Gazi Üniv. Müh. Mim. Fak. Der.
Cilt 25, No 3, 483-494, 2010

J. Fac. Eng. Arch. Gazi Univ.
Vol 25, No 3, 483-494, 2010

WEB TABANLI OTOMATİK DİL TANIMA VE ÇEVİRME SİSTEMİ

Uraz YAVANOĞLU ve Şeref SAĞIROĞLU
Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Maltepe, Ankara.
uraz@gazi.edu.tr, ss@gazi.edu.tr

(Geliş/Received: 09.07.2009 ; Kabul/Accepted: 19.12.2009)

ÖZET

Bu çalışmada, internet ortamında bulunan MS Word ve HTML sayısal belgelerinin dilini tanımak ve sunulan bilgilerin içeriğini farklı dillere çevirmek için bir sistem geliştirilmiştir. Dil tanıma sorunu aslında daha genel bir problem olan özniteliklerin sınıflandırılması olarak gözükmektedir. Geliştirilen sistem farklı işlemlerin yapılmasını kolaylaştırmak için kullanıcıların içerik dillerini hiç bilmedikleri internetteki sayısal dokümanların içeriklerinin yapay sinir ağları temelli zeki bir çözüm ile otomatik tespiti ve istenilen 40 dile otomatik olarak çevirebilecekleri sistemden oluşmaktadır. Yapılan testlerde 15 dil için dokümanlar kullanılmıştır. Bu testlerde sistemin gerçek zamanlı olarak çalışma başarısının beklenenin üzerinde olduğu gösterilmiştir. Bu çalışmanın, internet içeriklerinin daha etkin olarak kullanılmasını sağlaması beklenmektedir.

Anahtar Kelimeler: Dil tanıma, dil dönüştürme, web platformu, yapay sinir ağı, web tabanlı yazılım

AUTOMATIC WEB BASED LANGUAGE IDENTIFICATION AND TRANSLATION SYSTEM

ABSTRACT

This study presents new methods to identify web contents, containing MS Word and HTML documents in different languages and to translate them into specified languages. The identification problem can be seen as a specific instance of the more general problem of an item classification though its attributes. This novel approach is based on artificial neural network model to recognize the languages. Documents belonging to 15 languages were used in test. The results have shown that the approach presented in this work is very successful to meet the expectations in real-time language identification accuracy and translate into 40 different languages with the help of a developed platform. It is expected that this study will help users to use internet more effectively.

Key Words: Language identification, language translation, web platform, artifical neural network

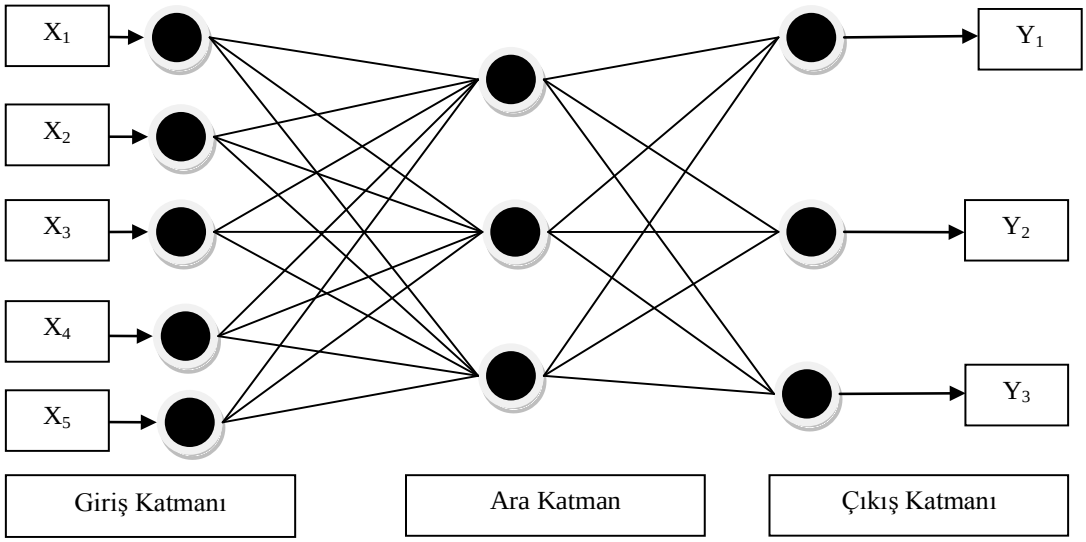
1. GİRİŞ (INTRODUCTION)

Dil tanıma üzerine literatürde pek çok yöntem geliştirilmiştir [1-15]. Padro ve Padro tarafından 2004 yılında yapılan çalışmada, istatistiksel yöntemlere dayalı üç farklı dil tanıma yöntemi incelenmiştir. Önemli sayılabilecek parametrelerden birisinin metin uzunluğu olduğu yine bu çalışmanın bulguları arasındadır. Başarı oranı, metin boyutu kısaldıkça doğru orantılı olarak azalmaktadır. Kısa metinlerde kullanılan yöntemlerde başarı oranı %60 seviyesine kadar gerileme göstermiştir [1]. Botha ve arkadaşlarının 2007 yılında yaptığı çalışmada, metin tabanlı 11 farklı Güney Afrika dili ele alınarak dil tanıma süreci araştırılmış ve n-gram istatistikleri sınıflandırma için kullanılmıştır [2]. Bu süreç yönetiminde, özellikle destek vektör makineleri (support vector machines) ve yakın komşuluk tabanlı sınıflandırıcılar (likelihood-based classifiers) karşılaştırılmıştır. Tanıma işlemi için giriş olarak uygulanan metinlerden az sayıda kelime ile tanıma işleminin mümkün olduğu belirlenmiştir. Destek vektör makineleri genel olarak daha başarılı sonuçlar vermesine rağmen sınıflandırma için gerekli olan yüksek işlem karmaşıklığı sebebiyle ancak basit ve kararlı bir sistem tasarımı için uygun olmadığı gösterilmiştir. Yapılan testlerde, daha büyük giriş setleri ile doğrusal olarak daha kesin sonuçlar elde edilmekle birlikte destek vektör makinelerinin karmaşıklık probleminin sistemi karasızlığa götürdüğü rapor edilmiştir [2]. El-Shishiny ve arkadaşları tarafından 2004 yılında yapılan diğer bir çalışmada, Arapça kökenli diller ve diğer diller arasında Arapça betiklerin kelime parçalama tabanlı bir metot ile nasıl ayrıştırabileceği araştırılmıştır [3].

Bu dillerden Arapça betiklerin tanınması ile %94 oranında dil tanıma başarısı elde edilmiştir.

Kruengkrai ve arkadaşları tarafından 2005 yılında yapılan başka bir çalışmada, otomatik dil tanıma için karakter dizisi çekirdeği (string kernels) kavramı tabanlı yeni bir yöntem önerilmiştir. Sonuçlar yeterli sayıda giriş setiyle 17 farklı dil için değerlendirilmiş ve elde edilen test sonuçlarında diller arası yakınsamanın azaltılması için eşleşen sözcük öbeklerine göre ağırlık atama yönteminin kullanılmasının gerektiği sonucuna varılmıştır [4]. Zavarsky ve arkadaşları tarafından 2005 yılında yapılan diğer bir çalışmada, kodlama örüntüleri karşılaştırılarak çok yüksek sayıda giriş setlerine ait dil tanıma işlemlerinin sonuçları sunulmuştur [5]. Çalışmada kullanılan büyük giriş kümeleri (milyonlarca) Dil İzleme Projesinin bir ürünü olarak ortaya çıkmaktadır. Bu proje milyonlarca metin tabanlı dokümandan oluşmaktadır. Bu dokümanları kullanarak, kullanıcılar n-gram ve kelime tabanlı araçları tanıtmak istedikleri dil için eğitirler. Kullanıcı her dil için gerekli olan materyali toparlayarak eğitim sürecini tamamlar. Bu yüzden sisteme giriş olarak uygulanan dokümanların çeşitli ölçütleri sağlaması gerekmektedir [5]. Peng ve arkadaşları tarafından 2003 yılında yapılan çalışmada dil bağımsız ve işlem bağımsız metin sınıflandırması ile karakter tabanlı n-gram dil modelleri kullanan bir metot önerilmiştir [6]. Bu yaklaşım en genel haliyle basit teorik ilkeler kullanan ve birçok dil sınıfında başarıyla çalışan bir yapıdır. Yapılan testlerde müzik veya DNA gibi ardışık veriler ile çoklu kategori sınıflandırması problemi (Reuters–21578 veri kümesi) gibi sorunların çözümünde önerilen yaklaşım başarılı bulunmuştur [6]. Nair ve arkadaşları tarafından 2007 yılında yapılan çalışmada, yazı dili İngilizce olmayan ülkelerin kendilerine özgü dil gruplarına karışan Latin kökenli metinlerin ayrıştırılarak gizli Markov Modelleri ile bu metinlere ait dillerin analizleri yapılmıştır [7]. Bu sayede değişik dil modelleri, ele alınarak oluşturulan metin editörleri ile dile bağlı renklendirme benzeri bir ayrıştırma aracı geliştirilmiştir. Geliştirilen yaklaşımın sözlük tabanlı metotlarla karşılaştırıldığında, zaman ve uzay karmaşıklığı bakımından üstün olduğu görülmüştür. Ahmed ve arkadaşları tarafından 2004 yılında yapılan çalışmada, etkili dil sınıflandırma için n-gram tasarsız birikimli frekans ekleme metodu kullanılmıştır [8]. Constable tarafından mevcut problemlerin çözülebilmesi için, dil tanıma sistemlerinde olması gereken bazı özelliklerin yazı sistemleri, varlık bilim, alan tanımlı veri kümesi, lehçeler ve diğer alt dil varyasyonları, üst dil bilim kategorileri, tarihsel dil değişimleri, dil bağımlı kategorileştirme ve yerelleştirme ile tüm bu sınıfların tartışıldığı konu başlıklarıyla ilgili çalışmalar yapılmıştır [9]. Adams ve arkadaşları tarafından 1997 yılında yapılan çalışmada Java programlama dilinde tasarlanmış deneysel bir sistem tanıtılmıştır. Bu çalışmanın oluşturulmasında dil tanıma sistemlerinde sıklıkla kullanılan n-gram analizleri sunulmuştur. Bu sayede konuşma etiketleme, tümce tanımlama, yabancı kelime çevirisi ve konu etiketleme gibi zorlukların aşılmasıyla birlikte web tabanlı zeki arama ve gösterimleri gibi konuların işlenebileceği önerilmektedir [10]. Ölvecky tarafından 2005 yılında yapılan çalışmada internetin büyümesiyle birlikte dil tanıma ve metin sınıflandırmanın öneminin artmasıyla birlikte bu alanda kullanılan n-gram varyasyonları incelenmiştir [11]. Bu alanda tartışılan bir başka konu ise n-gram giriş profili ile sistem eğitim büyüklüğünün ve yazım kalitesinin nasıl iyileştirilebileceğidir. Bilcu ve Astola tarafından 2006 yılında yapılan çalışmada yazınsal metinlerden dil tanıması yapabilecek yapay sinir ağı tabanlı hibrit bir metot önerilmiştir. Bu uygulama özellikle İngilizce ve Fransızcanın tanınması amacıyla geliştirilmiştir. Bu çalışma sayesinde iki dile ait metinlerden ses birimi hecelerine geçişin sağlanabileceği ve sistemin daha kararlı bir şekil alacağı savunulmuştur [12]. Liu ve arkadaşları ise karakterlerin resim verilerinden sınıflandırılması için yeni bir yöntem önerilmişlerdir [13]. Yapılan çalışmada Çince, İngilizce ve Japoncanın (Hiragana ve Katakana dâhil) dillerinin tanınması hedeflenmiştir. Zhu ve arkadaşları el yazısı ve bilgisayar yazıcı çıktılarının tanınmasında kullanılması için yeni bir yöntem önermişlerdir [14]. Yapılan testler sistemin yüksek karmaşıklıkta dokümanlar için iyi bir tanımlayıcı olduğunu göstermesine rağmen, kod çizelgelerinin oluşturulmasında karşılaşılan uzaysal karmaşıklık gibi sorunlar nedeniyle sistemin uygulama aşamasında bazı sorunların aşılması gerekmektedir [14]. Baykan ve arkadaşları, internet ortamında girilen tek bir adresten dil tanımanın mümkün olduğuna ilişkin öneriler sunmuşlardır [15]. Bu çalışmada kullanılan performans kriteri F-Ölçümü'dür. Bu ölçüm makine dil tanıma başarı kriteri olarak anılmaktadır. Bu testlerde, kişilerden ODP veri kümesinde kategorize edilmemiş 100 sayfayı tanımlamaları istenmiş ve ancak %50 başarı elde edilmiştir, sistem testlerinde ise bu başarı oranı %85 olarak gözlenmiştir [15]. Takçı ve Soğukpınar, harf tabanlı istatistiksel bir metot önermişlerdir [23]. Bu yöntem ile n-gram ve ortak sözcük metodu gibi yöntemlerde gerekli olan sözcükler yerine dil tanıması için kullanılacak metin içinde geçen harflere ait bir sıklık durum analizi yapmışlardır. Bu sayede inceledikleri İngilizce, Fransızca, Almanca ve Türkçe dilleri için alfabe bazında harflerin bulunma dağılımlarını çıkartarak dil tanıma için verilen bir metin içerisindeki harfleri ait olduklara dillere göre puanlayabilecek bir sistem geliştirmişlerdir.

Literatürde yapılan çalışmalar bu alanda yapılan çalışmalara büyük yön vermiş, dil tanımada farklı ve daha yüksek performanslar sunmuşsa da tüm dilleri tanımlayan yüksek başarımlı dil tanıma henüz gerçekleştirilememiştir. Bu makalede ise web



Şekil 1. Bir MLP yapısı; 5 giriş, 1 ara katman, 3 çıkış **(**As an example of Multi Layer Perceptron)

ortamındaki farklı dillerdeki içeriklerin otomatik olarak tanımlanması ve bu içeriklerin otomatik olarak istenilen dile dönüştürülmesi hedeflenmiştir.

2. YAPAY SİNİR AĞLARI (ARTIFICIAL NEURAL NETWORKS)

YSA'lar, yapılarına göre ileri beslemeli (feed forward) ve geri beslemeli (feed back) olarak ikiye ayrılırlar. İleri beslemeli ağlarda işaretler, giriş katmanından çıkış katmanına doğru tek yönde iletilirler. Geri beslemeli ağlarda, çıkış ve ara katman çıkışları kendinden önceki katmanlara ya da girişe geri beslenir. İleri beslemeli ağlara MLP (Multi Layer Perceptron), RBFN (Radial Basis Function Network) ve LVQ (Learning Vector Quantization) örnek verilebilirken, ART (Adaptive Resonance Theory), SOM (Self Organizing Maps) ve Elman ve Jordan ağları geri beslemeli ağlara örnek olarak verilebilir [16]. Bu YSA yapıları, Şekil 1'de verildiği gibi ileri beslemeli standart birleşmeli yapıda olabileceği gibi geri beslemeli dinamik yapıda da olabilirler.

Bu çalışmada, ileri beslemeli ağ yapılarından olan MLP yapısı kullanılmıştır. MLP yapısının tercih edilmesinin nedeni, bilinen en eski YSA modellerinden olması ve sınıflandırma problemlerinde başarılı sonuçlar üretmesidir. Bunların yanında, farklı öğrenme algoritmaları ile kullanıma uygun olması diğer bir tercih sebebidir. Şekil 1'de genel olarak gösterilen MLP modeli, bir giriş, bir veya daha fazla ara katman ve bir de çıkış katmanından oluşur. Bir katmandaki bütün nöronlar bir sonraki katmandaki bütün nöronlara bağlıdır. Giriş katmanındaki nöronlar tampon gibi davranırlar ve giriş sinyalini ara katmandaki nöronlara dağıtırlar. Ara katmandaki her bir nöronun çıkışı, kendine gelen bütün giriş sinyallerini takip eden bağlantı ağırlıkları ile çarpımlarının toplanması ile elde edilir. Elde edilen bu toplam, çıkışın toplam bir fonksiyonu olarak hesaplanabilir. MLP'de giriş katmanı hariç bir nöronun çıkışı

$$y_k = f(\sum_k w_k x) \quad (1)$$

ile hesaplanır. Burada, w nöronlar arasındaki ağırlık değerini, f ise kullanılan transfer fonksiyonunu gösterir.

YSA yapılarını eğitmede pek çok öğrenme algoritması kullanılmaktadır [16]. Bu çalışmada öğrenme algoritması olarak Levenberg-Marquardt, Geri Yayılım ve Momentumlu Geri Yayılım algoritmaları kullanılmıştır. Yapılan testler sonucunda sadece LM algoritması ile eğitilmiş YSA'larla işlem yapıldığı için sadece LM algoritması takip eden satırlarda kısaca özetlenmiştir.

Levenberg-Marquardt (LM) Algoritması

YSA'ları eğitmede başarı ile kullanılan algoritmaların başında gelmektedir [16,17]. LM algoritması, maksimum komşuluk fikri üzerine kurulmuş bir en az kareler hesaplama metodudur [16]. Bu algoritma, Gauss-Newton ve En Dik İniş (Steepest Descent) algoritmalarının en iyi özelliklerinden oluşur ve bu iki metodun kısıtlamalarını ortadan kaldırır. Genel olarak bu metot, yavaş yakınsama probleminden etkilenmez.

Hedef çıkışı hesaplamak için bir YSA'nın ağırlıklarının LM öğrenme algoritması kullanılarak eğitilmesi ağırlık dizisi w_0'a bir başlangıç değerinin atanması ile başlar ve hataların karelerinin toplamı E(w)'nin hesaplanmasıyla devam eder. Her E(w) terimi, hedef çıkış (y) ile gerçek çıkış (yd) arasındaki farkın karesini ifade eder. Bütün veri seti için E(w) hata terimlerinin tamamının elde edilmesiyle, ağırlık dizileri bir hesaplama akışı içerisinde, daha LM öğrenme algoritması adımlarının uygulanmasıyla adapte edilir.

LM öğrenme algoritmalarında hedef, parametre vektörü w'nin, E(w)'yi minimum yapacak veya arzu edilen çıkış (y) ile gerçek çıkış (yd) arasındaki farkın minimize olacak şekilde bulunmasıdır. LM algoritmasının kullanılmasıyla yeni vektör (w_{k+1}), farz edilen vektörden (w_k) Eş. 2 yardımıyla Eş. 3'deki gibi hesaplanabilir. Bu algoritmanın detayına [16] nolu kaynaktan erişilebilir.

$$E(w)=\sum_{i=1}^{m}(y_i - yd_i)^2 \quad (2)$$

$$w_{k+1} = w_k + \delta w_k \quad (3)$$

burada δw_k ifadesi

$$(J_k^T J_k + \boldsymbol{\lambda} I)\delta w_k = -J_k^T f(w_k) \quad (4)$$

Eşitliğinden faydalanılarak hesaplanır. Eş. 4'de;

$\boldsymbol{J_k}$: f'nin w_k değerlendirilmiş Jakobiyeni,
$\boldsymbol{\lambda}$: Marquardt parametresi ve
I: birim veya tanımlama matrisidir.

3. YSA ile doküman sınıflandırma
(Document Classification with ANN)

Bölüm 2'de açıklanan sebeplerden dolayı bu çalışmada doküman sınıflandırma için YSA'lardan faydalanılmıştır. Eğitim veri kümeleri oluşturulurken Latin alfabesi grubuna dâhil 5 farklı dil için giriş amacıyla kullanılacak sayısal (WORD) dokümanlar kullanılmıştır. Geliştirilen bu yaklaşımda test verilerinin üretilmesi için Erkek, Kadın, İnsan, Dünya, Ülke, Bilim, Kültür, Sanat, Politika, Hayat kelimeleri bir dil içerisinde en sık kullanılan kelimeler olarak önerilmiştir. Veri kümeleri oluşturulurken, belirlenen kelimelerin tanınması istenilen diğer dillerde çevirileri yapılmıştır. Bu çeviriler bir arama motorundan rastgele seçilen sayfalarda yer alan sonuçların her dilde benzer sayısal dokümanlara ulaşmayı sağlayacağı önerisiyle yapılmıştır. Bu sayede farklı kişiler tarafından yazılan ama benzer özellikler taşıyan toplumsal, bilimsel, politik vb. yazılar barındıran sayısal dokümanlar, veri kümelerinin oluşturulmasından kullanılmıştır. Bu kelimelerin kullanılmasıyla elde edilen sayısal dokümanların farklı dillerde aynı konuları içermesi ile önişlemler sonucunda elde edilecek eğitim veri kümelerinin rastlantısal sonuçlar içermemesi de hedeflenmiştir. Eğitim işlemi için sayısal dokümanların önişlemden geçirilmesi gerekmektedir. Bu önişlem sonucunda, Latin Alfabesi grubuna ait dillerin sınıflandırılması gerçekleşmektedir. Latin alfabesinin genişletilmiş grubunda 161 harf bulunmaktadır. Bu harflerden bazıları Almanca'da bulunan "ß" gibi sadece o dile özgü olmakla birlikte Almanca, Galce ve Türkçe'de olan "Ü" gibi birden fazla dilde ortak olarak bulunmaktadır. Bu aşamada sadece dile özgü harflerin kullanımı YSA'nın öğrenme sürecini etkileyerek, sürecin ezbere kaymasına neden olmaktadır. Bu yüzden, YSA'nın karar mekanizmasını iyileştirmek için tek dile özgü harfler kullanılmamıştır. Bunun için her harfin en az 2 dilde bulunması esas alınmıştır.

YSA uygulamalarının başarısı, uygulanacak yaklaşımlar ile yakından ilgilidir. YSA'larla çözümlenemeyecek problem sayısı çok az olmasına karşın başarılı bir tasarımı etkileyen pek çok faktör bulunmaktadır [16]. Bu tasarımlar genel olarak,

1. Uygun yapının seçimi
2. Bu yapıya uygun öğrenme algoritması ve uygulanan algoritmanın uygun parametrelerinin seçimi
3. Seçilen yapıya uygun giriş, ara katman ve çıkış sayılarının belirlenmesi
4. Ara katman nöron sayılarının yeteri kadar seçilmesi
5. Seçilen nöronlarda kullanılacak olan aktivasyon fonksiyonunun belirlenmesi
6. Eğitim ve test kümelerinin belirlenmesi ve bu kümelerde kullanılacak normalizasyon seviyelerinin belirlenmesi;

gibi zor ve zaman alıcı bir süreç gerektirmektedir. Bu süreç doğru şekilde işletilmediği durumlarda karmaşıklık artacağından YSA'nın performansı ciddi oranda etkilenecektir. Bu sebeple ağın büyüklüğü, öğrenme seviyesi ve ağırlıkların sınıflandırma seviyeleri de bir YSA'nın performansını ve genelleme kabiliyetini etkileyen başlıca faktörler arasındadır. Bir YSA uygun parametrelerle tasarlanırsa kararlı bir yapı ortaya çıkacağından kararlı sonuçlar elde edilecektir [16].

Bu makale çalışmasında, YSA parametrelerinin seçiminde literatürde önerilen yöntemlerden faydalanılmıştır. Bu parametreler öğrenme algoritması, giriş-çıkış sayıları, ara katman sayısı, transfer fonksiyonu ve ara katman nöron sayılarıdır.
İncelenen dillerin alfabelerinin farklı sayılarda birleşiminden oluştuğu için farklı giriş sayılarıyla yeniden tasarlanmıştır. Yapılan tasarımların başarıları Çizelge 1'de verilmiştir.

Çizelge 1'den de görülebileceği gibi en yüksek performansa sahip YSA modelinin belirlenmesi için yapılan testlerde 1. ara katmanda sigmoid transfer fonksiyonuna sahip 20 nöronlu yapı ile 2. ara katmanda tanjant hiperbolik transfer fonksiyonuna sahip 40 nöronlu yapının en iyi sonucu verdiği belirlenmiştir. Sistem sıfır hata oranına ulaşmak için 50 epok seviyesine kadar eğitilmiştir. Eğitim 1. epok seviyesinde ortalama hata kareleri toplamı 129.463, eğim ise 336.114/1e-010 olarak ölçülmüştür.

Çizelge 1. Uygun YSA parametrelerinin belirlenmesi için yapılan çalışmalar (Experimental Studies on selection of suitable Artificial Neural Network Structure)

YSA #	AKS	HKNS	TF	ÖA	MSE
1	2	5,5,1	S,TH,L	LM	3.5×10^{0}
2	**2**	**10,30,1**	**S,TH,L**	**LM**	**8.9×10^{-22}**
3	2	20,40,1	S,TH,L	LM	1.04×10^{-29}
4	2	30,40,1	S,TH,L	LM	9.97×10^{-24}
5	2	5,5,1	S,TH,L	GD	18.65×10^{0}
6	2	20,40,1	S,TH,L	GD	3.4×10^{0}
7	2	30,80,1	S,TH,L	GD	8.46×10^{0}
8	2	10,10,1	S,TH,L	GDM	18.66×10^{0}
9	2	20,20,1	S,TH,L	GDM	7.7×10^{0}
10	2	10,10,1	S,TH,L	GDM	8.25×10^{0}
11	2	30,50,1	S,TH,L	GDM	8.11×10^{0}
12	2	20,40,1	TH,S,L	LM	3.38×10^{-22}
13	3	20,30,40,1	S,TH,TH,L	LM	7.40×10^{-16}
14	3	20,30,40,1	S,S,TH,L	LM	2.28×10^{-26}
15	3	20,30,40,1	TH,S,TH,L	LM	3.54×10^{-19}

ÖA: Öğrenme Algoritması
AKS: Ara Katman Sayısı
HKNS: Her Katmanda Nöron Sayısı
TF: Transfer Fonksiyonu
TH: Tanjant Hiperbolik
S: Sigmoid
L: Doğrusal
MSE: Ortalama Hataların Karesi
LM: Levenberg-Marquardt geri yayılım
GD: Dereceli Azalan Geri Yayılım
GDM: Adaptif Öğrenme Oranlı Dereceli Azalan Geri Yayılım

Eğitim 5. epok seviyesine ulaştığı zaman ortalama hata kareleri toplamı 2.453 ve eğim 15.0804/1e-010 olarak ölçülmüştür. Eğitim 10. epok seviyesine ulaştığı zaman ortalama hata kareleri toplamı 8.11853e-006 ve eğim 0.00743012/1e-010 olarak ölçülmüştür.

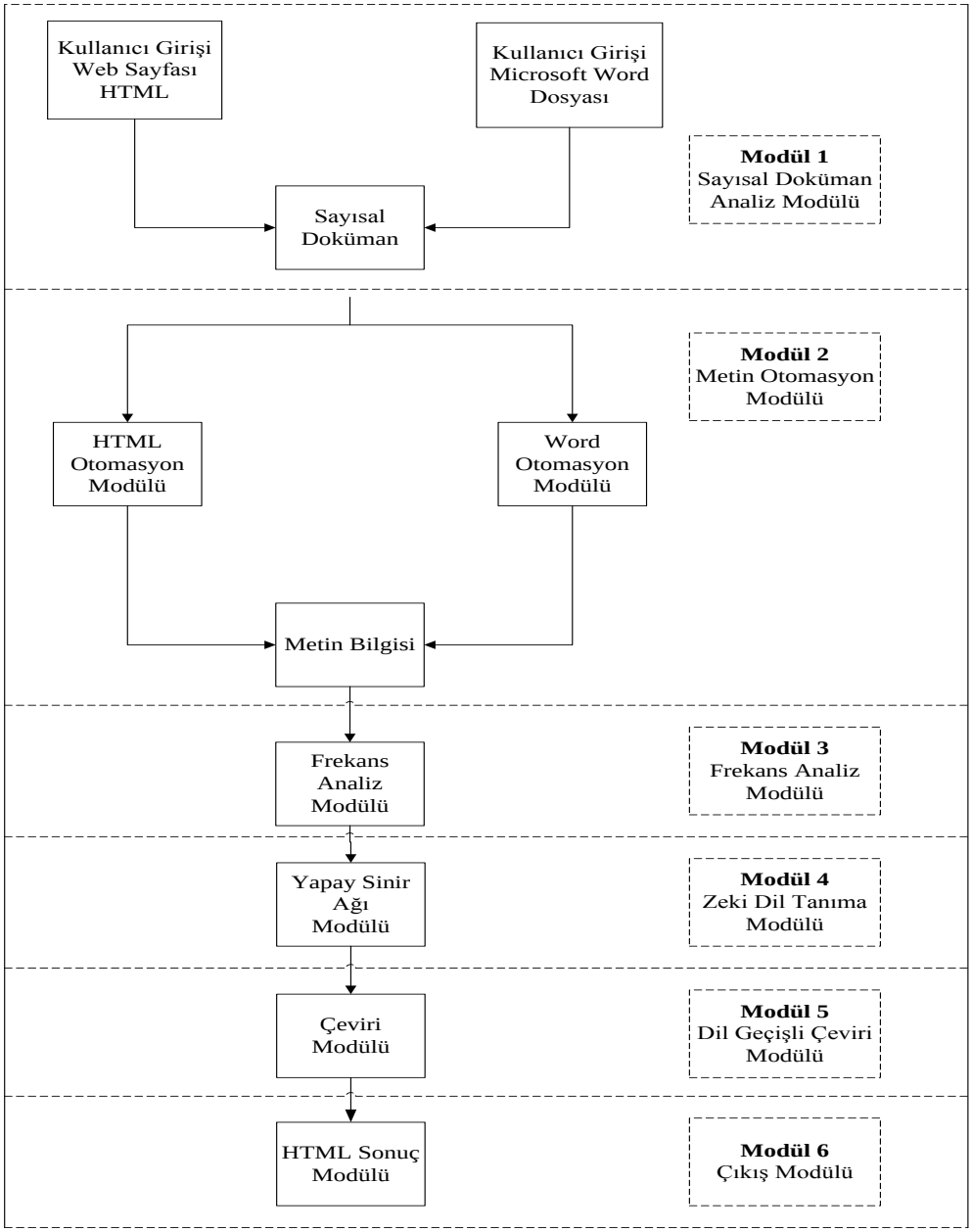
Eğitimin bu aşamasında hata oranının giderek düştüğü ve YSA modelinin uygulanan eğitim veri kümelerinden iyi bir genelleme yapabildiği görülmektedir. Hata oranı sıfıra yaklaşırken sistem varsayılan eğim minimuma ulaşarak 14. epok seviyesinde ortalama hata kareleri toplamı 1.04089e-029 ve eğim 9.32831e-015/1e-010 olarak ölçülmüştür. Eğitim, eğimin varsayılan değerine ulaşıncaya kadar

devam etmiştir. Bu eğitim esnasında temel almamız gereken en önemli kriter ortalama hata kareleri toplamıdır. Bu nedenle eğitimin son derece başarılı olduğu ortaya konmuştur.
GD, GDA ve GDM öğrenme algoritmalarının ise bu çalışmada LM algoritması kadar başarılı olmadıkları gösterilmiştir.

4. GELİŞTİRİLEN SINIFLANDIRMA ve OTOMATİK ÇEVİRİ SİSTEMİ (CLASSIFICATION and TRANSLATION SYSTEM)

Bu çalışma kapsamında dokümanları sınıflandırabilmek için bir sistem geliştirilmiş ve sistem performansını arttırmak için bir yöntem önerilmiştir. Önerilen yöntem kullanılarak, sayısal dokümanlardan elde edilen giriş verileriyle istenilen dile otomatik dönüşüm gerçekleştirebilen zeki bir dil tanıma uygulaması geliştirilmiştir. YSA kullanılarak gerçekleştirilen bu zeki dil tanıyıcı ve dönüştürücü yazılımının blok şeması Şekil 2'de verilmiştir. Şekil 2'de sunulan blok şemadan da görülebileceği gibi geliştirilen bu uygulama yazılımı, ön işlem, eğitim, test, sonuç-başarı oranı, otomatik dil çevirisi gibi işlemleri gerçekleştiren ve kesikli çizgiler ile birbirinden ayrılmış olan Sayısal Doküman Analiz Modülü, Metin Otomasyon Modülü, Frekans Analiz Modülü, Zeki Dil Tanıma Modülü, Dil Geçişli Çeviri Modülü ve Çıkış Modülü'dür.

Bu makale kapsamında gerçekleştirilen çalışma ile 15 dilde (Almanca, Arnavutça, Fransızca, Galce, Hırvatça, İngilizce, İrlandaca, İspanyolca, İtalyanca, Letonca, Macarca, Maltaca, Portekizce, Türkçe, Vietnamca) zeki dil tanıyıcı ve 41 dile dönüştürücü tasarımı için yapılan işlemler bu bölümde alt başlıklarla anlatılmıştır. Öncelikle önerilen yöntem olan Birleşim Tespit Yöntemi açıklanmıştır. Daha sonra da bu çalışma kapsamında geliştirilen sistemin modülleri tanıtılmıştır.



Şekil 2. Geliştirilen Yazılıma ait Blok Diyagram (Block Diagram of the Developed System)

4.1 Önerilen Yöntem: Birleşim Tespit Yöntemi (Suggested Approach: Unified Method)

Bu makale çalışmasında farklı dillerin tespitinde kolaylık sağlanması için bir yöntem önerilmiştir. Önerilen yöntem ortak alfabe kümesinin YSA üzerinde öğrenmeyi artırması ve matematiksel model oluşumunu etkilemesi sebebiyle patenti tarafımızca alınmış olan birleşim tespit yöntemleri kullanılmıştır. Bu yöntemlerin uygulanabilmesi için bir dile ait alfabe örüntüsünün hangi kurallar çerçevesinde birleştirileceğinin belirlenmesi gerekmektedir. Bu makale çalışmasında Latin Alfabesi ve bu alfabeden türeyen alfabelerin bağlı oldukları diller incelenmiştir. Eğitim veri kümeleri oluşturulurken, Latin alfabesi grubuna dâhil 15 farklı dile ait için sayısal (WORD) dökümanlar kullanılmıştır. Geliştirilen bu yaklaşımda test verilerinin üretilmesi için bir anahtar kelime tablosu ve Google arama motorundan faydalanılmıştır. Kelime tablosunda verilen Erkek, Kadın, İnsan, Dünya, Ülke, Bilim, Kültür, Sanat, Politika, Hayat kelimeleri bir dil içerisinde en sık kullanılan kelimeler olarak belirlenmiştir. Veri kümeleri oluşturulurken bu kelimelerin tanınması istenilen diğer dillerde çevirileri yapılmıştır. Bu çeviriler bir arama motorundan rastgele seçilen sayfalarda yer alan sonuçların her dilde benzer sayısal dokümanlara ulaşmayı sağlanacağı önerisiyle yapılmıştır. Bu sayede farklı kişiler tarafından yazılan ama benzer özellikler taşıyan toplumsal, bilimsel, politik vb. yazılar barındıran sayısal dokümanlar veri kümelerinin oluşturulmasında kullanılmıştır. Bu

anahtar kelimelerin kullanılmasıyla elde edilen sayısal dokümanların farklı dillerde aynı konuları içermesi ile önişlemler sonucunda elde edilecek eğitim veri kümelerinin rastlantısal sonuçlar içermemesini de hedeflenmektedir. Eğitim işlemi için sayısal dokümanların önişlemden geçirilmesi gerekmektedir. Bu önişlem sonucunda Latin Alfabesi grubuna ait dillerin sınıflandırılması gerçekleşmektedir.

Birleşim yöntemi, YSA'nın öğrenme performansını artırmak için önerilmiştir. Bu yöntem, tanınması istenilen dillere ait alfabelerin tek bir ortak alfabeye entegre edilme fikrine yeni bir bakış açısı getirmektedir. Bu yöntem kullanılarak oluşturulan ortak alfabe kümesinde standart ve genişletilmiş Latin alfabe kümesinin tanınması istenilen 15 dil için tekrar eden en az 2 harf seçilerek Birleşim Tespit Yöntemi Kümeleri oluşturulmuştur. Bu sayede bahsedilen 15 dil için her dil alfabesi içinde 2 kez ve daha fazla tekrar eden harflerden ortak alfabe kümesi oluşturulmuştur. Bu ortak alfabe kümesi ile tasarlanan sistem bütünü Birleşim Tespit Yöntemi (BTY) olarak anılmaktadır. Bu yöntemin adımları sırasıyla sunulmuştur.

1. Tanınması istenilen dillere ait alfabeler çıkartılır.
2. Bu alfabelere ait birleşim kümesi bulunur.
3. Giriş metni (web sayfası, word vs. dokümanlar) sisteme giriş olarak uygulanır.
4. Sistem alfabelerin birleşim kümesinde bulunan karakterlere bir sayıcı verir.
5. Bu sayıcı giriş metni içinde yer alan harfleri, alfabe birleşim kümesiyle karşılaştırarak tekrar eden harflerin sayıcı değerini bir artırır. Bu sayede alfabe birleşim kümesinde bulunan hangi harflerin metin içinde kaç kez yinelendiği bulunur.
6. Metin analizi tamamlandıktan sonra, harf sayıcıları toplanarak yinelenen toplam harf sayısı elde edilir.
7. Sistem her harf sayıcının giriş metni üzerindeki yüzdelik dağılımını hesaplar.
8. Eğitim kümesi oluşturulurken bu yüzdelik dağılımlara, incelenen dile ait bir sıra numarası ağın çıkışına verilir.
9. Eğitilen ağ, test edilirken ağa uygulanan yüzdelik değerler neticesinde ağın ürettiği çıkış, eğitimde kullanılan sıra numaraları ile karşılaştırılarak test metninin hangi dile en çok yakınsaması ve yüzdelik hatası bulunur.
10. Elde edilen sonuçlar çeviri programları, arama motorları, ofis programları gibi otomasyon yazılımlarında kullanılarak kullanıcılara sunulur.

Geliştirilen Sistem ve Modülleri (System and Modules)

Geliştirilen sistem ve modülleri Şekil 2'de verilmiştir. Şekil 2'de sunulan blok diyagram da bahsedilen Sayısal Doküman Analiz Modülü (SDAM), Metin Otomasyon Modülü (MOM), Frekans Analiz Modülü (FAM), Zeki Dil Tanıma Modülü (ZDTM), Dil Geçişli Çeviri Modülü (DGÇM) ve Çıkış Modülünün (ÇM) detayları aşağıda sırasıyla açıklanmıştır.

Sayısal Doküman Analiz Modülü

Bu modül, öncelikle içerik dili tanınması istenilen internet sitesinin veya WORD dokümanının varsayılan karakter kodlama kümesinin bir kaydını alır. Bu kayıt üzerinde bir analiz yapılarak sayfanın veya dokümanın karakter kodlama kümesi tespit edilir. Bu karakter kodlama kümesi ile sayfa bir kez daha sunucudan istenilir. Sisteme ulaşan kaynak koduna ek bir analiz daha yapılarak sayfaya ait bağlantı derinlikleri bulunur. Bu bağlantı derinlikleri sayesinde, istenilen sayfaya ait kaç alt bağlantı olduğu, daha hassas bir analiz için kullanıcının bilgisine sunulmaktadır.

Bu modül içerisinde kullanıcıların Microsoft Word dosyalarını sisteme giriş olarak uygulamalarını sağlayan iki alt sistem bulunmaktadır. Bu sistemler dosyaları BYTE tabanlı olarak işleme alarak akışları (stream) sonraki modüle iletirler. Ayrıca, bu modül de kullanıcı tarafından girilen düz metinler ise metin otomasyon modülüne doğrudan aktarılır.

Metin Otomasyon Modülü (MOM)

Sayısal doküman analiz modülü tarafından elde edilen internet sayfalarına ait kaynak kodları ile Microsoft Word dosya akışları bu modülde otomasyon sistemine girerken, kullanıcı tarafından girilen düz metin bilgisi sonraki modüle aktarılmaktadır. MOM tasarlanırken, Microsoft Ofis SDK paketi yerine açık kaynak kodlu dağıtımı yapılan Cellbi Microsoft Word Otomasyon Kütüphanesi [22] kullanılmıştır. Bu sayede kullanıcıların sistemlerine farklı yazılımlar kurması ya da bu yazılımları satın almadan kullanmaları sağlanmıştır. Geliştirilen otomasyon yazılımı sayesinde, ilk modülden akış dizisi olarak alınan dosyalar ilgili yapılarda işlendikten sonra düz metin bilgisi elde edilmektedir.

Frekans Analiz Modülü (FAM)

Frekans Analiz Modülü'nde daha önceden açıklandığı gibi birleşim tespit yöntemi kullanılmıştır. Birleşim tespit yöntemi aynı alfabelerde 2 kez ve daha fazla tekrar eden harf kümesini belirlemektedir. Bu yöntemde 60 karakter örüntüsü bulunmaktadır. Bu makale kapsamında geliştirilen yapay sinir ağı modeli metin-harf frekans bilgisi ile dil arasındaki matematiksel ilişkiyi öğrenmektedir. Bu matematiksel ilişkinin girişine uygulanacak harf frekans analizi ise bu modül tarafından yapılarak giriş metni üzerinden birleşim yöntemi kümesi esas alınarak harflerin yüzdelik frekans dağılımları hesaplanmaktadır.

Zeki Dil Tanıma Modülü (ZDTM)

Bu makale kapsamında önerilen birleşim yöntemi kullanılarak daha önceden dilleri bilinen Microsoft Word dosyalarının metin otomasyon modülünden elde

edilen metin bilgileri kullanılarak yapılan frekans analizleri ile ilgili dil arasındaki matematiksel ilişkiyi öğrenen YSA modeli bu modülün temelini oluşturmaktadır. Bu YSA modeli sayesinde sisteme uygulanan bir internet sayfası Microsoft Word dosyası ile düz metin bilgisinin dil tanıma sürecine aktarılması yapılabilmektedir. Bu YSA modelinin eğitim ve test süreci, dinamik bağlantı kütüphanesi yapısında tasarlanmıştır. FAM tarafından bu modüle gönderilen sayısal dokümana yüzdelik metin frekans bilgisi dinamik bağlantı kütüphanesi üzerinden daha önceden eğitilmiş ağ yapısına ulaşarak giriş olarak uygulanmakta ve tekrar aynı yolu izleyerek YSA çıkışından elde edilen sonucu, kullanıcı ara yüzüne iletmektedir. Bu sayede kullanıcılar giriş olarak uyguladıkları sayısal dokümanın dilini, bir yaklaşık sonuçtan n yaklaşık sonuca kadar olabilme yüzdelerini bu çalışmada sunulan zeki sistem sayesinde otomatik olarak elde edebilmektedir.

Dil Geçişli Çeviri Modülü (DGÇM)

Bu modül sayesinde kullanıcılar dillerini hiç bilmedikleri sayısal dokümanları literatürde önerilen dil çevirici sistemler ile dönüştürebilmektedir. Bu modül Google Translator tarafından geliştirilen yöntemi temel almaktadır. Google tarafından parametrik olarak kullanıma açılan sistem sayesinde ZDTM'den elde edilen dil bilgisi, kullanıcı ara yüzüne yansıtılarak, kullanıcıların metin bilgisini Google Translator tarafından desteklenen 41 farklı dile otomatik olarak dönüştürmektedir. Bu sayede Microsoft Word dosyaları ise akış düzenleri bozulmadan metin bilgisi olarak çeviri işlemine tabi tutulmakta ve otomatik olarak istenilen dile çevrilmektedir.

Çıkış Modülü

Bu modül, daha önceki işlemleri tamamlayan kullanıcıların dil tanıma işleminden sonra internet sayfalarını, Microsoft Word dokümanları ve düz metin dosyalarına ait çeviri metinlerini bütünleşik olarak internet tarayıcısında görüntüleyebilme, kayıt edebilme ve yazıcı çıktılarını kolaylıkla alınabilmesini sağlamak için tasarlanmıştır.

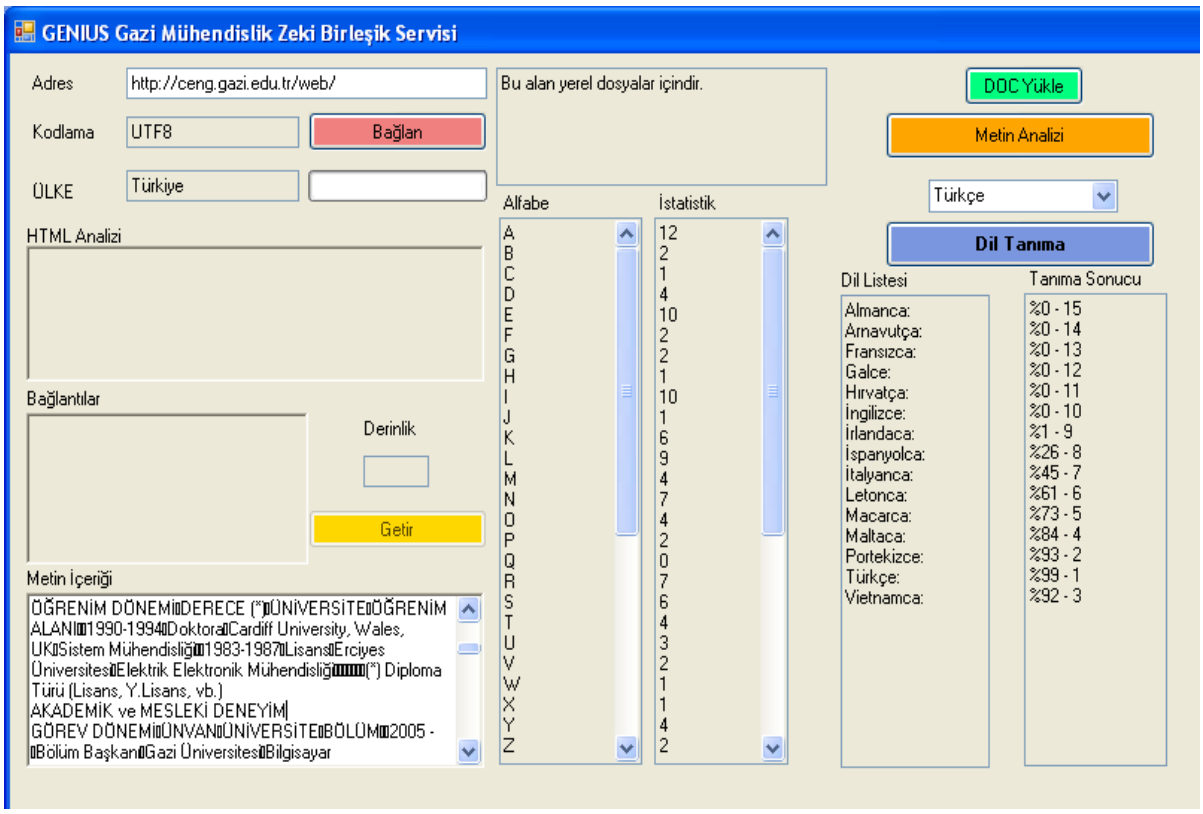
5. DENEYSEL SONUÇLAR (EXPERIMENTAL RESULTS)

Bu makale çalışmasında, kullanıcıların web ortamlarını daha etkin kullanabilmelerini sağlamak amacıyla web tabanlı otomatik dil tanıma ve çeviri sistemi geliştirilmiştir. Geliştirilen sistem, sadece bazı dillerde kullanılabilmektedir.

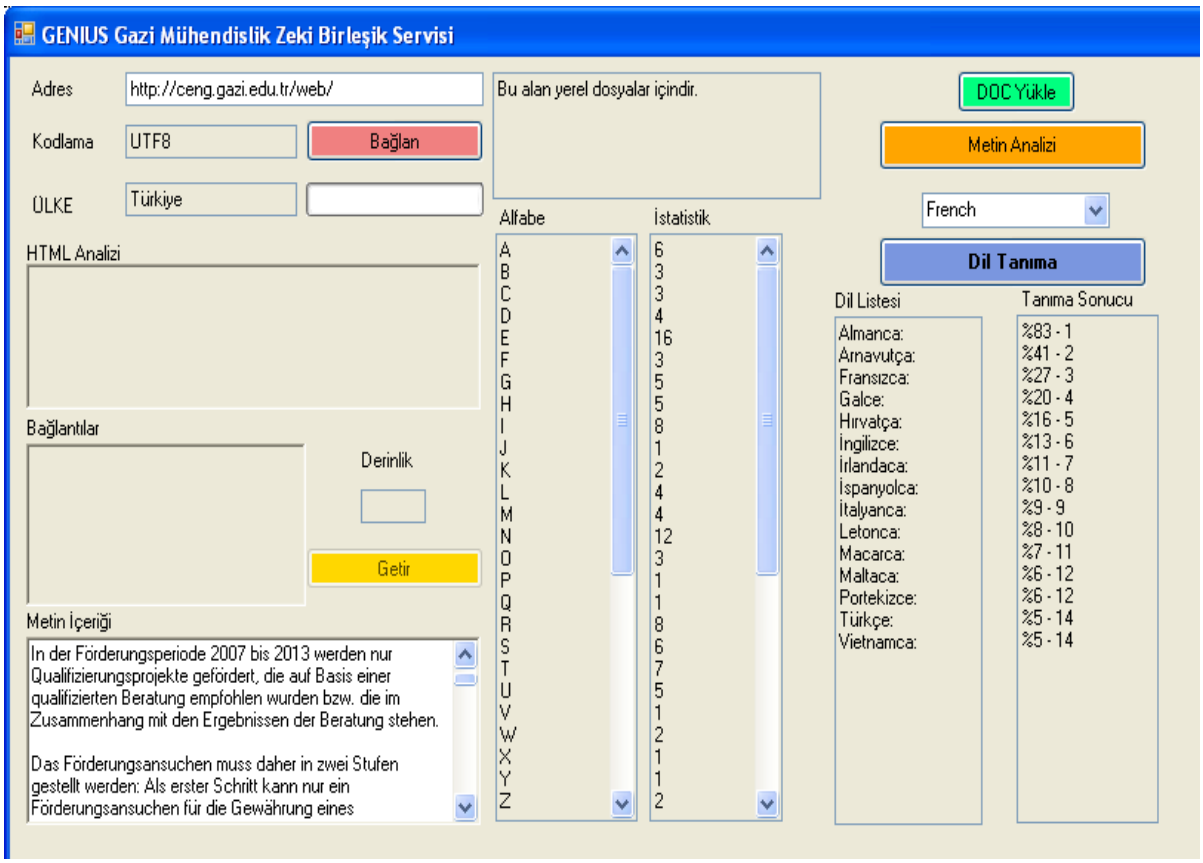
Bunun sebepleri ise YSA'nın öğrenme sürecinde performansının düşük olması, dil tanıma başarısının yetersizliği sadece web sitelerinde çalışması, çeviri yeteneklerinin kullanılan çeviri yazılımıyla sınırlı oluşu, sunucu bilgisayar ve bant genişliğine bağımlı olması ve tanınması istenilen dil sayısı arttıkça bu artışa paralel olarak artan eğitim veri kümesi ve YSA giriş sayısındaki artma ve en önemlisi bizlerin farklı dil içeriklerini analiz etmedeki yetersizliğimiz olarak sıralanabilir Karşılaşılan en büyük güçlük ise web sayfalarının aynı standart ve formatta bulunmamalarından kaynaklanan içeriğin doğru analiz edilememesi sorunudur.

Bu sebeple tanınması istenilen dil sayısı en çok kullanılan diller ile sınırlandırılmıştır. Geliştirilen sistemde önerilen yöntem ile farklı teknolojiler kullanılarak geliştirilen sistemin içerik tanımlama başarısı oldukça arttırılmıştır. Bu teknolojiler, bir bütün olarak entegre olduklarında önerilen alfabetik birleşim yöntemi ışığında yazılım tasarım ve test süreçleri gerçekleştirilmiştir. Bu yazılım sayesinde internet ortamında sıklıkla karşımıza çıkan ve literatürde geniş kabul görmüş olan Microsoft Word dosya formatı için dil tanımlama işlemleri başarıyla yapılabilmektedir. Sistemin test edilmesi amacıyla GENIUS (Gazi Engineering Intelligent Unified Service) yazılımı geliştirilmiştir. Bu yazılım kendi başına çalışır (standalone) olduğu için herhangi bir hizmet sağlayıcı sunucu veya bant genişliği kapasitesi gibi kısıtlar giderilmiştir. Bu çalışmada, kullanılan özel bileşenler sayesinde web içeriklerinin analiz edilmesinde yaşanan betik dil analizi gibi güçlüklerde ortadan kaldırılmıştır.

GENIUS yazılımı; HTML ve WORD dokümanlarını işleyebilen bir dil tanıma yazılımı olup Türkçe ve Almanca dilinde yazılmış Word belgeleri için iki örnek çalışma çıktısı Şekil 3'te verilmiştir. Şekil 3'de verilen yazılım hem HTML hem de Microsoft Word dokümanlarını analiz edebilmektedir. HTML dokümanlarının analiz edilmesi için "Adres" bölümüne analizi yapılmak istenilen web sayfasının adresi ve sayfa içerisinde incelenmesi istenilen alt link sayısı girilmektedir. Sistem ilgili web sayfasının kaynak koduna sayfaya ait karakter kodlaması ile gerçek zamanlı erişerek, elde ettiği betik kodlama içinden karakter analizi için kullanılacak metin girişleri elde edilir. Yazılımın "DOC Yükle" kısmında ise analizi yapılmak istenilen Microsoft Word belgesi içinden gerçek zamanlı olarak metinsel kısımlar çıkartılır ve sisteme giriş olarak uygulanır. "Metin Analizi" ile HTML ya da Word dokümanından elde edilmiş metin girişi önceki bölümlerde anlatılan Birleşim Tespit Yöntemi algoritması ile frekans analizine tabi tutulur. Sistemin son aşamasında ise "Dil Tanıma" butonu yardımıyla karakter frekans analizi ile YSA yapısı kullanılarak giriş dokümanına ait "Dil Listesi" ve dil "Tanıma Sonucu" açılır bir listeden seçilerek ekranda gösterilir. Geliştirilen sistemde elde edilen sonuçlar Word dokümanları için Çizelge 2'de HTML dokümanları için Çizelge 3'de verilmiştir.



(a) Türkçe Word belgesi uygulama çıktıları (As an example of Turkish Word Input and Output Result)



(b) Almanca Word belgesi uygulama çıktıları (As an example of Turkish Word Input and Output Result)
Şekil-3. GENIUS Yazılımı

Çizelge 2. Geliştirilen sistemin Microsoft Word belgeleri için sınıflandırma sonuçları (Experimental Results of Randomly Selected Microsoft Word Documents over Internet)

Format	DOC Dosya Tipi					
Örnek	#1	#2	#3	#4	#5	Tanıma Başarı Ortalaması
Almanca	92%	98%	98%	90%	83%	90%
Arnavutça	92%	93%	99%	98%	99%	96%
Fransızca	98%	87%	88%	85%	97%	91%
Galce	82%	99%	99%	96%	96%	94%
Hırvatça	97%	97%	98%	94%	95%	96%
İngilizce	98%	95%	95%	86%	95%	93%
İrlandaca	98%	98%	99%	99%	97%	98%
İspanyolca	99%	98%	99%	94%	96%	97%
İtalyanca	96%	96%	94%	93%	94%	94%
Letonca	95%	97%	84%	94%	96%	93%
Macarca	96%	97%	92%	99%	98%	96%
Maltaca	91%	83%	99%	98%	99%	94%
Portekizce	95%	91%	97%	90%	92%	93%
Türkçe	92%	99%	99%	99%	99%	98%
Vietnamca	99%	98%	99%	95%	97%	98%

Çizelge 3. Geliştirilen sistemin HTML belgeleri için sınıflandırma sonuçları (Experimental Results of Randomly Selected Web Pages)

Format	Her Dil için 15 Web Sitesi
Dil	Tanıma Başarı Ortalaması/Identification Performance
Almanca	93%
Arnavutça	99%
Fransızca	83%
Galce	94%
Hırvatça	75%
İngilizce	93%
İrlandaca	98%
İspanyolca	98%
İtalyanca	98%
Letonca	97%
Macarca	96%
Maltaca	99%
Portekizce	98%
Türkçe	99%
Vietnamca	99%

Çizelgelerde dosya tipine göre farklı dokümanları tanıma yüzdeleri verilmiştir. Çizelgelerde verilen değerler, eğitimlerden elde edilen en yüksek başarım değeri olduğu için tüm dokümanların başarı ortalamaları yüksektir. Bu çalışmada bir dili tanıma için en yüksek elde edilen sınıflandırma değeri dikkate alındığından, hem DOC hem de HTML dokümanları için tanıma oranı %100 olarak bulunmuştur. Diğer bir ifadeyle tanıma için elde edilen en yüksek başarım değeri, tanınacak olan dile ait olduğu için dillerin tamamı tanınmıştır.

6. SONUÇLAR ve ÖNERİLER (CONCLUSION and SUGGESTIONS)

Bu çalışmada geliştirilen ve başarıyla sunulan YSA tabanlı zeki bir Dil Tanıyıcı ve Dönüştürücü sisteminde;

- YSA'nın Dil Tanıyıcı tasarımında başarılı olduğu,
- YSA'ların ara katman sayısı, nöron sayısı ve seçilecek olan fonksiyon tipine bağlı olarak başarıların arttığı,
- Eğitim ve test için gerekli olan ortak alfabe kümelerinin birbirlerine yakınsayan dillerde karmaşıklığın artmasına neden olduğu,

- YSA'ların farklı dillerde başarılarının ölçülmesinin şart olduğu,
- Tasarlanan YSA yapısının kolaylıkla gerçek-zamanlı uygulamalarda kullanılabileceği,
- Ülkemizde ve dünyada zeki dil tanıma sistemlerine yönelik yeterli çalışma bulunmadığı, bu çalışmanın bütün çalışmaların önünü açacağı
- Literatürde dil tanıma üzerine birçok çalışma bulunmasına rağmen veri kümeleri elde etmek için kullanılan yöntemlerin yetersiz olduğu görülmüştür.

Bu makale çalışmasında önerilen sistemin en büyük avantajı dil tanıma sistemi için kullanılan yöntemin dil grubuna özgü olmayışı ve diğer diller içinde sistemin dönüştürülme sürecinin esnekliğidir. Dil tanıma veritabanı için önerilen birleşim yöntemlerinin tanınması istenilen dillere ait alfabelerde tekrarlanan harf sayıları hedef alınmaktadır.

Gerçekleştirilen çalışma ile literatüre, istatistiksel dil tanıma metotlarının yanında farklı bakış açısı kazandırmaktadır. Literatürde de belirtildiği gibi zeki yaklaşımların DT tasarımında kullanılmasının doğru bir tercih olduğu, farklı yapılar seçerken MLP yapısının LM gibi güçlü algoritmalarla eğitilmesiyle YSA'ların DT tasarımında başarıyı arttırdığı tespit edilmiştir.

Bu çalışma sonucunda geliştirilen sistem ile kullanıcılardan Word dokümanları veya web sayfaları verileri giriş olarak aldıktan sonra, dil tespiti sonucunda kullanıcının dönüştürmek istediği dile yapılan tercüme yine yazılım üzerinde bulunan dâhili web göstericisi arabiriminde sunulmaktadır. Bu nedenle kullanıcılar farklı dillerde yazılmış dokümanları istedikleri dile başka hiçbir araç kullanmadan zeki yöntemler ile dönüştüre-bilmektedirler.

Bu çalışmanın internetin daha verimli kullanılmasına, farklı dil ve kültürlerde yapılan çalışmaların kolaylıkla okunup öğrenilmesine ve internette karşılaşılan pek çok problemin çözümüne büyük katkılar sağlayacağı değerlendirilmektedir. Bu sistem sayesinde, internetin geniş kitlelere ve halk gruplarına yayıldığı bir dönemde küreselleşme ile ortaya çıkan ana dil yayınlarının başka milletler tarafından takip edilmesinin önünü açacağı değerlendirilmektedir. Sistemin henüz kendi kendine öğrenme sürecini bulunmadığından ölçeklenebilir veritabanı ve ağ yapısının sonraki çalışmalarda kullanıcılardan gelen geri beslemeler doğrultusunda güncellemesi gerekebileceği değerlendirilmektedir.

7. KAYNAKLAR (REFERENCES)

1. Padro M., Padro L., "Comparing Methods for Language Identification" ***Procesamiento del Lenguaje Natural,*** Barcelona, 33-35 (2004).

2. Botha G.R., Zimu V.Z., Barnard E., "Text- based language identification for the South Africanlanguages", ***SAIEE Africa Research Journal***, Cape Town, 141-146 (2007).

3. El-Shishiny H., Troussov A., McCloskey DJ., Takeuchi M., Nevidomsky A., Volkov P., "Word Fragments Based Arabic Language Identification", ***NEMLAR Conference on Arabic Language Resources and Tools***, Mısır, 23-26 (2004).

4. Kruengkrai C., Srichaivattana P., Sornlertlamvanich V., Isahara H., "Language Identification Based on String Kernels" ***Communications and Information Technology***, Pekin, 896-899 (2005).

5. Zavarsky P., Wada S., Mikami Y.,"Language and Encoding Scheme Identification of Extremely Large Sets of Multilingual Text Documents", ***The 10th Machine Translation Summit***, Puket, 354-355 (2005).

6. Peng F., Schuurmans D., Wang S.,"Language and Task Independent Text Categorization with Simple Language Models", ***North American Chapter of the Association for Computational Linguistics - Human Language Technologies***, Edmonton, 110-117 (2003).

7. Nair A.S., Nair V. V., Chandra V. S. S., "Hidden Markov Model Based Identification of Transliterated Regional Language Words in Text Documents", ***Twentieth International Joint Conference on Artificial Intelligence***, Haydarabad, 87-91 (2007).

8. Ahmed B., Cha S-H.,, Tappert C.,"Language Identification from Text Using N-gram Based Cumulative Frequency Addition", ***Student/Faculty Research Day,*** New York, 121-128 (2004).

9. Constable P.G., "Toward a Model for Language Identification", ***Summer Institute of Linguistics International Working Papers"***, *D*ublin (2002).

10. Adams G., Resnik P., "A Language Identification Application Built on the Java Client/Server Platform", ***The European Chapter of the Association of Computaional Linguistics Workshop***, İspanya (1997).

11. Ölveck T., "N-Gram based Statistics Aimed. at Language Identification", ***Student Research Conference in Informatics and Information Technologies***, Brastilava, 1-7 (2005).

12. Bilcu, E.B., Astola J., "A Hybrid Neural Network for Language Identification from Text", ***Machine. Learning for Signal Processing Conference***, Maynooth, 253-258 (2006).

13. Liu Y-H., Chang F., Lin C-C., "Language Identification of Character Images Using Machine Learning Techniques", ***International Conference on Document Analysis and Recognition***, Seul, 630-634 (2005).

14. Zhu G., Yu X.,Li Y., Doermann D., "Unconstrained Language Identification Using A

Shape Codebook", ***The 11th International Conference on Frontiers in Handwritting Recognition***, Montreal, 13-18 (2008).

15. Baykan E., Henzinger M., Weber I., "Web Page Language Identification Based on URLs", ***International Conference on Very Large Data Bases***, Auckland, 176-187 (2008).

16. Sağıroğlu, Ş., Beşdok, E., Erler, M., "Mühendislikte Yapay Zeka Uygulamaları-1:Yapay Sinir Ağları", ***Ufuk Kitabevi***, Kayseri, 10-100 (2003).

17. Sağıroğlu Ş., Yavanoğlu U., Güven E.N., "Web Based Machine Learning for Language Identification and Translation", ***International Conference on Machine Learning and Applications***, Ohio, 280-285 (2007).

18. Aldrich, A., "R. A. Fisher on Bayes and Bayes'Theorem", **Bayesian Analysis**, 3, No. 1, pp.161–170, (2008)

19. İnternet: Google Yazılım "Web Tabanlı Dil Çeviri Aracı Web Sayfası" http://translate.google.com/ (2008)

20. İnternet: Microsoft Yazılım "Visual Studio 2005 C# Windows Form Uygulaması Yazılım Geliştirme Aracı", http://msdn.microsoft.com/en-us/vstudio/default.aspx (2005).

21. İnternet: Mathworks Yazılım "Matlab R2007B Deployment Tool, Dinamik bağlantı Kütüphanesi Geliştirme Aracı", http://www.mathworks.com/products/new_products/ release2007b.html (2007).

22. İnternet: Cellbi Yazılım "Microsoft Word OLE Otomasyon Bileşeni" http://www.cellbi.com/products/docframework.aspx (2008).

23. Takçı H., Soğukpınar İ. "Letter Based Text Scoring Method for Language Identification", **Springer Lecture Notes in Computer Science,** Vol. 3261/2005 283-290 (2004).