileride yapılacak olan İnternet üzerindeki doğal dil işleme projeleri için sadece Türkçe yazılmış olan sitelerin veri kaynağı olarak kullanılması mümkün olacaktır. Buna göre İnternet üzerinde Türkçe kaynaklara dayanılarak anlambilimsel (sematic) çalışmalar yapılabilir.

Ayrıca insanların sadece kendi dillerindeki sonuçları görmesi sağlanarak arama motorlarındaki başarı oranı arttırılabilecek ve bu sayede topluma da bir fayda sağlanmış olacaktır.

8. Referanslar

[1]Özlem Çetinoğlu, "A Prolog Based Natural Language Processing Infrastructure for Turkish", Yüksek Lisans Tezi, Boğaziçi Üniversitesi, 2003

[2]Şadi Evren ŞEKER, Birol Aygün, "Design and Implementation of a Personal Calendar with a Natural Language Interface in Turkish", http://www.shedai.net/tusa, Yüksek Lisans Tezi, Yeditepe Üniversitesi, 2003

[3]Penelope Sibun & A. Lawrence Spitz, "Language Determination: Natural Language Processing from Scanned Document Images", Fuji Xerox Palo Alto Laboratory, 1994

[4] Izumı Suzukı, Yoshıkı Mıkamı, Arıo Ohsato, "A Language and Character Set Determination Method Based on N-gram Statistics", Nagaoka University of Technology, 2002

[5]A M. Turing, On Computable Numbers, with an Application to the Entscheidungsproblem, *Proc. London Math. Soc.* , 1937

[6]Stephen J Rymill, Neil A. Dodgson, "Psychologically-based vision and attention for the simulation of human behaviour", Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia, Sayfalar 229-236, 2005

[7]Covington M. A. Natural Language Processing for Prolog Programmers, Prentice Hall, New Jersey, 1994

[8]Sadi Evren SEKER, A. C. Cem Say, Birol Aygun, "Türkçe Dogal Dil Arayüzülü bir Kisisel Takvim Programinin, Tasarim ve Kodlamasi",TAINN,2003, sayfa 90

[9]Şadi Evren ŞEKER, Ender Özcan, Z. İlknur Karadeniz, "Design and Implementation of a Turkish Speaking JAVA Code Generator and Code Teller", http://www.shedai.net/tuja, ICEIS, Proto, Protogual, NLUCS Workshop April 2004

[10] Şeniz Demir, "Improved Treatment of Word Meaning in a Turkish Conversational Agent", Master Tezi, Boğaziçi Üniversitesi, 2003

[11]Oflazer K. (1993) Two-Level Description of Turkish Morphology, Proc. Second Turkish Symposium on Artificial Intelligence and Neural Networks, BU Press, pp. 86-93.