Elektronik Sosyal Bilimler Dergisi www.esosder.org
Electronic Journal of Social Sciences info@esosder.org
Güz-2008 C.7 S.26 (064-070) ISSN:1304-0278 Autumn-2008 V.7 N.26

BİLGİSAYARLI DİL TANIMLAMADA DİLBİLİMSEL YAKLAŞIMLAR VE BİR YAZILIM DENEMESİ[1]

LINGUISTIC TECHNICS ON LANGUAGE IDENTIFICATION AND A SOFTWARE PROJECT

Yrd.Doç.Dr.Ahmet TARCAN tarcan@dicle.edu.tr
Uzman Fahri ÇAKAR fcakar@dicle.edu.tr

ÖZ

Doğal Dil İşleme alanında *N-Gram, Markov Model* gibi çeşitli teknikler kullanılmaktadır. Bu çalışmada dilbilimsel bazı ölçütler kullanılarak algoritma oluşturulmaya çalışılmış ve ana dili Türkçe olmayan bilgisayar kullanıcılarına yönelik bir yazılım geliştirilmiştir. Sözü edilen dilbilimsel ölçütler arama motorlarına uygulandığında da aynı şekilde etkili sonuçlar elde edilebilmektedir.

Anahtar Kelimeler: Doğal Dil İşleme, Dil Tanımlama, Dilbilim, İletişim

ABSTRACT

Various techniques, such as *N-Gram* and *Markov Model* are used in natural Language Processing area. In this study, using some linguistic criteria it was aimed to compose algorithm and software has been developed for computer users whose native language is not Turkish. When the mentioned linguistic criteria are also applied to the search engines, same results are obtained in the same way.

Keywords: Natural Language Processing, Language Identification, Linguistic, Communication

1.GİRİŞ/TARİHÇE

1949 Warren Weaver'ın *Çeviri* adlı çalışması Doğal Dil İşleme alanında önemli bir tarihtir. 1950'li yıllarda Amerika da otomatik çeviri ile başlayan doğal dil işleme, bu yıllarda ABD hükümetinden büyük fonlar bulabiliyordu. İngilizce-Rusça makine çevirisi alanında araştırma projelerine ümit bağlayan devlet bu projeleri desteklemekten çekinmiyordu. Kamu fonları yaklaşık 48 ayrı çalışma gurubunu finanse ediyordu.

Rakamlar 1960'lı yılların ortalarında 20 milyon doları bulmuştu. Ancak çalışmalardan beklenen elde edilememiş, harcanan paraların boşuna harcandığı duygusu ortaya çıkmıştı. US Bilimler Akademisinin 1966 yılında Makine Çevirisine yönelik hazırladığı olumsuz rapor bu alandaki fonların bir anda durdurulmasına yol açmış, dolayısıyla Doğal Dil İşleme araştırmaları önemli bir kesintiye uğramıştı. ABD hükümeti uzun bir süre bu fonları askıya almış ve bu alandaki hiçbir projeyi desteklememişti. Bu süreç 1968 yılında Dr.Peter Toma'ın Kaliforniya da bir şirket kurup *Systran* adlı bir sistem geliştirmesi ile yeniden bir ivme

[1] Bu çalışma 12-13 Mayıs 2006 tarihlerinde Maltepe Üniversitesinde düzenlenen 20. Ulusal Dilbilim Kurultayı Programında sunulan "Bilgisayarlı Dil Tanımlamada Dilbilimsel Yöntemler" adlı bildiriden hareketle hazırlanmıştır.

kazanabilmişti. Şirketin geliştirdiği programlar 1970'li yıllarda Amerika Birleşik Devletlerinin hava kuvvetlerinde Rusça-İngilizce çeviri amaçlı kullanılmaya başlanmış, 1974–1975 yılları arasında ise aynı şirketin hazırladığı sistemler NASA tarafından Sovyetler Birliğinin Apollo-Soyouz uzay projesine ulaşmak için kullanılmıştı. 1975 yılına gelindiğinde ise Dr.Toma'ın hazırladığı Systran makine çevirisi programları Avrupa Birliği komisyonlarında İngilizce-Fransızca çeviri amaçlı kullanıldı.

Doğal Dil İşleme Tarihinde edebiyatın özelliklede romancıların, bilim kurgu yazarlarının da payını unutmamak gerekir. İkinci Dünya Savaşında İngiliz ordusunda radarlardan sorumlu subay olarak görev yapan Clarke, 1968 yılında daha sonra filme dönüştürülen ünlü *Uzay Macerası* adlı bilim kurgu filminin eserini kaleme aldı. Filmde akıllı bilgisayar HAL astronotlarla konuşuyor, çevresindeki hemen her şeyi algılıyor, satranç oynadığı astronotlardan birisini de sonunda öldürüyordu(Makine çevirisinin tarihi ile ilgili genel bilgiler kaynakçada geçen internet sitelerinden özetlenerek aktarılmıştır. 1950-1966 yılları arasında bu alanda Amerika'da harcanan para miktarı ve çalışma gruplarının sayısı ile ilgili farklı rakamlar veren kaynaklar vardır. Daha geniş bilgi için bkz. www.hutchinsweb.me.uk/PPF-4.pdf).

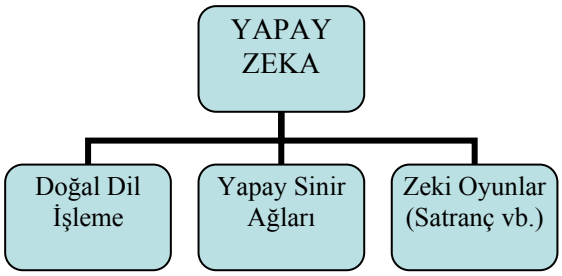
2.DİLBİLİM VE BİLGİSAYAR BİLİMLERİ ARASINDA YENİ BİR DİSİPLİN

Günümüzde Doğal dil işleme Dilbilim, Yazılım Mühendisliği ve Matematik bilimlerinin bir alt dalı olarak ortaya çıkmıştır.

Dil işleme ve tanımlama teknikleri çeşitli düzeylerde gerçekleşmektedir;

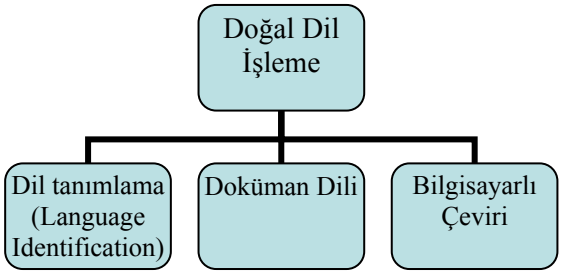
1-Konuşma Dilini tanıma
2-Yazılı Dili Tanıma
3-El yazısını tanıma
4-Metni konuşmaya çeviren yazılımlar (Microsoft Sam, Cepstral Swift Talker vb.)

İngilizce Doğal Dil işleme başlığının açılımı olarak, NLP (Natural Language Processing) yapay zekâ ve dilbilimin alt kategorisi olarak tanımlanmaktadır. Türkçe, İngilizce, Almanca, Fransızca gibi doğal dillerin (insana özgü tüm diller) işlenmesi ve kullanılması amacı ile araştırma yapan bilim dalının adıdır.



"*Doğal Dil İşleme, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır. Bu çözümlemenin insana getireceği kolaylıklar, yazılı dokümanların otomatik çevrilmesi, soru-cevap makineleri, otomatik konuşma ve komut anlama, konuşma sentezi, konuşma üretme, otomatik metin özetleme, bilgi sağlama gibi birçok başlıkla özetlenebilir. ….Örneğin, tüm kelime işlem yazılımları birer imla düzeltme aracı taşır. Bu araçlar aslında yazılan metni çözümleyerek dil kurallarını denetleyen doğal dil işleme yazılımlarıdır.*"[2]

[2] http://tr.wikipedia.org/wiki/Do%C4%9Fal_Dil_%C4%B0%C5%9Fleme



3- KAVRAMSAL SORUN

Batı dillerinden Türkçeye Doğal Dil İşleme şeklinde aktarılan bu disipline komşu bir takım farklı kavramlarda kullanılmaktadır. İngilizce ve Fransızcada genellikle mühendislik alanına yakın terimlerin kullanıldığı dikkati çekmektedir:

İngilizce:
Natural Language Processing (NLP)
Computational Linguistics
Quantitative Linguistics
Linguistics Engineering
Automatic treatment of the natural language
Engineering of the language
Computerized Linguistics

Fransızca
Linguistique Informatique
Informatique Linguistique
Traitement Automatique des Langues Naturels(TALN)
Linguistique Computationalle
Le calcul linguistique
L'Ingénierie de la Langue
L'ingénierie linguistique
Linguistique Mathematique
Linguistique Algorithmique
Linguistique Quantitative
Linguistique Calculatoire

Türkçe
Doğal Dil İşleme
Bilişimsel Dilbilim
Dil Mühendisliği
Hesaplamalı Dilbilim

Bu kavramlardan her biri bir bilim dalını mı, bir teknolojiyi mi, yoksa belli bir alanda çalışan araştırma guruplarını mı ifade ediyor? Kesin olarak bildiğimiz şey bu kavramların genellikle Dilbilim, Yazılım mühendisliği ve Matematik bilimleri (Cebir, Mantık, İstatistik) çevresinde gruplandığıdır. Boğaziçi Üniversitesinden Prof. Dr.Bülent SANKUR' un hazırladığı Bilişim Sözlüğünde *Computational Linguistics* Hesaplamalı Dilbilim şeklinde değil de Bilişimsel Dilbilim olarak Türkçeye kazandırılmıştır. Sözlükte söz konusu alan bir bilim dalı olarak değil *çalışma alanı* olarak tanımlanmıştır:

"Computational Linguistics: Bilişimsel Dilbilim.Genellikle bilgisayar kullanımıyla gerçekleştirilen istatistiksel dilbilim çalışmalarıyla, biçimbilgisi, sözdizimi, anlambilim düzeylerinde otomatik çözümlemeleri içeren ve doğal dilden doğal dile otomatik çeviri çalışmalarını, otomatik biçem araştırmalarını kapsayan çalışma alanı.(Sankur-2004)"

Aynı sözlükte Natural Language Processing ise Doğal Dil İşleme şeklinde Türkçeye aktarılmış ve bağımsız bir bilim dalı şeklinde tanımlanmıştır.

"Natural Language Processing: Doğal Dil İşleme, insan dilini çözümleme ve anlama tekniklerini konu edinen ve insan makine iletişimi amaçlı doğal dile benzer diyalog üretmeye

*çalışan bilim dalı. Doğal dil işleme ile bilgisayara insanların dili kullanma yaklaşımlarını da öğretmiş olur (*A.g.e s.531*)"*

4.UYGULAMA ALANLARI

Doğal Dil İşleme Bilgisayarların insan dilini algılamaları gereken her yerde kullanılabilir, muhtemel uygulama alanları şu şekilde sıralanabilir:

1.İnternet Ortamında gittikçe artan dokümanların değerlendirilmesinde
2.Uluslar arası Çalışan şirketlerin müşteri profilini belirlemede
3.Elektronik Ticarette
4.Savunma ve İstihbarat Alanlarında(Güvenlik ve suçlu teşhisi)
5.Yabancı Dil Öğretiminde
6.Makine Çevirisinde
7.Elektronik Sözlüklerde
8.İmla hatalarının otomatik düzeltilmesinde
9.Film ve Sinema Sektöründe[3]
10.Mobil Telefonların Konuşma Algılama Sistemlerinde
11.Otomatik Özet Çıkarma
12.Bilgi Aramada
13.Görme engellilerin bilgisayar kullanmalarında

5.DİL TANIMLAMADA TEMEL YAKLAŞIMLAR

1994 yılında Cavnar, W. B. ve J. M. Trenkle tarafından geliştirilen *N-GRAM* yöntem dünyada bu alanda en sık yöntemlerin başında gelmektedir. MM şeklinde kısaltılarak yazılan *Markov* model genellikle konuşma dilinin ayırt edilmesinde kullanılır, Yazı dilini ayırt eden uygulamaları da mevcuttur. Doğal dil işleme alanında *Ted Dunning, Compression Based Approch (PPM-Teahon)* gibi yöntemlerde kullanılmaktadır.

6-N-GRAM YÖNTEM

N-Gram yöntem verilen metinde tek harfli (mono-gram), çift harfli(b,-gram), tri-gram vb kombinezonları taramaktadır. Yöntemin başlıca savunusu her dilde bazı n-gram'ların diğer dillere oranla frekansının yüksek olduğudur. Az sayıda dile uygulandığında sorunsuz bir şekilde başarı % 100'e yakın olmaktadır. Benzer dillerde ise N-gram tekniğin başarısı oldukça azalmaktadır.Örneğin Algoritma Farsça ve Urducayı ayırt ederken net bir başarı ortaya koyamamaktadır. Asya dillerine uygulanamadığı iddia edilmektedir, ancak Japon ekiplerin bu konudaki çalışmalarına bakıldığında bu sorunun aşılmış olduğu düşünülmektedir.[4]
N-Gram yöntem kullanılarak hazırlanan dil analizörleri genellikle PERL programlama dilinde yazılabilmekte, ancak bu amaçla PHP, DELPHİ gibi programlama dilleri de kullanılabilmektedir.

[3] Konuşmacı dönüştürme sistemi kullanılarak bir filmdeki aktris ve aktörlerin orijinal ses tonuyla başka bir dilde konuşması planlanmaktadır. Örnek olarak orijinal seslendirmesi İngilizce olan bir filmi ele alalım. Bu film başka bir dilde, örneğin Türkçede seslendirileceği zaman, filmdeki metin Türkçeye çevrilip Türkçe konuşan seslendirmeciler tarafından ses kayıtları yapılmaktadır. Bu durumda, doğal olarak filmdeki aktör ya da aktristin ses tonu ve ses özellikleri kaybolmaktadır. Konuşmacı dönüştürme sistemi devreye sokularak aktör ve aktrislerin hiç bilmedikleri bir dilde konuşabilmeleri sağlanabilecektir. İngilizceden http://www.sestek.com.tr/voice_conversion/docs/report_kd_2002.pdf

[4] http://gii.nagaokaut.ac.jp/gii/lopdiary.php?blogid=8

7.DİL TANIMLAMADA DİLBİLİMSEL ÖLÇÜTLER

Bugün Doğal Dil İşleme alanında kullanılan N-Gram uygulamaları, *monogram*, *bigram*, *trigram* gibi birimlerin bir dilbilimci için çok fazla anlamı yoktur. Dilbilimcinin dünyasında daha çok anlam birimler (monemler), sesbirimler (fonemler), önekler, sonekler vb kavramlar yer alır. Hatta bu yüzden dilbilimcilerin pek çoğu daha başlangıçtan itibaren, N-gram, Matematiksel dilbilim, Hesaplamalı dilbilim gibi kavramlarından ürkerler. Metin dili tanımlamada kullanılan bazı dilbilimsel kriterler etkili sonuçlar verebilmektedir.

Muhtemel Dilbilimsel Kriterler

1.Her Dile Özgü Karakterler (İ, Ğ, oe)
2.Frekansı Yüksek Birimler (Bağlaçlar, Tanımlık vb. (The, de, del)
3.Önekler veya son ekler(li, lı, eur, er)
4.Gramer yapısında her dile özgü kombinezonlar(Özne, fiil, nesne sıralaması vb.)
5.Ses yapısındaki özellikler, örneğin Türkçede iki sesli harfin yan yana gelmemesi gibi. Ünlü, ünsüz uyumu vb.
6.Her dile özgü özel isimler
7.Her dilin alfabesini algoritmaya tanıtmak

Bu arada dilbilimsel tekniklere getirilen bazı eleştiriler de dikkati çekmektedir. Bunlar:

1-Ortak kelime ya da frekansı yüksek kelimeler geniş veri tabanlarının test edilmesini gerektiriyor
2-Özel isimler: sayısız özel isim olduğu için bunları referans noktası olarak kabul etmek oldukça güç ve birden fazla dilde bulunma olasılığı yüksek.

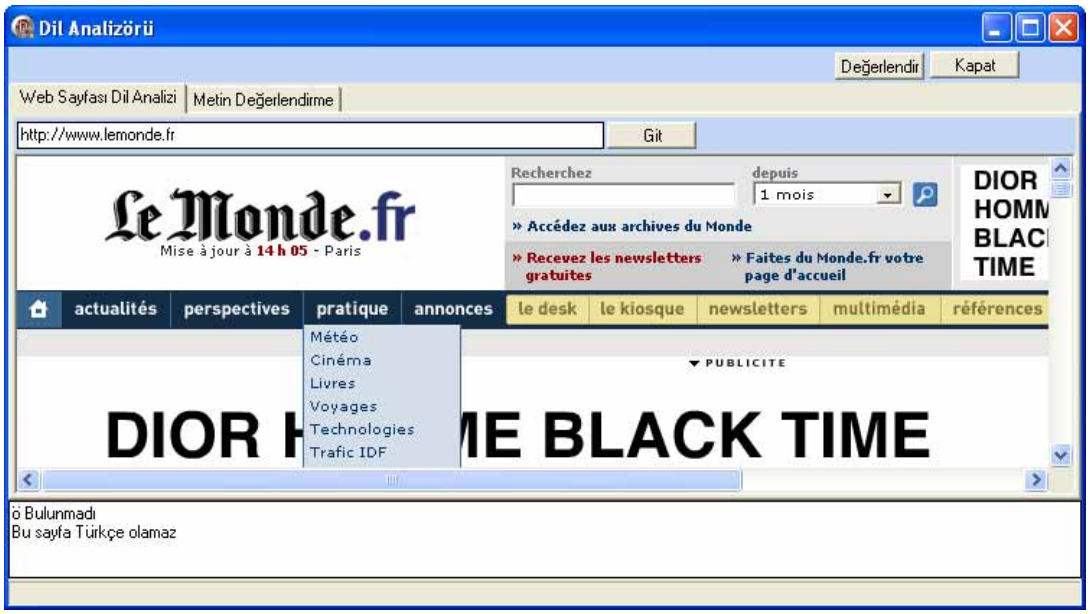
8.ÖRNEK ALGORİTMANIN UYGULANMASI

Günümüzde bilgisayarların dil tanımlama modüllerinde genellikle *N Gram* yöntemler kullanılmaktadır. Bu yöntem bilgisayara tek harfli, (Monogram), çift harfli(Bigram) ve üç harfli (Trigram) birimler verilerek yazılım oluşturulmaktadır.

N-Gram ve diğer bilinen tekniklerin yanında dilbilimsel bazı tekniklerden de yararlanılabileceği varsayılmaktadır. Dilbilimsel ölçütler algoritmaya eklendiğinde bilgisayarların son derece kolay bir şekilde dilleri birbirinden ayırt ettiği gözlenmektedir. Bu ölçütler her dile özgü karakterler, ön ekler, son ekler , her dilde frekansı yüksek olan kelimeler vs.

Bu çalışmada Türkçe kullanıcılar için web sitelerinin dilini ayırt edecek bir yazılım geliştirildi. Diğer dillere de uygulanabilecek olan bu algoritma Delphi'de hazırlandı.

Yazılımın araç çubuğuna web sitesinin adı yazılarak analizöre web sitesinin dili sorulmaktadır. Şu an için sadece Türkçe için geliştirilen dil analizörü web sitesinin Türkçe olup olmadığını veya % kaç ihtimalle sitenin Türkçe karakterler içerdiğini vb. bilgileri kullanıcıya vermektedir. Yapılan ölçümlerde geliştirilen bu analizörün % 100'e yakın oranda Türkçe web sitelerini doğru bir şekilde tahmin ettiği tespit edilmiştir.



Fransız *Le Monde* gazetesinin web sitesinin (www.lemonde.fr) analizördeki görüntüsü yukarda yer almaktadır.

KAYNAKÇA

AYDIN Doğan Emin,(1999) Bilişim ve Telekomünikasyon Terimler Sözlüğü(Telsim) İstanbul

DANLOS Laurence, (2002), Linguistique informatique-traduction automatique, Université de tous les savoirs, France

LAWLER John and DRY Aristar, (1998) Using Computers in Linguistics, Routledge, Newyork,

ÖZBALKON Nuri, (2000) Teknik Terimler Sözlüğü, Alfa Yayınları, İstanbul

RONI AMELAN, (2003) From Information Society to knowledge society, The New Courier, October UNESCO p.32

SANKUR Bülent (2004) İngilizce-Türkçe Ansiklopedik Bilişim Sözlüğü, s.158 Pusula Yayıncılık, İstanbul

SUMY MOUHOUBI, (2005) Languages used on the web, The New Courier, November Unesco,p.62

TAVUKÇUOĞLU Cengiz, (2004) Bilişim Terimleri Sözlüğü, Asil Yayın Dağıtım, Ankara

İnternet Kaynakları

http://www.sestek.com.tr/voice_conversion/docs/report_kd_2002.pdf

http://www.yapay-zeka.org

http://tr.wikipedia.org/wiki/Do%C4%9Fal_Dil_i%C5%9Fleme

http://translate.google.com/translate?hl=en&sl=fr&u=http://www.lri.fr/~heitz/&prev=/search%3Fq%3DLinguistics%2BIngenierie%26hl%3Den%26lr%3D%26sa%3DX (Enginering of Language)

http://www.atilf.fr/atilf/presentation_ang.htm (K)

http://www.ofil.refer.org/tribune/n25/profil.htm (Kavram)

http://web.uni-marburg.de/linguistik//lingengi.html#

http://www.admin.ch/ch/f/bk/sp/indus/indus1.html#1

http://tr.wikipedia.org/wiki/Do%C4%9Fal_Dil_i%C5%9Fleme#Uzman_Sistemler_ve_Do.C4.9Fal_Dil_.C4.B0.C5.9Fleme

http://en.wikipedia.org/wiki/Peter_Toma

http://www.ict4lt.org/en/en_mod3-5.htm#machinetrans

http://www.attiaspace.com/Publications%5CDissertation.pdf

http://en.wikipedia.org/wiki/AI_winter#Early_episodes

http://www.worldlingo.com/ma/enwiki/fr/ALPAC (Alpac raporunun tam metni)

http://www.hutchinsweb.me.uk/PPF-4.pdf