T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ŞARTLI RASTGELE ALANLAR İLE TÜRKÇE WIKIPEDIA SAYFALARINDAN SEMANTİK İLİŞKİLERİN ÇIKARILMASI

Canan GİRGİN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN
DOÇ. DR. BANU DİRİ

İSTANBUL, 2014

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ŞARTLI RASTGELE ALANLAR İLE TÜRKÇE WIKIPEDIA SAYFALARINDAN SEMANTİK İLİŞKİLERİN ÇIKARILMASI

Canan GİRGİN tarafından hazırlanan tez çalışması 18.02.2014 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Aysun GÜRAN

Doğuş Üniversitesi _____

Yrd. Doç. Dr. Mehmet AKTAŞ

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Semantik arama teknolojilerinin altyapılarını oluşturan varlıklar arası ilişkiler ve ontolojilerin kullanımı ve önemi gün geçtikçe artmaktadır. Varlıklar arasındaki ilişkiler yapısal metinlerden kolaylıkla çıkartılabilirken, yapısal olmayan metinlerde bu işlem zorlaşmaktadır. Yapısal olmayan metinlerdeki varlıklar arasındaki ilişkilerin çıkarımında çeşitli "İlişki Çıkarımı" (Relation Extraction) uygulamaları yapılmaktadır.

Bu çalışmada, Türkçe Wikipedia sayfalarından Şartlı Rastgele Alanlar ile varlıklar arasındaki ilişkilerin çıkarımı amaçlanmıştır.

Bu çalışmanın başından sonuna kadar yapmış olduğu öneriler, karşılıklı tartışmalar ve yönlendirmeleri ile tezde yapmış olduğu detaylı incelemeler ve katkıları için hocam Doç. Dr. Banu Diri'ye çok teşekkür ederim.

Hayatım boyunca desteklerini hiçbir zaman eksik etmeyen sevgili anneme, babama ve beni yüreklendiren ve çalışmalarım boyunca bana destek olan eşime ve oğluma çok teşekkür ederim.

Şubat, 2014

Canan GİRGİN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

x	Gözlem dizisi
y	Etiket dizisi
t_j	Geçiş özellik işlevi
s_k	Durum özellik işlevi
λ_i	Çalışma verilerinden tahmin edilebilecek parametre
$Z(x)$	Normalleştirme faktörü
p	Dağılımın beklentisi

KISALTMA LİSTESİ

CRF	Conditional Random Field-Şartlı Rastgele Alanlar
DT	Doğru Tespit
GMM	Gizli Markov Model
GÖBL	GNU Özgür Belgeleme Lisansı
GT	Gerçek Tespit
MEMM	Maksimum Entropi Markov Model
MUC	Message Understanding Conference
MUC-6	Sixth Message Understanding Conference
MUC-7	Seventh Message Understanding Conference
P	Precision-Tutturma
R	Recall-Bulma
TDK	Türk Dil Kurumu
UT	Uygulama Tespiti

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

ŞARTLI RASTGELE ALANLAR İLE TÜRKÇE WIKIPEDIA SAYFALARINDAN SEMANTİK İLİŞKİLERİN ÇIKARILMASI

Canan GİRGİN

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç.Dr. Banu DİRİ

Varlıklar arası ilişkiler semantik arama teknolojilerindeki en önemli yapı taşlarını oluşturmaktadır. Semantik arama teknolojisini kullanan ürünler, altyapılarında varlıklar arasındaki ilişkilerin tutulduğu veri depolarını barındırmaktadırlar. Varlıklar arasındaki ilişkilerin çıkarımında çeşitli "İlişki Çıkarımı" (Relation Extraction) uygulamaları yapılmaktadır. Bu çalışmada, Türkçe Wikipedia sayfalarından varlıklar arasındaki ilişkilerin çıkarımı amaçlanmıştır.

Gerçekleştirilmiş olan çalışma genel hatları ile 4 modülden oluşmaktadır.

1- Pars (Wikipedia Parser)

2- CAT (CRF Automatic Trainer)

3- Köstebek (Relation Extractor)

4- Terazi (Evaluator)

Türkçe Wikipedia sayfalarının tamamının elde edilebilmesi için Wikipedia dump'larından yararlanılmıştır[i]. Wikipedia dump'larının ayrıştırılması ve çalışma

[i] http://dumps.wikimedia.org/trwiki/

esnasında kullanılacak verilerin veri tabanına aktarılması için “Pars” uygulaması gerçekleştirilmiştir.

Makalelerde bulunan varlıklar arası ilişkilerin çıkarımı esnasında Şartlı Rastgele Alanlar (CRF) kullanılmıştır.

Şartlı Rastgele Alanlar altyapısının kullanılabilmesi için etiketlenmiş bir eğitim setine ihtiyaç vardır. Wikipedia sayfalarında metinlerde anlatılan konular ile ilgili özet bilgilerin yapısal olarak kişiler tarafından doldurulduğu bilgi kutusu bölümleri bulunmaktadır. “Pars” uygulaması ile bu veriler ayrıştırılmıştır. Otomatik olarak eğitim setinin oluşturulabilmesi için geliştirilmiş olan “CAT” uygulaması ile bilgi kutularından çıkarılan veriler kullanılarak Şartlı Rastgele Alanlar altyapısı için gerekli olan eğitim seti üretilmiştir.

Wikipedia metinlerinden Şartlı Rastgele Alanlar altyapısı ile ilişkilerin çıkarımı için “Köstebek” uygulaması gerçeklenmiştir. Eğitim setine dâhil edilmemiş Wikipedia verileri üzerinden sistem çalıştırılarak, sistemin çıktılarının doğruluğunu otomatik olarak ölçümleyebilmek için “Terazi” uygulaması gerçeklenmiştir. Bu uygulamada çıktılar ile metinlere ait bilgi kutusunda belirtilen değerler karşılaştırılarak ölçümleme yapılmıştır.

Anahtar Kelimeler: İlişki Çıkarımı, Doğal Dil İşleme, Şartlı Rastgele Alanlar, Wikipedia, Semantik Arama

ABSTRACT

SEMANTIC RELATION EXTRACTION BY CONDITIONAL RANDOM FIELDS FROM TURKISH WIKIPEDIA PAGES

Canan GİRGİN

Department of ComputerEngineering

MSc. Thesis

Advisor: Assoc.Prof.Banu DİRİ

Relations between entities constitute the most important fundamental parts of semantic search technologies. The products that use semantic search technologies include datastores which keep relations between entities in their infrastructures. Various Relation Extraction applications are done in the extraction of the relations between entities. In this work, it is aimed to extract relations between entities from Turkish Wikipedia pages.

The work done in this paper mainly consist of 4 modules.
1- Pars (Wikipedia Parser)
2- CAT (CRF Automatic Trainer)
3- Köstebek (Relation Extractor)
4- Terazi (Evaluator)

Wikipedia dumps are used in order to obtain all Turkish Wikipedia pages. "Pars" application is implemented to parse Wikipedia dumps and transfer the data, which is to be used during the study, to the database. Conditional Random Fields (CRF) is used during the extraction of relations between entities in the article.

A tagged training set is needed for use of Conditional Random Fields infrastructure. Wikipedia pages include information boxes which consist of text summaries filled by human beings constitutionally. This data is indexed by using "Pars" application. By using the “CAT” application, which is developed for creating training sets automatically, data is extracted from these information boxes and the training set, which is required for Conditional Random Fields infrastructure, is produced.

"Köstebek" application is implemented in order to extract the relations from Wikipedia texts by using Conditional Random Fields infrastructure. By operating the system on the Wikipedia data that excluded from training set, "Terazi" application is implemented to evaluate the correctness of system outputs automatically. Basically in this application, the values of the information boxes belong to Wikipedia texts and these outputs are compared and ended up with an evaluation.

Keywords: Relation Extraction, Natural Language Processing, Conditional Random Fields, Wikipedia, Semantic Search

BÖLÜM 1

GİRİŞ

İlişki Çıkarımı olarak Türkçe 'ye çevrilen "Relation Extraction - RE", veri ve doküman madenciliği, doğal dil işleme, bilgi çıkarımı ve bilgiye erişim gibi birçok disiplinle ilişkilidir. İlişki çıkarımının amacı resmi olan veya olmayan bir dilde yazılmış, belli bir çalışma alanına bağlı veya bağımsız olan tüm dokümanlar içerisinden, dile bağımlı veya bağımsız olarak varlıklar arasındaki ilişkileri ve bu ilişki türlerini bulmaktır.

Semantik arama teknolojileri yaygınlaştıkça ilişki çıkarımı yöntemlerine olan ihtiyaç artmıştır. Semantik arama teknolojilerinin istenilen başarımı sağlayabilmesi için altyapılarında varlıklar arasındaki ilişkilerin depolandığı veritabanları bulunmaktadır.

1.1. Literatür Özeti

"Extraction of Relations from Web and Wikipedia KnowItAll" [1] ve"TextRunner"[2] genel web dokümanları içerisinden ilişki çıkarımı yapan uygulamalardır. Catriple [3]uygulaması Wikipedia sayfalarında bulunan kategorilerin analiz edilmesi ve hiyerarşisinin çıkartılması üzerine geliştirilmiştir. Bu uygulama Wikipedia sayfalarında bulunan bilgi kutularındaki bazı bilgilerin çıkarılması için kullanışlı bilgiler sunmaktadır. Ancak uygulamanın odak noktası kategorileri kapsamaktadır.

Ruiz-Casado[4] geliştirdiği eğitim tabanlı sistemde elinde bulunan kelime öbekleri arasındaki ilişkileri bulmak için Wikipedia makalelerini kullanmıştır. Sistem kural tabanlı olup, elle etiketlenen bir eğitim setine ihtiyaç duymaktadır. Wang'da [5]geliştirdiği sistemde kural tabanlı yaklaşımları kullanmıştır ancak, çıkartılan desenler üzerinde seçimli kısıtlar eklemişlerdir. Herbelot ve Copestake [6]cümlelerdeki semantik ilişkileri

basit bir şekilde gösterebilmek ve cümlelerden is-a ontolojisini çıkarabilmek için bir metot önermişlerdir.

Son olarak, Nguyen [7]varlıkları etiketleyerek ve varlıklar arasındaki bağımlılıkları analiz ederek ilişkilerin çıkarımını sağlamıştır. Sistemleri 13 adet ilişki türünü içermektedir.

Yukarıda anlatılan sistemler oluşturulurken Wikipedia metinlerinin ve bilgi kutularının karakteristiği göz önünde bulundurulmamıştır.

Wu ve Weld [8], Wikipedia metinlerinden otomatik olarak bilgi kutusundaki bilgilerin çıkarımını sağlamak amacıyla "Kylin" sistemini geliştirmişlerdir.

iPopulator [9] uygulaması da Kylin ile benzer yapıda geliştirilmiştir. Her ikisi de altyapılarında CRF kullanmışlardır.

Wu ve Weld[8],dört adet özel bilgi kutusu şablonu ile çalışmalarını sürdürmüşlerdir. Kylin 0.74- 0.97 tutturma (precision) ve 0.61-0.96bulma (recall) değeri elde etmiştir.

Daha sonraki çalışmalarda Wu, Kylin'nin ontolojileri ve web arama sonuçlarını da kapsayan yeni bir sürümünü çıkartmıştır.

Gerçekleştirilen sistemler İngilizce Wikipedia metinleri üzerinde geliştirilmiş ve test edilmiştir. Türkçe Wikipedia metinleri üzerinden ilişkilerin çıkartıldığı bir çalışmaya rastlanmamıştır.

1.2. Tezin Amacı

Semantik arama teknolojilerinin altında yatan en önemli yapıtaşlarından biri varlıklar arası ilişkilerdir. Bu ilişkilerin kişiler tarafından elle tanımlanması uzun ve zahmetli süreçler gerektirmektedir.

Mevcut arama motorlarının bazılarında semantik arama teknolojileri kullanılmaktadır. Semantik arama teknolojilerini kullanabilmek için çeşitli kaynaklardan elde edilmiş olan ilişkileri barındıran zengin bir veri tabanı kullanılır. Bu veritabanları çeşitli dillerde yazılmış metinlerden çıkartılmıştır.

Web üzerindeki çeşitli sitelerde bulunan Türkçe metinlerdeki ilişkilerinde çıkartılıp ilişki veritabanlarına eklenmesi ile Türkçe sayfalardaki semantik arama sonuçları daha başarılı sonuçlar sağlayacaktır.

Bu çalışmada, Türkçe Wikipedia metin sayfalarından Şartlı Rastgele Alanlar kullanılarak ilişkilerin çıkartılıp depolanması amaçlanmıştır. Bu doğrultuda Wikipedia sayfalarındaki kişilere ait sayfalar ele alınmıştır. 56 adet kategorideki kişilere ait Wikipedia metinlerinde Şartlı Rastgele Alanlar kullanılarak kişilerin doğum yeri, doğum tarihi, ölüm tarihi bilgilerinin Şartlı Rastgele Alanlar ile çıkartılması gerçeklenmiştir. Bu çalışmada ayrıca, otomatik olarak Wikipedia sayfalarındaki "bilgi kutusu" verileri ile eğitilen bir makine öğrenme yöntemi kullanılmasına odaklanılmıştır.

Yapılan çalışma genel hatları ile dört adımdan oluşmaktadır. İlk olarak, Türkçe Wikipedia sayfalarının dump olarak xml formatında tamamı alınmış ve kullanılabilir veriler şeklinde ayrıştırılarak veri tabanına yazılmıştır.

Çalışmanın ikinci adımında veri tabanındaki bilgilerden kişilere ait olanları tespit edilmiş ve tespit edilen kişilere ait metinlerdeki "bilgi kutusu" verileri analiz edilmiştir. Bilgi kutusu verilerinden daha önce tanımlanmış ise "doğum yeri", "doğum tarihi", "ölüm tarihi" verileri ayrıştırılmıştır. Ayrıştırılan bu veriler kullanılarak CRF uygulamasının eğitim seti otomatik olarak oluşturulmuştur.

Çalışmanın üçüncü adımında, eğitim setini kullanarak ve eğitim setinin yapısı analiz edilerek CRF modeli kurulmuştur.

Çalışmanın son adımında ise otomatik hazırlanan eğitim seti ile eğitilmiş CRF uygulaması ile eğitim setine dâhil edilmemiş olan Wikipedia metinleri ile sistem test edilmiş ve başarım ölçümü yapılmıştır. Elde edilen sonuçlar karşılaştırmalı olarak Bölüm 5'de sunulmuştur.

1.3. Hipotez

İlişki Çıkarımı uygulamalarında varlıklar arasındaki ilişkilerin çıkartılması amaçlanmaktadır. Türkçe için yapılan bu tür çalışmalarda çoğunlukla kural tabanlı yaklaşımlar kullanılmıştır.

Kural tabanlı yaklaşımlarda çıkarım yapabilmek için net bir şekilde örüntü oluşturulması gerekmektedir. Bu durumda resmi olmayan dokümanlarda başarım oranları düşmektedir.

Ayrıca, günümüzde kural tabanlı yaklaşımların çoğu belli bir alana göre hazırlanmış doküman türleri için kullanılmaktadır. Bu şekilde geliştirilen bir sistem başka bir alana ait dokümanlar için uygulanmak istendiğinde, tanımlanan kurallar yetersiz kalabilmektedir.

Bu çalışmada Wikipedia'daki kişilere ait sayfalarda yer alan bilgi kutularındaki eksik bilgilerin otomatik olarak çıkarılması amaçlanmıştır. Başlangıç olarak bilgi kutusunda yer alan doğum yeri, doğum tarihi ve ölüm tarihi bilgilerinin wikipedia sayfalarından çıkarılarak bilgi kutularının doldurulması hedeflenmiştir. Wikipedia metinlerinden ilişkilerin çıkarımı için geliştirilen çalışmada Şartlı Rastgele Alanların kullanılması tercih edilmiştir. Bilgi kutusu içerisinde yer alan varlık isimleri ile otomatik etiketleme yapılarak, Şartlı Rastgele Alanlar için gereken eğitim seti oluşturulmuş ve bu eğitim seti ile sistem eğitilmiştir.

BÖLÜM 2

WIKIPEDIA

2.1. Wikipedia Nedir?

Wikipedia, kullanıcıları tarafından ortaklaşa olarak birçok dilde hazırlanan, özgür, bağımsız bir internet sitesidir. Wiki teknolojisi kullanılarak hazırlanmaktadır.

"Wiki", dilindeki "wiki wiki" (hızlı veya bilgi amaçlı) sözcüğünden,"Pedi" ise Antik Yunan Medeniyetinde "kapsamlı kültürel eğitim sistemi" anlamına gelen paideia kelimesinden gelmektedir[10]. Wikipedia sözcüğü ”wiki” ve ”pedi” kelimelerinin birleşiminden oluşur.

Wikipedia, Nupedia'nın[1] yan kuruluşu olarak kurulmuştur. Nupedia belli bir zaman sonrada kapatılmıştır. Nupediaki maddelerin yazımı oldukça yavaştı.2000yılında yeni bir sistem gereksinimi ortaya çıkmış ve Nupedia'nın ilk wikisi 10 Ocak'ta İnternete konulmuştur.

15 Ocak 2001 tarihinde yeni proje Wikipedia.com sayfasından yayına başlamıştır.

Mayıs 2001'de Wikipedia birçok dilde yazılmaya başlandı ve İngilizce olmayan Wikipedia’larda da yayın hayatına başlandı (Katalanca, Çince, Felemenkçe, Almanca, Esperanto, Fransızca, İbranice, İtalyanca, Japonca, Portekizce …)[11].

[1] http://nupedia.wikia.com

2.2. Bilgilerin Güvenilirliği

Wikipedia da bulunan bilgilerin güvenilirliği ve doğruluğu üzerine tartışmalar mevcuttur ve site yoğun olarak vandalizme maruz kalmaktadır. Ancak2005'te yapılan bir araştırmada, İngilizce Wikipedia'daki[1] "doğal bilgiler üzerine" girdilerin doğruluğu, Britannica Ansiklopedisi[2] ile aynı seviyede bulunmuştur[12].

2.3. İstatistikler

Wikipedia da 100'den fazla dilde[3],3.800.000 madde üzerinde çalışan 48.000 aktif editör[4]vardır. Bugün itibariyle Türkçe Wikipedia' da[5] 220.394 madde bulunmaktadır.

2.4. Wikipedia 'nın Özellikleri

2.4.1. Yazarlar

Wikipedia, gönüllülerin ortaklaşa çabası doğrultusunda ve hemen hemen herkesin web sitesine ulaşıp değiştirmesiyle yazılmaktadır. Wikipedia'ya bilgi girişi yapabilmek için üye olmak zorunluluğu yoktur. Bu durum bilgilerin herkes tarafından anında eklenebilmesine olanak verir. Ancak üye olmanın, yapılan değişikliklerin kaydını tutabilmek, seçkin madde vb. oylamalarına katılmak, vandalizmle mücadeleyi kolaylaştırmak gibi avantajları vardır.

Diğer İnternet ansiklopedi projelerinin çoğu üyelik gerektirir ve yalnızca uzmanların yazılarına izin vermektedir: Stanford Encyclopedia of Philosophy[6], Nupedia[7], h2g2ve Everything vb. gibi. Bazı ansiklopediler de Wikipedia gibi wiki kullanmaktadır. Diğer

[1] http://tr.wikipedia.org/wiki/%C4%B0ngilizce_Vikipedi

[2] http://tr.wikipedia.org/wiki/Britannica_Ansiklopedisi

[3] http://stats.wikimedia.org/EN/Sitemap.htm

[4] http://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm

[5] http://tr.wikipedia.org/wiki/Ana_Sayfa

[6] http://plato.stanford.edu/

[7] http://nupedia.wikia.com

ansiklopedilerden farklı olarak, Wikipedia GNU Özgür Belgeleme Lisansı[1](GÖBL) kullanmaktadır.

2.4.2. Değiştirme ve Düzenleme

Hemen hemen tüm konuklar Wikipedia'nın içeriğini değiştirebilir; yeni kullanıcılar da yeni madde başlatabilirler. Sayfada yapılan değişiklik derhal yansır. Bu ortaklaşa ve anında değişiklikler, editörlerin mevcut maddeleri hızlıca geliştirmesine ve yenilerinin gerçekleşir gerçekleşmez ortaya çıkmasını sağlar.

Maddeler istisnalar dışında her zaman değiştirilmeye açıktır. Eğer ki, vandalizm ya da "revert war" (geri dönüştürme savaşı) yaşanırsa madde koruma altına alınır. Wikipedia hiçbir maddesinin tamamlanmış ya da bitmiş olduğunu iddia etmez. Maddeler bir kullanıcı ya da editör grubu tarafından kontrol edilmez ya da telif altına alınmaz; içerik hakkındaki kararlar Wikipedia'nın editör kurallarına göre fikir alma ortak kararıyla ya da bazen oyla karar verilir.

2.4.3. Özgür İçerik

Wikipedia'da ve kardeş projelerinde, bütün katkıların yasalara uygun olarak yapılması önemlidir. Wikipedia'daki maddeler GÖBL[2] ile yazılmakta ve her yazar telif haklarından vazgeçmektedir. GÖBL, birçok telifi özgür bırakılmış lisans gibi, yazılanın dağıtılmasını, benzeri dokümanların çıkmasını ve kâr amaçlı kullanıma izin verir. Bu lisans ayrıca yeniden katılım gösterenlerin aynı lisans ile değişiklikler yapmasına izin verir.

2.4.4. Türkçe Wikipedia Nedir?

Wikipedia projesinin, 2003'te hayata geçen Türkçe koludur. Türkçe Wikipedia, diğer tüm dillerdeki Wikipedia'larla birlikte, kâr amacı gütmeyen Wikipedia Vakfı'nın[3] tescilli markasıdır.

---

[1] http://tr.wikipedia.org/wiki/GNU_%C3%96zg%C3%BCr_Belgeleme_Lisans%C4%B1

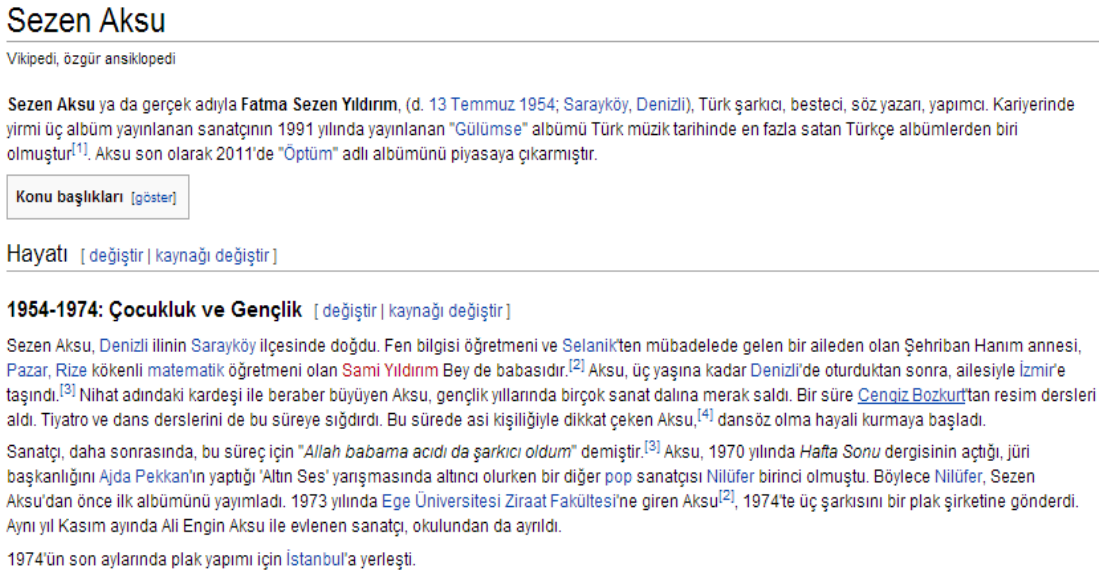
[2] http://tr.wikipedia.org/wiki/GNU_%C3%96zg%C3%BCr_Belgeleme_Lisans%C4%B1

[3] http://wikimediafoundation.org

2.4.5. Wikipedia Bilgi Kutuları

Wikipedia web sayfalarının genel şablonunda maddelere ait metinler ve bu metinlerin özet bilgilerinin yapısal olarak çıkartıldığı “bilgi kutusu” bölümü bulunmaktadır.

Bilgi kutusundaki yapısal veriler yazarlar tarafından çıkartılmaktadır. Sezen Aksu’ ya ait Wikipedia sayfası Şekil 2.1‘de ve bu sayfaya ait yapısal olarak çıkartılmış olan bilgilerin bulunduğu bilgi kutusu Şekil 2.2’de örneklenmiştir.



Sezen Aksu

Vikipedi, özgür ansiklopedi

Sezen Aksu ya da gerçek adıyla **Fatma Sezen Yıldırım**, (d. 13 Temmuz 1954; Sarayköy, Denizli), Türk şarkıcı, besteci, söz yazarı, yapımcı. Kariyerinde yirmi üç albüm yayınlanan sanatçının 1991 yılında yayınlanan "Gülümse" albümü Türk müzik tarihinde en fazla satan Türkçe albümlerden biri olmuştur[1]. Aksu son olarak 2011'de "Öptüm" adlı albümünü piyasaya çıkarmıştır.

Konu başlıkları [göster]

Hayatı [değiştir | kaynağı değiştir]

1954-1974: Çocukluk ve Gençlik [değiştir | kaynağı değiştir]

Sezen Aksu, Denizli ilinin Sarayköy ilçesinde doğdu. Fen bilgisi öğretmeni ve Selanik'ten mübadelede gelen bir aileden olan Şehriban Hanım annesi, Pazar, Rize kökenli matematik öğretmeni olan Sami Yıldırım Bey de babasıdır.[2] Aksu, üç yaşına kadar Denizli'de oturduktan sonra, ailesiyle İzmir'e taşındı.[3] Nihat adındaki kardeşi ile beraber büyüyen Aksu, gençlik yıllarında birçok sanat dalına merak saldı. Bir süre Cengiz Bozkurt'tan resim dersleri aldı. Tiyatro ve dans derslerini de bu süreye sığdırdı. Bu sürede asi kişiliğiyle dikkat çeken Aksu,[4] dansöz olma hayali kurmaya başladı.

Sanatçı, daha sonrasında, bu süreç için "*Allah babama acıdı da şarkıcı oldum*" demiştir.[3] Aksu, 1970 yılında *Hafta Sonu* dergisinin açtığı, jüri başkanlığını Ajda Pekkan'ın yaptığı 'Altın Ses' yarışmasında altıncı olurken bir diğer pop sanatçısı Nilüfer birinci olmuştu. Böylece Nilüfer, Sezen Aksu'dan önce ilk albümünü yayımladı. 1973 yılında Ege Üniversitesi Ziraat Fakültesi'ne giren Aksu[2], 1974'te üç şarkısını bir plak şirketine gönderdi. Aynı yıl Kasım ayında Ali Engin Aksu ile evlenen sanatçı, okulundan da ayrıldı.

1974'ün son aylarında plak yapımı için İstanbul'a yerleşti.

Şekil 2.1 Sezen Aksu Wikipedia sayfası

Şekil 2.2 Sezen Aksu bilgi kutusu

Wikipedia sayfalarında bulunan bilgi kutularındaki veriler belirli şablonlara uygun olarak doldurulmaktadır. Bu şablonlar ihtiyaca göre tanımlanmıştır. Şablonlara yenileri eklenebildiği gibi gerektiği zaman var olan şablonlar üzerinde güncelleme, ekleme ve çıkartma yapılabilmektedir. Türkçe Wikipedia maddeleri için tanımlanmış olan tüm bilgi kutusu şablonlarına Wikipedia sayfasından[1] erişilebilmektedir. Bir müzik sanatçısına ait Wikipedia sayfası için doldurulması gereken müzik sanatçısı bilgi kutusu şablonu Şekil 2.3' de örneklenmiştir.

[1] http://tr.wikipedia.org/wiki/Kategori:Bilgi_kutusu_%C5%9Fablonlar%C4%B1

```
{{Müzik sanatçısı bilgi kutusu
| ad                  =
| resim               =
| resim boyutu        =
| genişlet            =
| altyazı             =
| artalan             = <!-- **Zorunlu** Birini seçin: tek şarkıcı, vokalsiz çalgıcı, çalmayan görevli, geçici -->
| doğum adı           =
| takma adı           =
| doğum tarihi        = <!-- {{Doğum tarihi ve yaşı|YIL|AY|GÜN}} -->
| yer                 =
| köken               =
| ölüm tarihi         = <!-- {{Ölüm tarihi ve yaşı|YIL|AY|GÜN|YIL|AY|GÜN}} (ilk olarak ölüm tarihi yazın) -->
| ölüm yeri           =
| tarz                =
| meslek              =
| çalgı               =
| etkin yıllar        = <!-- YYYY-YYYY (veya: -günümüz) -->
| plak şirketi        =
| ilişkili hareketler =
| web sitesi          = <!-- {{URL|örnek.com}} -->
| önemli çalgılar     =
}}
```

Şekil 2.3 Şablon: Müzik sanatçısı bilgi kutusu şablonu

Bilgi Kutusu şablonları tanımlanırken iç içe şablonlarda tanımlanabilmektedir. Bu durumda hiyerarşik bir yapı izlenmektedir. Örn: Wikipedia'da "Kişi Bilgi Kutusu Şablonları" adı altında bir kategori tanımlanmıştır. Bu kategori altında "Kurgusal karakter bilgi kutusu şablonları"[1]alt kategorisi ve ek olarak 56 adet şablon tanımlanmıştır. Hiyerarşinin son elemanı olan "Kurgusal karakter bilgi kutusu şablonları" kategorisinde ise 22 adet şablon tanımlanmıştır. Tanımlı olan en alt seviyedeki şablonlar Şekil 2.4'de görülmektedir.

[1] http://tr.wikipedia.org/wiki/Kategori:Kurgusal_karakter_bilgi_kutusu_%C5%9Fablonlar%C4%B1

Kategori:Kurgusal karakter bilgi kutusu şablonları

Vikipedi, özgür ansiklopedi

Kategorideki sayfalar

Bu kategoride toplam 22 sayfa bulunmaktadır ve şu anda bunların 22 tanesi görülmektedir.

A
- Şablon:Animanga/karakter
- Şablon:Avrupa Yakası karakteri bilgi kutusu

B
- Şablon:Bleach karakteri bilgi kutusu
- Şablon:Buffyevreni karakter bilgi kutusu

D
- Şablon:Digimon Türleri

H
- Şablon:Harry Potter karakteri bilgi kutusu

K
- Şablon:Karakter Bilgi kutusu
- Şablon:Karakter bilgi kutusu

K (devam)
- Şablon:Kurgusal karakter bilgi kutusu

L
- Şablon:Lost karakteri bilgi kutusu

O
- Şablon:Orta Dünya karakteri bilgi kutusu

P
- Şablon:Parmaklıklar Ardında karakteri bilgi kutusu
- Şablon:Pokémon türü
- Şablon:Prison Break karakteri bilgikutusu

S
- Şablon:Simpsonlar karakteri
- Şablon:Sonic anime

U
- Şablon:Uzay Yolu karakteri bilgi kutusu

X
- Şablon:X Files karakteri

Y
- Şablon:Yalan Dünya karakteri bilgi kutusu
- Şablon:Yıldız Geçidi ırkları bilgi kutusu
- Şablon:Yıldız Geçidi karakteri
- Şablon:Yıldız Savaşları karakteri bilgi kutusu

Kategori: Kişi bilgi kutusu şablonları

Şekil 2.4 Kategori: Kurgusal karakter bilgi kutusu şablonları

Wikipedia maddelerindeki bilgi kutusu bölümleri yazarlarca doldurulurken herhangi bir kısıt bulunmamaktadır. Bu durumun sonucu olarak sayfalarda bulunan bilgi kutularındaki bilgiler bazen şablonlara uygun olarak doldurulmamaktadır.

Wikipedia sayfalarındaki bilgi kutusunda bulunan veriler yazarlarca doğru bilgiler ile doldurulmaya çalışılmaktadır. Ancak, bu bilgiler içinde yanlış olarak doldurulmuş verilere de rastlanabilmektedir.

Tüm bunlara ek olarak Wikipedia maddelerinde bulunan bilgi kutularındaki yapısal olarak eklenmiş veriler bazen maddenin geniş anlatımında bulunmayabilmektedir.

BÖLÜM 3

ŞARTLI RASTGELE ALANLARLA VARLIK İSMİ TANIMA

Doğal dil işlemede bilgi çıkarımı yöntemleri kullanılarak bir doküman içerisindeki belirli kriterlere uyan bilgilere erişim yapılabilmektedir. Bu işlem sırasında örneğin, bir kalıba uygun olan verilerin çıkartılması istenebilir. Bilgi çıkarım işleminde amaç çok miktardaki veriyi otomatik olarak işleyen bir yazılım üreterek, insan müdahalesini minimum seviyeye indirmektir. Bilgi çıkarımın bir alt dalı olan İlişki Çıkarımında da bir veri kaynağında geçen varlıklar arasındaki ilişkilerin çıkartılması hedeflenir. Bu ilişkiler, meslek, doğum yeri, eğitim durumu gibi bilgiler olabilir.

İlişki çıkarımında kullanılan çeşitli yöntemler olmasına karşın günümüzde makine öğrenmesi teknikleri kullanılarak yeni sistemler geliştirilebilmektedir. Eğitilen sistemlerden Şartlı Rastgele Alanlar kullanılarak Wikipedia metinlerindeki ilişkiler belirlenebilmektedir.

3.1 Sıralı Verinin Etiketlenmesi

Bir dizi gözlem dizilerine etiket sıralarının atanması biyobilişim, bilişimsel dilbilim ve konuşma tanıma [13] [14] [15] dâhil olmak üzere birçok alanda yapılmaktadır. Doğal dil işlemede sözdizimsel analiz ile sıralı veri etiketlenebilmektedir. Örneğin;

<İsim>Ahmet</İsim><TamlayanEki>in</TamlayanEki><İsim>baba</İsim><TamlananEki>sı</TamlananEki>,<Sıfat>beyaz</Sıfat><İsim>koyun</İsim><NesneEki>u</NesneEki><İsim>araba</İsim><DiğerZarflar>ile</DiğerZarflar><İsim>köy</İsim><DolaylıTümleçEki>e</DolaylıTümleçEki><Yüklem>getir</Yüklem><ZamanEki>di</ZamanEki>

Cümlelerin bu şekilde etiketlenmesi, daha üst düzey doğal dil işleme görevleri için faydalı bir işleme adımıdır: Sözdizimsel analiz, açık olarak dilin özünde bulunan bazı yapılara sadece işaret etmek suretiyle sözcüklerin içerdiği bilgileri arttırır [16].

Böyle bir etiketleme ve parçalara ayırma görevlerinin gerçekleştirilmesi için kullanılan en yaygın yöntemlerden birisi, herhangi bir cümledeki sözcükler için en olası etiket dizisini tanımlamak amacıyla, Gizli Markov Modelleri (GMM) [17]ya da olasılıksal sonlu özdevinirin kullanılmasıdır.

Gizli Markov Modeli doğal dil işleme, ses tanıma, video işleme ve bunun gibi zamana bağlı değişkenlerin olduğu alanlarda kullanılan bağlantı tabanlı bir modeldir. GMM yapısı itibariyle Markov Model teorisine dayanmaktadır. Markov Model yapısı ardışık olarak gelen düğümlerin (çizge modeli) olasılığının bulunmasında kullanılır. Markov Model yapısında düğümler sadece gözlemlenen varlıkları ifade eder. Ancak, GMM gözlemleyemediğimiz ve varsaydığımız bağlantılar için saklı düğümler oluşturarak bağlantıları bizim varsayımımıza göre yapar.

Markov Model ile bir denklemin matematiksel bir modeli çıkarılırken, GMM ile bu denkleme yakınsayan bir model çıkarılır [18].GMM iki stokastik süreç içerir. İlk olan Markov süreci, zaman ile ilgili değişikliklerde kullanılır ve durumları içeren bir Markov zinciri üretir. Diğer süreç gözlemlenebilir olan özellik parametrelerini veya gözlemler denilen rastgele değişkenleri içerir [19].

GMM'ler, gözlem dizileri ve bunlarla ilişkili etiket sıraları boyunca sırasıyla yayılan$\boldsymbol{X}$ve $\boldsymbol{Y}$'nin rastlantısal değişkenler olduğu bir $p(\boldsymbol{X}, \boldsymbol{Y})$ortak olasılık dağılımını tanımlayan üretken bir model biçimindedir. Bu yapıdaki bir ortak dağılımı tanımlamak için üretken modellerin tüm olası gözlem öğelerini sıralaması gerekir ve bu görev, bir gözlem dizisinde diğer elemanlardan bağımsız olarak gözlem elemanları izole edilmiş birimler şeklinde temsil edilmedikçe birçok alan için zorlu bir görevdir. Zaman içinde herhangi bir andaki gözlem elemanı sadece o zamandaki bir duruma ya da etikete doğrudan bağlı olabilir. Bu, birkaç basit veri dizileri için uygun bir varsayımdır [20].

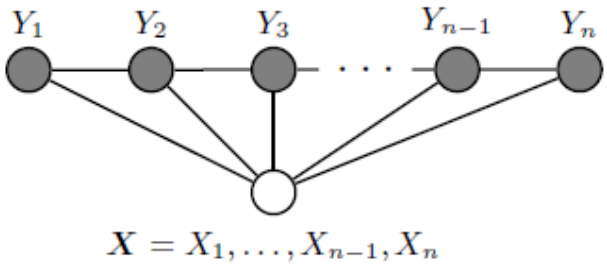
Bu temsil sorunu sıralı verilerin etiketlenmesinde en temel problemlerden birisidir. İzlenebilir çıkarımı destekleyen bir model gereklidir ama nedensiz bağımsız varsayımlar yapmadan verileri temsil eden bir model de arzu edilmektedir. Bu her iki kriteri yerine

getiren bir yol, hem etiket hem de gözlem dizileri boyunca bir ortak dağılım yerine, belirli bir x gözlem dizisinde etiket dizileri boyunca $p(\boldsymbol{Y}|\boldsymbol{x})$ şartlı olasılığını tanımlayan bir modelin kullanılmasıdır. Şartlı modeller, $p(\boldsymbol{y}_*|\boldsymbol{x}_*)$ şartlı olasılığını arttıran $\boldsymbol{y}_*$ etiket dizini seçmek suretiyle $\boldsymbol{x}_*$yeni gözlem dizini etiketlemek için kullanılır. Bu tür modellerin koşullu doğası, gözlemleri modellemede herhangi bir çaba harcanmadığı ve bu diziler hakkında nedensiz bağımsız varsayımlar yapmak zorunda olmadığı anlamına gelir; modelleyicinin bu özelliklerin nasıl ilişkili olduğu hakkında endişelenmesine gerek kalmadan, gözlem verisinin rastgele özellikleri bu modelle elde edilebilir.

Şartlı Rastgele Alanlar [20]sıralı verilerin etiketlenmesi ve parçalara ayrılması için, koşullu yaklaşıma dayalı bir olasılıksal çerçevedir. Şartlı Rastgele Alanlar, belirli bir gözlem dizisindeki etiket dizileri boyunca bir tek log-lineer dağılımı tanımlayan yönsüz bir grafik modeli biçimidir. Şartlı Rastgele Alanların Gizli Markov Modele göre temel avantajı, GMM'lerin gerektirdiği bağımsız varsayımlardan rahatlamaya neden olacak şekilde şartlı yapıda, olmalarıdır. Buna ek olarak, Şartlı Rastgele Alanlar, Maksimum Entropi Markov Modelleri [14] (MEMM'ler) ve yönlendirilmiş grafik modellerine dayalı diğer şartlı Markov modelleri tarafından sergilenen bir zayıflık olan etiket önyargı problemini önlerler. Şartlı Rastgele Alanlar çok sayıda dizi etiketleme görevlerinde hem MEMM'ler hem de GMM'lerden üstündür [21][22][23].

3.2 Yönsüz Grafik Modelleri

Şartlı Rastgele Alan, gözlem dizilerini temsil eden rastgele değişken olan $\boldsymbol{X}$üzerinde global olarak devam eden yönsüz bir grafik model ya da Markov Rastgele Alan [24] olarak görülebilir. Formal olarak, Y'nin bir Y_v elemanını temsil eden her bir rastgele değişkene denk düşen bir $v \in V$ düğümü olacak şekilde bir yönsüz grafik olarak $G= (V, E)$ şeklinde tanımlanır. Şayet, her bir Y_v rastgele değişkeni, *G* yönünden Markov özelliğine uyarsa, bu durumda $(\boldsymbol{Y}, \boldsymbol{X})$ bir şartlı rastgele alandır. Teoride, modellenmekte olan etiket dizilerinde koşullu bağımsızlıkları temsil etmesi şartıyla, *G* grafiğinin yapısı rastgele olabilir. Ancak, diziler modellenirken karşılaşılan en basit ve en yaygın grafik yapısı, Şekil 3.1'de gösterildiği gibi, $\boldsymbol{Y}$ biçimi elemanlarına denk düşen düğümlerin basit birinci derece zincir oluşturmasıdır. Şekil 3.1'de görüldüğü gibi $\boldsymbol{Y}$, basit birinci dereceden zincir oluşturur[16].



Şekil 3.1 Diziler için zincir yapılı Şart Rastgele Alanların rastgele yapısı

Gölgesiz düğümlere denk düşen değişkenler model tarafından üretilmemiştir.

3.3 Potansiyel Fonksiyonlar

Şartlı Rastgele Alanın grafik yapısı, $\boldsymbol{Y}$'nin Y_v elemanları boyunca ortak dağılımını, koşullu bağımsızlık kavramından türetilmiş pozitif gerçek değerli potansiyel işlevlerin normalleştirilmiş bileşkesini çarpanlara ayırmak için kullanılabilir. Her bir potansiyel işlev, G'deki köşe noktaları tarafından temsil edilen rastgele değişkenler alt dizisinde işlev görür. Yönsüz grafik modelleri için şartlı bağımsızlık tanımına göre, G'deki iki köşe noktası arasında bir köşenin yokluğu, bu köşe noktaları tarafından temsil edilen rastgele değişkenlerin bu modeldeki tüm rastgele değişkenlerde şartlı olarak bağımsız olduğunu işaret eder. Bu nedenle potansiyel işlevlerin, şartlı bağımsız değişkenlerin aynı potansiyel işlevde görünmeyecek şekilde, ortak olasılığı çarpanlara ayırmanın mümkün olduğunu sağlaması gerekir. Bu gerekliliği yerine getirmenin en kolay yolu, her bir potansiyel işlevin ilişkili köşe noktalarının G içinde maksimal grup oluşturan rastgele değişkenler dizisinde işlev görmesini gerektirmektir. Bu, hiçbir potansiyel işlevin, köşe noktaları doğrudan bağlantılı olmayan herhangi bir rastgele değişkenler çiftine işaret etmemesini ve şayet iki köşe noktası bir grupta görünürse, bu ilişkinin açık olmasını sağlar. Şekil 3.1'de sergilendiği gibi zincir yapılı Şartlı Rastgele Alanlar durumunda ise her bir potansiyel işlev, Y_i ve Y_{i+1} komşu etiket değişkenlerinde işlev görmektedir.

İzole edilmiş potansiyel bir işlevin doğrudan olasılıksal yorumu olmayıp, bunun yerine tanımlanan işlevdeki rastgele değişkenler konfigürasyonunda sınırlamaları temsil etmektedir. Bu da sonuçta global konfigürasyonların olasılığını etkiler; yani yüksek

olasılığa sahip bir global konfigürasyon, düşük olasılığa sahip bir global konfigürasyona göre bu sınırlamalardan bir çoğunu yerine getirmesi olasıdır.

3.4 Şartlı Rastgele Alanlar

Lafferty ve diğerleri [25] $x$gözlem dizisindeki belirli bir y etiket dizisinin, her birinin Eşitlik 3.1 'de oluşturulduğu gibi potansiyel işlevlerin normalleştirilmiş bileşkesi olacağını tanımlar:

$$\exp\left(\sum_j \lambda_j t_j\ (y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \quad (3.1)$$

Burada $t_j(y_{i-1}, y_i, x, i)$ tüm gözlem dizisinin ve etiket dizisindeki $i$ve$i$— 1 konumlarındaki etiketlerin geçiş özellik işlevidir; $s_k(y_i, x, i)$ ise i konumundaki etiket ve gözlem dizisinin durum özellik işlevidir; λ_i ve μ_k çalışma verilerinden tahmin edilecek parametrelerdir.

Özellik işlevlerini tanımlarken, model dağılımını tutması gereken eğitim verilerinin ampirik dağılımının bazı karakteristiklerini ifade etmek için, gözlemin gerçek değerli özellikler $b(x, i)$ dizisi oluşturulur. Böyle bir özelliğin örneği şu şekildedir:

$$b(x, i) = \begin{matrix} 1 & i \text{ konumundaki gözlem "İstanbul" kelimesi ise} \\ 0 & \text{Diğer durum} \end{matrix}$$

Şayet mevcut durum (durum işlevi durumunda) ya da önceki ve tüm durumlar (geçiş işlevi durumunda) belirli değerleri alırsa, her bir özellik işlevi, bu gerçek değerli gözlem özelliklerinden $b(x, i)$'den birisinin değerini alır. Bu nedenle tüm özellik işlevleri gerçek değerlidir. Örneğin, aşağıdaki geçiş işlevini düşünün:

$$t_i(y_{i-1}, y_i, x, i) = \begin{cases} b(x, i) \text{ Eğer} y_{i-1} \text{sıfat ise}, y_i \text{isimdir} \\ 0 \qquad\qquad\qquad\qquad \text{Diğer durum} \end{cases}$$

Ayrıca, bu notasyon aşağıdaki gibi yazılarak basitleştirilmiştir:

$$s(y_i, x, i) = s(y_{i-1}, y_i, x, i)$$

ve

$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$$

Burada her bir $f_j(y_{i-1}, y_i, x, i)$'ye ya bir $s(y_{i-1}, y_i, x, i)$durum işlevidir ya da bir $t(y_{i-1}, y_i, x, i)$ geçiş işlevidir [26].Bu durum ise bir x gözlem dizisindeki y etiket dizisi olasılığının Eşitlik 3.2'deki gibi yazılmasına izin verir:

$$p(y|x, \lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(\boldsymbol{y}, \boldsymbol{x})) \quad (3.2)$$

$Z(x)$ Bir normalleştirme faktörüdür.

3.5 Maksimum Entropi

Şartlı Rastgele Alanlar örüntü tanımada kullanılan istatistiksel modellemelerden biridir. Şartlı Rastgele Alanlar sıralı veriyi işaretlemek ve bölümlere ayırmak için kullanılan Maksimum Entropi Markov Model ve Gizli Markov Modelin genel halini yansıtan bir olasılık sunar.

Şartlı Rastgele Alanlarda ağırlıklı olarak Maksimum Entropi ilkesi kullanılarak, eğitim veri setinden tahmini olasılık dağılımı hesaplanmaktadır. Olasılık dağılım entropisi bir belirsizlik ölçüsüdür ve söz konusu dağılım mümkün olduğu kadar homojen olduğunda maksimize edilir. Maksimum Entropi prensibi sonlu eğitim verileri gibi yetersiz bilgi üzerinden elde edilen olasılık dağılımında bile mevcut kısıtlamaları temsil eden bir dizi maksimum entropi olduğunu belirtmektedir. Başka bir dağılım yersiz varsayımlar içerecektir [27].

Eğitim veri seti içerisinde bulunan bilgiler bir dizi özellik fonksiyonu ile sunulmaktadır. Maksimum Entropi Model dağılımı eğitim verisinde her bir özellik fonksiyonunun sunduğu ampirik dağılımın mümkün olduğunca homojen olmasını beklemektedir[28][29].

3.6 Şartlı Rastgele Alan Kütüphaneleri

Şartlı Rastgele Alanların kullanımı için çeşitli kütüphaneler ve araçlar hali hazırda geliştirilmiştir. Bunlar arasında CRFsuite, CRF ++, Crfsgd, Flexcrf, Mallet gibi kütüphaneler bulunmaktadır. Çalışma kapsamında yukarıda ismi sayılan kütüphaneler incelenmiştir.

CRF ++ kütüphanesi 2005 yılında geliştirilmiştir. C++ tabanlıdır. Dosyalar ile çalışmaktadır. Analiz yapılmak istenilen metinler dosyalar ile sisteme verilmelidir. Şablonlar kullanılabilmektedir. Bu işlem için şablon dosyaları oluşturulmaktadır. Paralel çoklu iş parçacıkları ile çalışabilmeyi desteklemekte ve yaygın olarak kullanılmaktadır.

Crfsgd 2006 yılında C++ programlama dili ile geliştirilmiştir. Özellik şablonları kullanılabilmektedir.

Flexcrf 2005 yılında çoklu iş parçacıkları ile paralel olarak çalışabilmekte ve C++ programlama dili ile geliştirilmiştir.

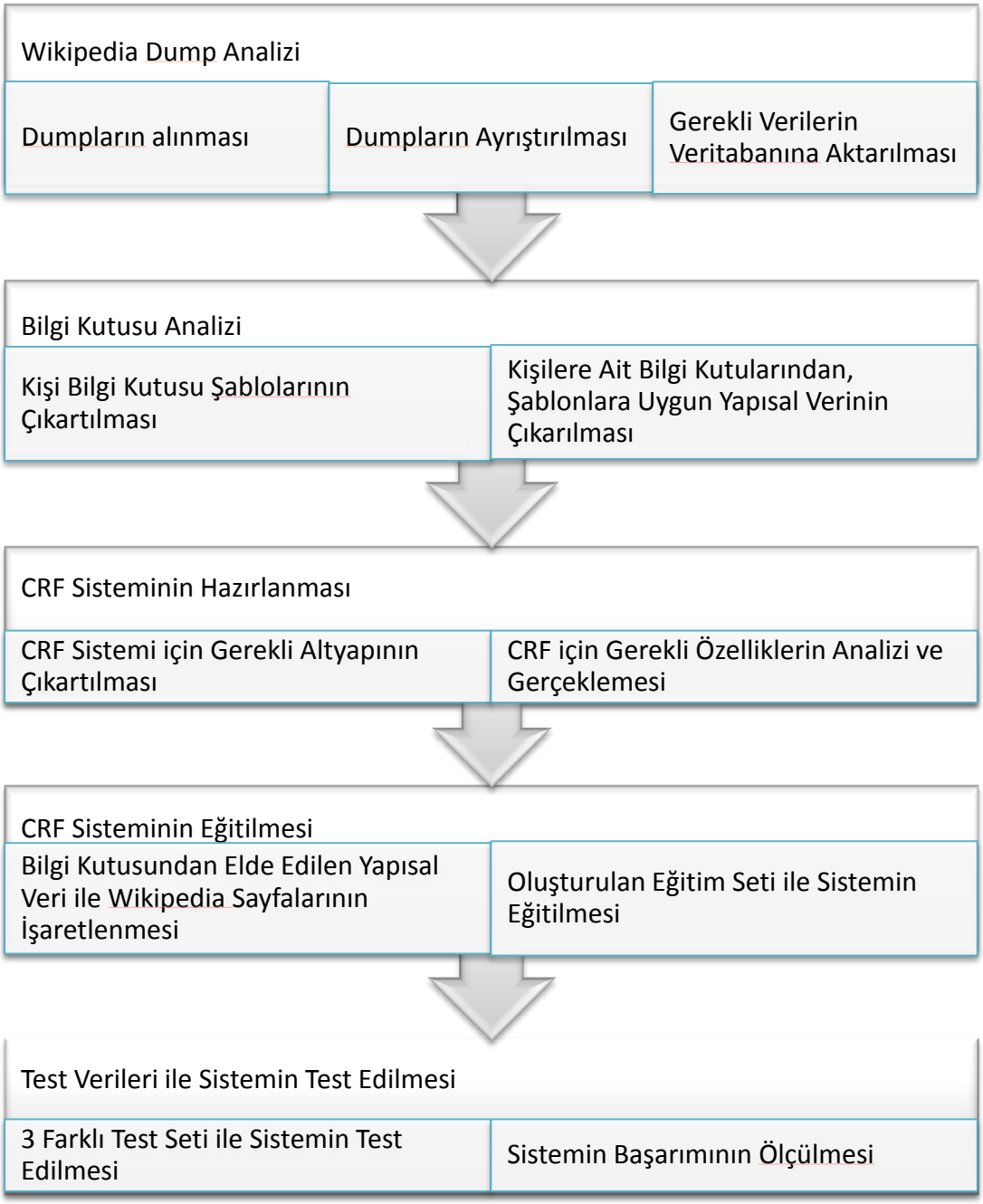
Crfsuite 2007 yılında geliştirilmiştir. Hızlı olması içi C programlama dili ile yazılmıştır.

Mallet Java ile yazılmış olup, Şartlı Rastgele Alanların yanı sıra birçok NLP işlemi için kullanılabilmektedir. Mevcut sistemler arasında en iyi sonuç döndüren kütüphanedir[30]. Şablon dosyaları hazırlamaya gerek yoktur. API olarak kullanılabilmektedir. Çoklu iş parçacıkları ile paralel çalışabilmeyi desteklemektedir. Kullanılabilir fonksiyon sayısı da oldukça fazladır.

BÖLÜM 4

GELİŞTİRİLEN SİSTEM

Tasarlanan sistemde semantik aramalar için gereken ilişkilerin yapısal olmayan metinlerden çıkartılabilmesi amaçlanmıştır. Bu amaca yönelik olarak Wikipedia sayfalarında yer alan ,“Wikipedia Bilgi Kutuları” bölümün detaylı olarak açıklanan, dolu bilgi kutusu bilgileri çıkartılmıştır. Bilgi kutusu dolu olmayan kişilerin ise “doğum yeri”, “doğum tarihi” ve “ölüm yeri” bilgilerinin, otomatik eğitilen CRF sistemi ile yapısal olmayan Wikipedia sayfalarından çıkartılması üzerine odaklanılmıştır. Geliştirilen sistemin tasarımı Şekil 4.1’de gösterilmektedir.



Şekil 4.1 Sistem tasarımı

4.1 Wikipedia Dump Analizi

Web ortamındaki bazı dillerdeki Wikipedia sayfalarının tamamı, belirli periyotlarla xml formatında sıkıştırılmış olarak paylaşılmaktadır. Bunlar Wikipedia dump olarak nitelenmektedir.

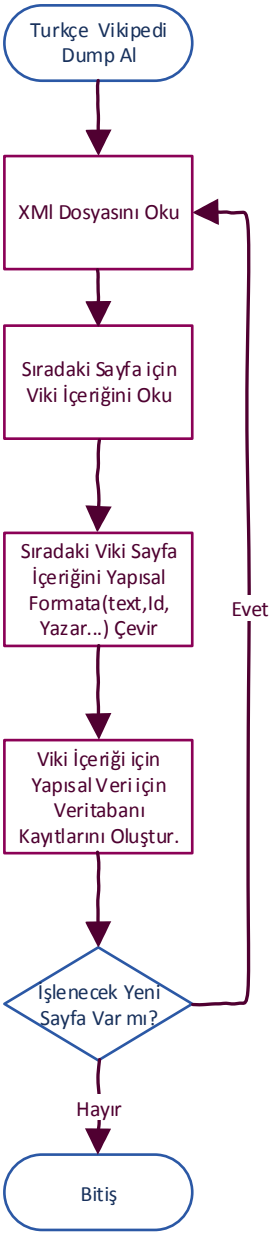
Semantik aramalar için gerekli bilgi çıkarımı yapabilmek için geliştirilen sistemde Wikipedia'da ki maddeler kullanılmıştır. Bu amaca yönelik olarak öncelikle Türkçe

Wikipedia sayfalarının en son yayınlanmış olan, içerisinde metinlerin, bilgi kutusunun ve kategorilerin bulunduğu dumplar web ortamından indirilmiştir.

Türkçe Wikipedia'da çeşitli konularda yazılmış olan 599.048 adet madde bulunmaktadır. İndirilen dump xml formatında ki dosya bu verileri içermektedir. Sistemin çeşitli aşamalarında sayfaların farklı verilerine ihtiyaç duyulacağından, her seferinde ayrıştırma işlemine gerek olmaması için bu veriler yapısal bir formatta ayrıştırılıp, veritabanına yazılmıştır.

Bu aşamadaki işlem adımları Şekil 4.2'de gösterilmiştir.

Bu işlem için java uygulaması geliştirilmiş ve Mysql veri tabanı kullanılmıştır.



Şekil 4.2 Dump analizi işlem adımları

4.1.1 Sayfalarda Tanımlı Bilgi Kutusu Verilerinin Ayrıştırılması

Geliştirilen sistemde CRF için gerekli eğitim verisi otomatik olarak üretilmiştir. Otomatik etiketleme işlemi yapılırken, Wikipedia sayfalarında yazarlarca tanımlanmış olan bilgi kutusu verileri alınarak bu veriler ilgili sayfanın içeriğinde bulunarak etiketleme yapılmıştır.

Wikipedia dump içerisindeki xml dosyasında Wikipedia metinleri ve bu metinlere ait yazar adı, değişiklik tarihi gibi alanlar ayrı ayrı tutulmuştur. Ancak, metinlere ait bilgi kutuları ayrı bir alan olarak değil metin içeriği ile birlikte bulunmaktadır.

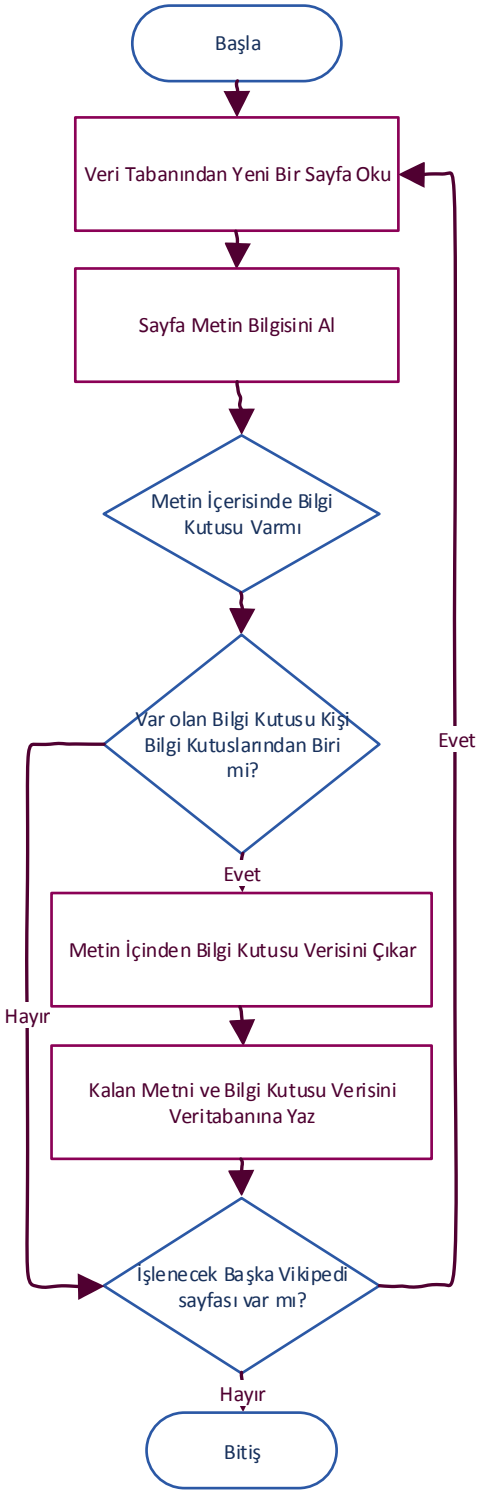
Sistemin eğitim verisinin oluşturabilmesi için her sayfa için tanımlı bilgi kutusu kullanılacağından bu veri metin verisi içerisinden ayrıştırılmıştır. Geliştirilen sistemde kişilere ait bilgilerin içerik çıkarımı hedeflendiğinden bu tez çalışmasında kişilere ait bilgi kutusu şablonlarına odaklanılmıştır. Kişi bilgi kutularının neler olduğu Wikipedia sayfasından [1] alınmıştır.

Toplamda 57 adet kişi bilgi kutusu bulunmaktadır. Üzerinde çalışılan bilgi Kutusu Şablonları EK A 'da listelenmiştir. İşlem Adımları Şekil 4.3'de gösterilmiştir.

Sayfalar içerisinden bilgi kutularının çıkartılması işleminin yapılmasında, Wikipedia sayfalarının bazılarının farklı formatlarda olmasından kaynaklanan zorluklar yaşanmıştır. Farklı formatlardaki sayfalar incelenmiş ve düzenlemeler yapılarak başarım arttırılmıştır.

Bu işlem sonucunda 2.465 adet kişi bilgi kutusuna sahip kayıt bulunmuş ve bu kayıtların bilgi kutusu verileri veritabanında ayrı bir alanda daha sonra kullanılmak üzere tutulmuştur.

[1] http://tr.wikipedia.org/wiki/Kategori:Ki%C5%9Fi_bilgi_kutusu_%C5%9Fablonlar%C4%B1



Şekil 4.3 Bilgi kutusu çıkarımı işlem adımları

4.2 Bilgi Kutusu Analizi

Bu bölümde bilgi kutusu şablonlarının çıkarılması ve şablonlara uyan yapısal verinin bilgi kutularından nasıl çıkarıldığı hakkında bilgi verilmektedir.

4.2.1 Kişi Bilgi Kutusu Şablonlarının Çıkartılması

Bölüm 4.1'de belirtilen adımlar ile kişilere ait bilgi kutuları metinler içerisinden ayrıştırılmıştır. Kişilere ait şablonlarda belirtilen veriler, yazarlarca bilgi kutularında doldurulmuş olarak bulunmaktadır. Şablonla ilgili detaylı bilgi Bölüm 2.4.5'de verilmiştir.

Gerçeklenen sistem kişilerin doğum yeri, doğum tarihi, ölüm tarihi bilgileri üzerine çalıştırılacağından bilgi kutularından bu üç bilginin çıkartılması için şablonlarda bu bilgilerin nasıl bulunduğu incelenmiştir.

Şablonlarda bilgilerin sabit formatlarda olması beklenirken, farklılıklar içerdiği görülmüştür. Bu farklılıklardan dolayı tüm şablonlar incelenerek verilerin bu şablonlarda ne tur formatlarda bulunduğu çıkartılmıştır. Şablonlarda üç verinin ne tür formatlarda bulunduğu Çizelge 4.1'de listelenmiştir.

Çizelge 4.1 Şablon veri formatları

Doğum Yeri	Doğum Tarihi	Ölüm Tarihi
Doğduğuşehir	doğumtarihi	ölümtarihi
doğduğuülke	doğum	ölüm_tarihi
doğum yeri	doğum_tarihi	Ölüm tarihi ve yaşı
doğum	Doğum tarihi	Ölüm yılı ve yaşı
doğum_yeri	Doğum tarihi ve yaşı	ölüm
Doğum yeri	Doğum yılı ve yaşı	Ölüm tarihi
Doğumyeri	birth_date	death_date
birth_place	birthdate	death date
Birth place	birth date	deathdate
Birthplace	Born	
Yer		

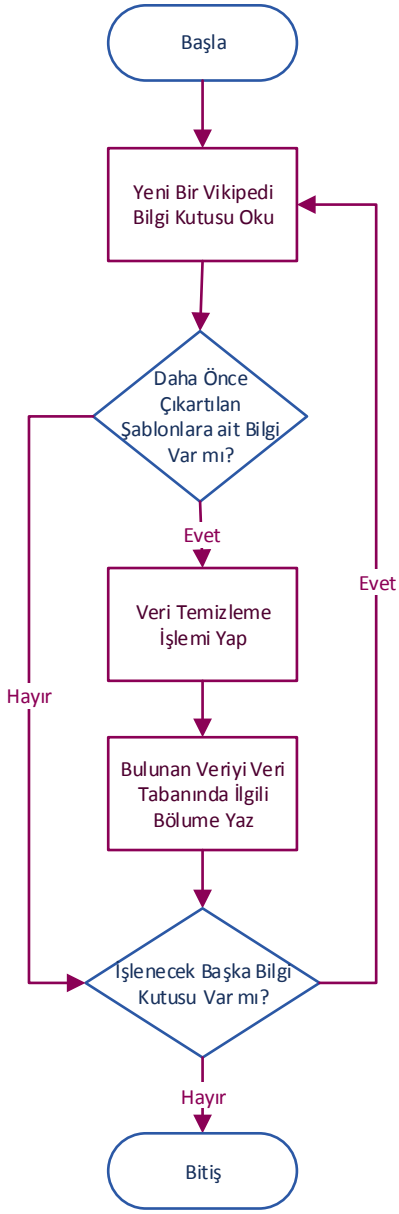
4.2.2 Kişilere Ait Bilgi Kutularından, Şablonlara Uygun Yapısal Verinin Çıkarılması

Bölüm 4.2.1 ‘de belirlenen şablonlardaki formatlar kullanılarak bilgi kutusuna sahip Wikipedia sayfalarındaki bilgi kutularından kullanılacak olan alanlar çıkartılmıştır. Bu aşamada bilgi kutusu şablonlarına uygun olarak girilen verilerdeki tutarsızlıklar fark edilmiş ve bunlar üzerine ek çalışmalar yapılması gerekmiştir. Örneğin: Doğum yeri alanına “İzmir, 1995” gibi bilgilerin girildiği tespit edilmiştir. Bir başka örnek olarak doğum yeri alanında “İzmir ilinin Bornova ilçesi” gibi bilgiler bulunabilmektedir. Bu tür bilgiler doğum tarihi ve ölüm tarihi alanlarında da bulunmaktadır.

Bu tür verilerden dolayı şablona uygun doldurulan veriler çıkartıldıktan sonra sistem tarafından veri temizleme işlemleri uygulanmıştır.

Veri Temizleme işlemi tamamlandıktan sonra, çıkartılan veriler veritabanında ayrı ayrı bölümler olarak tutulmaktadır. Böylelikle Wikipedia ’da metni olan bir kişi ile ilgili, eğer bilgi kutusunda verileri doldurulmuş ise, doğum yeri, doğum tarihi, olum tarihi bilgileri elde edilmiş olmaktadır.

Bu aşamada yapılan işlem ile ilgili işlem adımları Şekil 4.4‘de gösterilmiştir.



Şekil 4.4 Bilgi kutusu veri çıkarımı

4.3 Sisteminin Hazırlanması

Bu bölümde CRF'in çalıştırılabilmesi için gerekli olan arka planda yer alan iş adımları anlatılmaktadır.

4.3.1 CRF Sistemi için Gerekli Altyapının Çıkartılması

Gerçeklenen sistemde CRF altyapısının kullanılması kararlaştırılmıştır. Bu doğrultuda öncelikle mevcutta bulunan CRF araçları incelenmiştir. CRF ++, Crfsgd, Flexcrf, Crfsuite

ve Mallet örnek olarak verilebilir. Java dilinde yazılmış olmasından ve sağladığı esnekliklerden dolayı Mallet kullanımı uygun görülmüştür.

4.3.2 CRF için Gerekli Özelliklerin Analizi ve Gerçeklemesi

Şartlı Rastgele Alanlar yönteminde kullanılmak üzere belirlenen temel özellikler, herhangi bir doküman türüne bağlı olmaksızın çıkartılmıştır. Temel özellikler Şartlı Rastgele Alanlar 'da eğitim veri setinden olasılık hesaplama için kullanılmaktadır. Dokümandaki her kelime için özelliklere bakılır. Şartlı Rastgele Alanlarda her kelime için özellik bilgileri etiket bilgisi ile belirtilmektedir. Kelimelere ait özelliklerin çıkarımı ile ilgili adımlar Şekil 4.6'da gösterilmiştir.

4.3.2.1 Kelimenin Kendisi

Kelimenin kendisi, kelimeye W="kelime" özelliği olarak eklenir. Böylelikle aynı kelimelerin aynı varlık tipine işaret etme olasılığı hesaplanmış olur.

4.3.2.2 Büyük Harfle Başlama

Türkçe bir dokümanda kelimenin büyük harf ile başlayıp başlamamasına dikkat edilmektedir. Bir kelime büyük harf ile başlıyorsa bu durumda "STARTCAPITALIZE" özelliği bu kelimeye atanır.

4.3.2.3 Parantez İçinde Olma

Bir kelime metin içerisinde parantez arasında bulunuyor ise "INBRACKET" özelliği bu kelimeye atanır.

4.3.2.4 Kelimenin Yıl Formatında Olması

Bir kelime Tarih formatına uygun bir yapıya sahip ise "BWC= DATE" özelliği kelimeye atanır.

4.3.2.5 Kelimenin Sayısal Bir Değer İçermesi

Bir kelime içerisinde herhangi bir sayısal değer içeriyorsa, "HASDIGIT" özelliği kelimeye atanır.

4.3.2.6 Kelimenin Tamamen Sayısal Değerlerden Oluşması

Bir kelime sadece sayısal değerlerinden oluşuyor ise “ALLDIGITS” özelliği kelimeye atanır.

4.3.2.7 Kelimenin Tamamının Büyük Harf İçermesi

Bir kelime tamamen büyük harflerden oluşuyor ise “CAPITALIZED “ özelliği kelimeye atanır.

4.3.2.8 Kelimenin Noktalama İşareti İçermesi

Kelimeler kesme (‘), çift tırnak (“ ”), nokta (.), virgül (,), noktalı virgül (;), iki nokta (:) şeklindeki noktalama işaretleri içeriyor ise, içerdiği noktalama işaretine göre farklı özellikler atanır. Örn: ”HAS_DASH”,” HAS_SLASH”, “START_MINUS”.

4.3.2.9 Kelimenin NGram’ları

Geliştirilen sistemde bazı son ekleri belirleyebilmek için n-gram çıkarımına ihtiyaç duyulmuştur. N-gram, bir karakter katarının n adet karakter dilimidir [17]. Geliştirilen sistemde özellik çıkarımında kullanmak üzere n-gram’ın farklı birkaç uzunluğu olarak 2-gram, 3-gram, 4-gram’lar çıkartılmıştır.

Kelimenin 2-gram, 3-gram, 4-gram’ları kelimeye özellik olarak atanır. Örneğin: İzmir kelimesinde “izm”, “zmi”,”mir”,”izm”,”zmir” özellikleri kelimeye özellik olarak atanmıştır.

4.3.2.10 Kelimenin Önceki ve Sonraki Kelime Özellikleri

Kelimeden önce gelen beş kelime ve kelimeden sonra gelen beş kelime kelimeye özellik olarak eklenir.

Örn: Kelimeden önce ve sonra gelen beş kelimenin “WORD=” özellikleri kelimeye özellik olarak eklenir. Böylelikle “izmir ilinde ölmüştür” şeklinde bir kelime geldiğinde

"WORD=ilinde" ve "WORD=ölmüştür" özellikler cümledeki "İzmir" kelimesine özellik olarak eklenmiş olacaktır.

4.3.3 Olasılık Hesabı

Geliştirilen sistemde örüntü tanıma için olasılık değerlerinin belirlenmesinde Şartlı Rastgele Alanlar kullanılmaktadır. Şartlı Rastgele Alanlarda ağırlıklı olarak Maksimum Entropi ilkesi kullanılarak, eğitim veri setinden tahmini olasılık dağılımı hesaplanmaktadır.

Eğitim verilerinin $\{(\mathrm{x}^{(\mathrm{k})}, \mathrm{y}^{(\mathrm{k})})\}$ bağımsız ve eşit dağıtılmış olduğu varsayıldığında, Eşitlik 4.1' deki fonksiyon parametresi olan λ'nın tüm eğitim setinin olabilirliği (likelihood) $\mathrm{p}(\{\mathrm{y}^{(\mathrm{k})}\}|\{\mathrm{x}^{(\mathrm{k})}\}, \lambda)$'de ifade edildiği şekildedir. Maksimum olabilirlik eğitimi olabilirlik logaritmasında (log-likelihood), olasılığı maksimize edecek şekilde parametre değerleri seçer [23]. Şartlı Rastgele Alanlarda olabilirlik logaritması (4.1)'deki gibidir.

$$\mathcal{L}(\lambda) = \sum_k \left[\log \frac{1}{Z(x^{(k)})} + \sum_j \lambda_j F_j\,(y^{(k)}, x^{(k)})\right] \tag{4.1}$$

Burada verilen fonksiyon, global maksimum için yakınsama garanti, iç bükey fonksiyonudur [29].

Olabilirlik logaritmasının farklılaştırılması λ_j parametresi ile belirtilir;

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda_j} = E_{\tilde{p}(Y,X)}[F_j(Y,X)] - \sum_k E_{p(Y|x^{(k)},\lambda)}\,[F_j(Y, x^{(k)})] \tag{4.2}$$

Eşitlik 4.2'de verilen $\tilde{p}(Y,X)$ eğitim veri setinin ampirik dağılımını, $E_p[.]$'de p dağılımının beklentisini gösterir. Buradaki türevin sıfır verimi için Maksimum Entropi Model'in kısıtı; her özelliğin model dağılımı eğitim veri setinin ampirik dağılımının altındaki değere eşit olmasıdır [29].

Şartlı Rastgele Alanlarda en yüksek olabilirlik parametre değerlerini belirleyebilmek için eğitim veri seti içerisindeki her gözlem dizisinin her özellik işlevinin dağılımını hesaplamak mümkün olmalıdır.

4.4 CRF Sisteminin Eğitilmesi

Bu bölümde CRF sisteminin eğitimi için gerekli olan veri setinin etiketlenmesinden bahsedilecektir.

4.4.1 Bilgi Kutusu Verileri ile Wikipedia Sayfalarının İşaretlenmesi

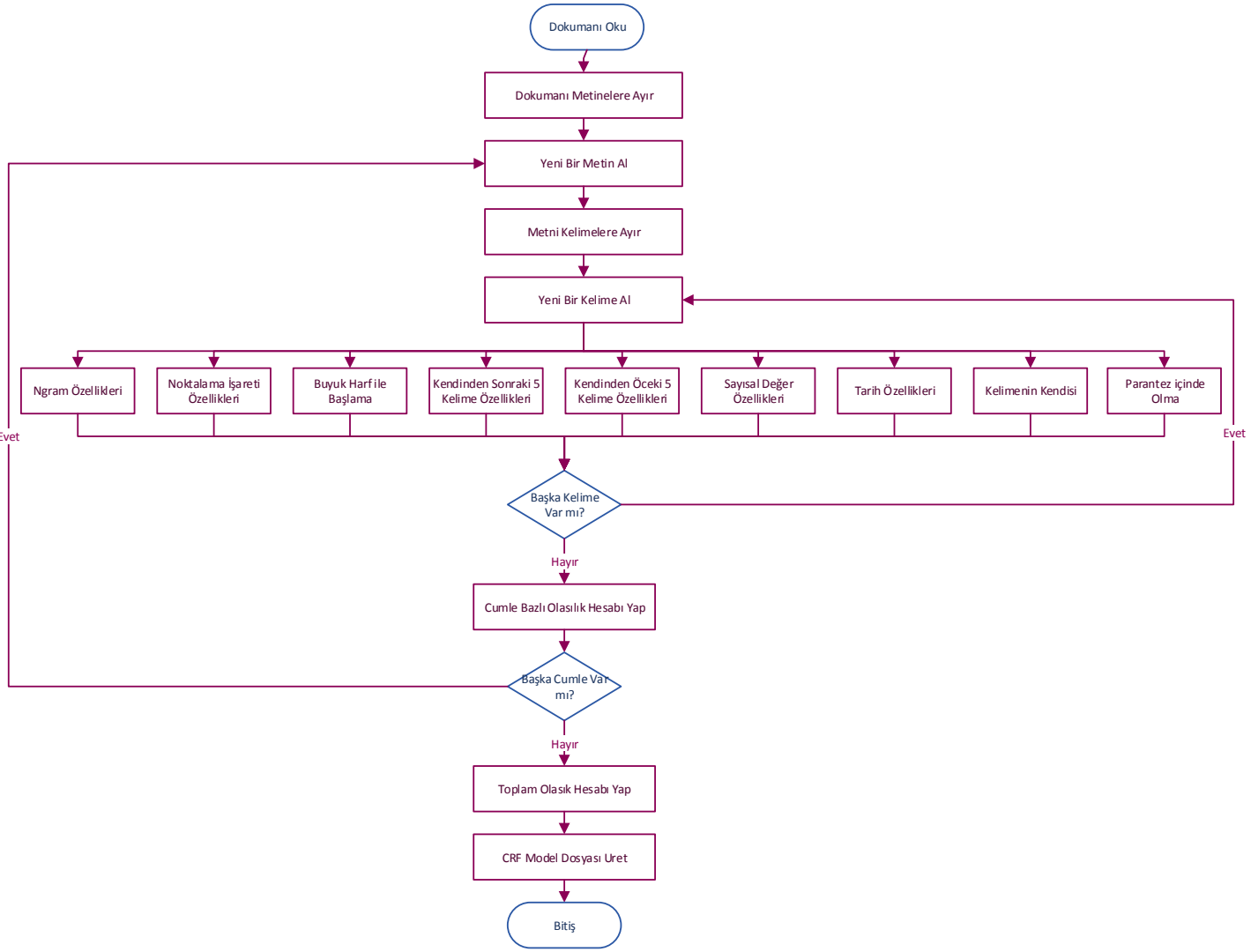
Bölüm 4.1.1 'de belirtilen adımlar ile elde edilen yazarlarca doldurulmuş bilgi kutusu verileri ve bu bilgi kutularının ait oldukları kişilerin detaylı metinleri kullanılarak CRF veri seti oluşturulmaktadır. Bu aşamada bilgi kutusunda bulunan bilgiler metin içerisinde aranmakta ve bulunduğunda bilgi kutusundaki hangi bölümde bulunduğuna göre etiketleme yapılmaktadır. Etiketli veri seti oluşturulurken metinlerdeki etiketlenmiş kelimeler ve bu kelimelerden sonra ve önceki beş kelime eğitim setine eklenmektedir. Eğitim veri seti oluşturulurken standartlara uygun şekilde etiketleme yapılmıştır. Şekil 4.5'de otomatik oluşturulan eğitim veri seti örneği bulunmaktadır.

```
Yaşamı ==Zakir Ramiyev 23 Kasım <ENAMEX_TYPE="p_dogum_tar">1859</ENAMEX> tarihinde, günümüzde Rusya Federasyonu
bu döneme aittir. 9 Ekim <ENAMEX_TYPE="p_olum_tar">1921</ENAMEX> tarihinde Orsk'ta vefat ettiği zaman
1920 - ö. 1 Nisan <ENAMEX_TYPE="p_olum_tar">1987).Çek</ENAMEX> asıllı yazar, filozof.Freud'un fikirlerini tamam
bulunan toplumcu tiyatro. 1 Mayıs <ENAMEX_TYPE="p_dogum_tar">1989'da</ENAMEX> Hüseyin Hilmi Bulunmaz tarafından
kutusu}}'''Volkan Tokcan''' (d. 11 Ocak <ENAMEX_TYPE="p_dogum_tar">1988</ENAMEX> - <ENAMEX_TYPE="p_dogum_yer">İ
'''Kemal Bayram Çukurkavaklı''' (d. <ENAMEX_TYPE="p_dogum_tar">1934-</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">Konya,
kutusu}}'''Halim Baykuş''' (d. 5 Haziran <ENAMEX_TYPE="p_dogum_tar">1956,</ENAMEX> Nobrain), Türk vatandaşı|Tür
{{kayda değerlik}}'''Murat Soyak''' (doğum. <ENAMEX_TYPE="p_dogum_tar">1971,</ENAMEX> Niğde), Türk Şair-Yazar{{
kutusu}}'''Nilüfer Yenidoğan''' (d. 1 Nisan <ENAMEX_TYPE="p_dogum_tar">1975-</ENAMEX> <ENAMEX_TYPE="p_dogum_yer
-->{{Müzik sanatçısı bilgi kutusu}}'''Gökhan Büyükkara''' <ENAMEX_TYPE="p_dogum_tar">1983</ENAMEX> <ENAMEX_TYPE
vefat etmiştir.Zübeyir Gündüzalp 2 Nisan <ENAMEX_TYPE="p_olum_tar">1971</ENAMEX> Cuma günü İstanbul'da vefat et
Beder''', yemek yazarı.{{Kişi bilgi kutusu}} <ENAMEX_TYPE="p_dogum_tar">1960</ENAMEX> yılında Konya'nın Akşehir
kutusu}}'''Serhan Kavut''' (d. 21 Eylül <ENAMEX_TYPE="p_dogum_tar">1981,</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">Bu
Bey'in Yargılanması</ref>, (ö. 10 Nisan <ENAMEX_TYPE="p_olum_tar">1919),</ENAMEX> I. Dünya Savaşı'nın son yılla
kutusu}}'''Daniel Lavoie''' (d. 17 Mart <ENAMEX_TYPE="p_dogum_tar">1949,</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">Ma
{{Kayda değerlik|biyografi}}{{Kişi bilgi kutusu}}Özgür Oktel <ENAMEX_TYPE="p_dogum_tar">1974</ENAMEX> <ENAMEX_T
Capli)'''Kerim Çaplı''' (d. 13 Ocak <ENAMEX_TYPE="p_dogum_tar">1949,</ENAMEX> Karşıyaka, İzmir - ö. 2 Kasım <EN
Causes dergisinin çizeridir.== Özgeçmiş == <ENAMEX_TYPE="p_dogum_tar">1976</ENAMEX> yılında Ankara'da doğdu. Ça
kutusu}}'''Nurettin Yılmaz''', (d. 1 Ocak <ENAMEX_TYPE="p_dogum_tar">1962,</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">
{{Kaynaksız}}{{Basketbolcu bilgi kutusu}}'''Serhat Büker''' (d. <ENAMEX_TYPE="p_dogum_tar">1981,</ENAMEX> <ENAM
sanatçısı bilgi kutusu}}'''Sibel Egemen''' (d. <ENAMEX_TYPE="p_dogum_tar">1958,</ENAMEX> <ENAMEX_TYPE="p_dogum_
bilgi kutusu}}Jamey Jasta 7 Ağustos <ENAMEX_TYPE="p_dogum_tar">1977'de</ENAMEX> West Haven, Amerika Birleşik <E
'''Burak Güven''' (d. 19 Ekim <ENAMEX_TYPE="p_dogum_tar">1975,</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">İstanbul)</E
olan (doğum tarihi 8 aralık <ENAMEX_TYPE="p_dogum_tar">1967)</ENAMEX> Hollandalı elektronik müzik sanatçısı. Ju
kutusu}}Kamber Arslan (d. 5 Ocak <ENAMEX_TYPE="p_dogum_tar">1980,</ENAMEX> <ENAMEX_TYPE="p_dogum_yer">Kayseri)<
kutusu}}'''Louis de Berniéres''' 8 Aralık <ENAMEX_TYPE="p_dogum_tar">1954'te</ENAMEX> <ENAMEX_TYPE="p_dogum_yer
({{lang-mk|Игор Николовски}}; d. 16 Temmuz <ENAMEX_TYPE="p_dogum_tar">1973),</ENAMEX> Makedonyalı eski millî fu
```

Şekil 4.5 Eğitim veri seti örneği

4.4.2 Oluşturulan Eğitim Seti ile Sistemin Eğitilmesi

Oluşturulan eğitim seti ile sistemin eğitilebilmesi için Mallet kütüphanesi kullanılarak bir java uygulaması geliştirilmiştir. Uygulamanın eğitim seti oluşturulduktan sonraki işlem adımları Şekil 4.6'da detaylı olarak gösterilmiştir. Uygulama kütüphane olarak da kullanılabilecek şekilde düzenlenmiştir. Eğitim seti ile sistem eğitildiğinde olasılık hesaplarının bulunduğu bir model dosyası üretilmektedir.



Şekil 4.6 Sistemin eğitilmesi

BÖLÜM 5

DENEYSEL SONUÇLAR

Sistemin test edilebilmesi için üç farklı veri seti (VS1, VS2 ve VS3) oluşturulmuştur. Oluşturulan veri setleri ile sistem test edilmiş ve başarımları ölçülmüştür.

Sistemin test edilmesi için oluşturulmuş olan her veri setinin hem eğitim hem de test aşamasında sahip olduğu özellikler Çizelge 5.1'de verilmiştir.

Çizelge 5.1 Veri setlerinin genel yapısı

		# wikipedia sayfa	#doğum yeri	#doğum tarihi	#ölüm tarihi
VS1	**Eğitim**	2174	999	1915	663
	Test	2174	999	1915	663
VS2	**$Eğitim_1$**	1474	631	1335	516
	$Test_1$	700	673	669	311
	$Eğitim_2$	1474	696	1320	413
	$Test_2$	700	656	671	376
	$Eğitim_3$	1400	671	1255	397
	$Test_3$	774	727	749	398
VS3	**$Eğitim_1$**	2138	984	1955	663
	$Test_1$	46	46	46	46
	$Eğitim_2$	2039	998	1828	663
	$Test_2$	145	145	145	145
	$Eğitim_3$	2129	998	1955	617
	$Test_3$	55	55	55	55

5.1 Test Seti (VS1)

Wikipedia sayfalarında kişi bilgi kutusuna sahip tüm sayfalar ele alınmıştır. Bu sayfalarda bulunan bilgi kutusundaki yazarlar tarafından etiketlenmiş varlık isimleri (doğum yeri, doğum ve ölüm tarihi) ile sistem eğitilmiştir.

İkinci aşamada kişi bilgi kutusuna ait bilgilerin bulunduğu sayfalardaki metinler çıkartılmıştır. Metinler eğitmiş olduğumuz sisteme gönderilerek içerisinden doğum yeri, doğum tarihi ve ölüm tarihi varlık isimlerini çıkartması sağlanmıştır.

Son aşamada sistemin metinlerde bulduğu varlık isimleri ile Wikipedia sayfalarında bulunan bilgi kutularındaki yazarlar tarafından doldurulan varlık isimleri karşılaştırılmış ve sistemin başarımı ölçülmüştür. Sistemin test sonuçlarında aşağıdaki durumlar ortaya çıkmaktadır:

C: Bilgi Kutusunda yer alan bilgiyi wiki sayfasında doğru olarak buluyorum

S: Bilgi Kutusunda yer alan bilgiyi wiki sayfasında yanlış olarak buluyorum

D: Bilgi Kutusunda yer alan bilgiyi wiki sayfasında bulamıyorum

I: Bilgi Kutusunda bilgi yok ancak, wiki sayfasından bilgi çıkarıyorum.

Bu notasyon benzer sistemler olarak geliştirilmiş IPopulator [9] ve Makhoul [31] çalışmalarında da kullanılmıştır. Tutturma (Precision) eşitlik (5.1), Bulma (Recall) (5.2) ve F-ölçüm (5.3) hesaplanmıştır.

$$Tutturma = \frac{|C|}{|C|+|S|} \quad (5.1)$$

$$Bulma = \frac{|C|}{|C|+|S|+|D|} \quad (5.2)$$

$$F - \text{ölçüm} = 2 * Tutturma * Bulma\ /(Tutturma + Bulma\ ) \quad (5.3)$$

Hesaplamalar yapılırken, Wikipedia bilgi kutularında olmayan ancak, sistemin bulduğu varlık isimlerinin doğruluğu otomatik olarak tespit edilemediğinden bu değerler yanlış ya da doğru olarak kabul edilmemiş ve ölçümlere dahil edilmemiştir.

Ölçümlemeler neticesinde elde edilen test sonuçları Çizelge 5.2'de gösterilmiştir.

Çizelge 5.2 VS1 Test sonuçları

Varlık	Tutturma	Bulma	F-ölçüm
Doğum Yeri	0,91	0,71	0,79
Doğum Tarihi	0,97	0,93	0,94
Ölüm Tarihi	0,97	0,96	0,96

Bu adımda sistem elimizde bulunan tüm kişi Wikipedia sayfalarının bilgi kutularında yazarlarca etiketlenmiş varlık isimleri ile eğitilmiş ve aynı sayfaların metinleri ile de test edilmiştir. Sistemin ezberleyeceğini düşünerek her varlık ismi için yüzde yüze yakın bir başarı alınması bekleniyordu. Ölçümlemeler sonucunda doğum yeri bilgisindeki başarımın diğer alanlara oranla daha düşük olduğu gözlemlenmiştir. Bu sonucun en önemli sebebi doğum yeri bilgisinin metin içerisinde diğer alanlara oranda, çok farklı formatlarda bulunabilmesidir. Örneğin, doğum yeri bilgisi metin içerisinde "İzmir ilinde doğmuştur", "İzmir ilinin Bornova semtinde dünyaya gelmiştir.", "İzmirde hayata gözlerini açmıştır." gibi. Farklı formatlardan yeteri miktarda eğitim verisi sağlandığında bu alandaki başarımda artış görülecektir. Doğum yeri alanının sistem tarafından tespitinin düşük olmasının bir diğer sebebi ise, metin içerisinde geçen doğum yeri varlık isminin, bilgi kutusunda etiketlenirken farklı şekilde tanımlanmış olmasıdır. Örneğin: Doğum yeri bilgisi bilgi kutusunda "İstanbul" olarak tanımlanmış ancak, Wikipedia sayfasının içeriğinde bu bilgi Üsküdar olarak geçmektedir. Bu durumda eğitim seti otomatik olarak üretilirken bilgi kutusunda belirtilen bilgi metinde bulunamadığından eğitim setine dahil edilmemektedir. Bu gibi durumlar doğum yeri alanının tespitinde, sistem başarımını etkilemektedir.

5.2 Test Seti (VS2)

İkinci veri setinde elimizde bulunan kişi bilgi kutusuna sahip Wikipedia sayfaları üç gruba ayrılmıştır. Bu şartı sağlayan toplamda 2.174 adet Wikipedia metni bulunmaktadır. Bu metinler 1. grup= 700, 2. grup 700 ve 3. grup 774 adet olacak şekilde çapraz geçerleme yapabilmek için ayrılmıştır.

Testlerde sırası ile 1. grup veri, eğitim setinden çıkartılarak kalan iki gruptaki Wikipedia bilgi kutularındaki yazarlar tarafından etiketlenmiş varlık isimleri ile eğitim seti otomatik olarak oluşturulmuş ve sistem eğitilmiştir. Bu işlem her 3 set için tekrarlanmış ve karşılaştırmalı sonuçlar üretilmiştir. Başarım ölçümü yapılırken VS1 de kullanılan Bulma, Tutturma, F-ölçüm hesaplamaları kullanılmıştır. Sonuçlar Çizelge 5.3, Çizelge 5.4, Çizelge 5.5' de gösterilmektedir.

1. **Birinci grup 700 test sonucu:**

Çizelge 5.3 VS2-1 Test sonuçları

Varlık	Tutturma	Bulma	F-ölçüm
Doğum Yeri	0,92	0,55	0,69
Doğum Tarihi	0,93	0,84	0,89
Ölüm Tarihi	0,92	0,81	0,87

2. **İkinci grup 700 test sonucu:**

Çizelge 5.4 VS2-2 Test sonuçları

Varlık	Tutturma	Bulma	F-ölçüm
Doğum Yeri	0,78	0,55	0,65
Doğum Tarihi	0,89	0,83	0,86
Ölüm Tarihi	0,89	0,85	0,87

3. **Üçüncü grup 774 test sonucu:**

Çizelge 5.5 VS2-3 Test sonuçları

Varlık	Tutturma	Bulma	F-ölçüm
Doğum Yeri	0,72	0,47	0,57
Doğum Tarihi	0,91	0,87	0,89
Ölüm Tarihi	0,88	0,86	0,87

Sistemin ortalama F-ölçüm değeri Doğum yeri için 0.64, Doğum tarihi için, 0.88 ve Ölüm tarihi için 0.87 olarak elde edilmiştir. VS2 için gerçekleştirilen testlerin başarımının, VS1 için gerçekleştirilen testlerin başarımından düşük olduğu

görülmektedir. Bunun sebebi, VS1'de sistem eğitilirken elimizde bulunan bilgi kutularında yer alan varlık isimlerinin tamamı kullanılmıştır. VS2 testlerinde ise çapraz geçerleme yapılacağından eğitim seti hazırlanırken, elimizdeki etiketlenmiş verinin 2/3'ü kullanılmıştır. Eğitim setindeki verinin 1/3 oranında azalması sistemin başarımını etkilemektedir.

5.3 Test Seti (VS3)

Veri Seti 3'te üç farklı test yapılmıştır. Her test seti için kişi bilgi kutusuna sahip Wikipedia sayfalarından, doğum yeri, doğum tarihi, ölüm tarihi varlık isimlerinden sadece birinin bilgi kutusunda yazarlar tarafından etiketlenmiş olduğu sayfalar eğitim setinden çıkartılmıştır. Daha sonra eğitim seti kalan Wikipedia metinleri ile eğitilip, test için ayrılan sayfaların metinleri ile sistem çalıştırılarak ölçümlemeler yapılmıştır.

VS1 ve VS2 testlerinde sistem başarımı ölçülürken bilgi kutularındaki etiketlenmiş varlık isimleri ile sistemin bulduğu varlık isimleri karşılaştırılmıştı. VS3 testleri yapılırken test setinde bulunan metinler okunarak varlık isimleri elle etiketlenmiştir. Başarı ölçümlemeleri yapılırken sistemin bulduğu varlık isimleri ile elle etiketlenmiş varlık isimleri karşılaştırılmıştır.

Test sonuçlarının ölçümlemesinde hata matrisi kullanılmıştır (Çizelge 5.6). Tutturma (5.4) ve Bulma (5.5), ve Doğruluk (5.7) ile hesaplanmıştır.

Test seti VS1 ve test seti VS2 'de tüm veri seti üzerinde sistem test edilmiştir. Sistemin bulduğu varlık isimleri ile bilgi kutularındaki işaretlenmiş varlık isimleri karşılaştırılmıştır. Bilgi kutusunda bulunmayan bir varlık ismi sistem tarafından metin içerisinden çıkartıldığında, bulunan varlık isminin doğruluğu net bir şekilde bilinmediğinden VS1 ve VS2 için hata matrisi kullanılmamıştır.

Çizelge 5.6 Hata matrisi detayı

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	TP (doğru pozitif)	FP(yanlış pozitif)
	Varlık İsmi Değil	FN(yanlış negatif)	TN(doğru negatif)

TP: Sayfada bulunan varlık ismini sistem doğru olarak buluyor.

TN: Sayfada olmayan varlık ismini sistem bulamıyor.

FN: Sayfada bulunan varlık ismini sistem ya hatalı olarak buluyor ya da bulamıyor.

FP: Sayfada olmayan bir varlık ismi sistem tarafından bulunuyor.

$$Tutturma = \frac{TP}{TP+FP} \quad (5.4)$$

$$Bulma = \frac{TP}{TP + FN} \quad (5.5)$$

$$F - \text{ölçüm} = \frac{2 * Tutturma * Bulma}{Tutturma + Bulma} \quad (5.6)$$

$$Doğruluk = \frac{TP+TN}{TP + TN + FP + FN} \quad (5.7)$$

1. **Birinci grup: Bilgi Kutusunda Sadece Doğum Yeri Bilgisi Bulunan Metinler**

İşlem Adımları:

1. Bilgi Kutusunda sadece doğum yeri bilgisi dolu olan, kişi bilgi kutusuna sahip 46 adet Wikipedia metni tespit edilmiştir.
2. Kişi bilgi kutusuna sahip, bilgi kutusunda doğum yeri, doğum tarihi ve ölüm tarihi bilgisinden herhangi birisinin tanımlı olduğu ve birinci adımda tespit edilen 46 adet Wikipedia metni dışında kalan metinler ile eğitim seti otomatik olarak hazırlanmıştır.

3. Oluşturulan eğitim seti ile sistem eğitilmiştir.

4. Eğitim setinden hariç tutulan 46 adet Wikipedia metni için doğum tarihi, ölüm tarihi ve doğum yeri bilgisi çıkarımı test edilmiş ve ölçümlemeler yapılmıştır. Testler sırasında bilgi kutusunda işaretlenmiş bilgiler yerine, wikipedia metinleri okunarak doğum tarihi, ölüm tarihi ve doğum yeri bilgileri de elle etiketlenmiştir. Sistemin bulduğu değerler ile elle etiketlenmiş doğum tarihi, ölüm tarihi ve doğum yeri etiketleri karşılaştırılmıştır. Doğum yeri için yapılan test sonuçlarının hata matrisi Çizelge 5.7'de, doğum tarihi için yapılan test sonuçlarının hata matrisi Çizelge 5.8'de, ölüm tarihi için yapılan test sonuçlarının hata matrisi Çizelge 5.9'da ve gösterilmiştir. Testler sonucu elde edilen F-ölçüm değerleri ve doğruluk değerleri Çizelge 5. 10'da gösterilmektedir.

Çizelge 5.7 VS3-1 Doğum yeri için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	21	0
	Varlık İsmi Değil	12	13

Çizelge 5.8 VS3-1 Doğum tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	26	5
	Varlık İsmi Değil	3	12

Çizelge 5.9 VS3-1 Ölüm tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	20	3
	Varlık İsmi Değil	13	10

Çizelge 5.9'da ölüm tarihi bilgisinin başarımı örneklenmiştir. Bu sonuçta FN değerinin yüksekliği göze çarpmaktadır. FN değeri metinde ölüm tarihi değerinin bulunduğu ve sisteminde bu değeri bulduğu ancak yanlış bulduğu değerleri içermektedir. Ölüm tarihi bilgisi wikipedia metinlerindeki bilgi kutularında az sayıda işaretmiş olarak bulunmaktadır. Elimizdeki tüm bilgi kutularında toplamda 663 adet ölüm tarihi bilgisi dolu olarak bulunmaktadır. Bu durumda sistemin eğitiminde kullanılan eğitim setinde ölüm tarihi bilgisi sayıca az olmaktadır. Bu durum olasılık hesaplamalarındaki başarımın düşmesine neden olmaktadır.

Çizelge 5. 10 VS3-1 Ölçüm değerleri

Varlık	Tutturma	Bulma	F-ölçüm	Doğruluk
Doğum Yeri	1.0	0,63	0.77	0.73
Doğum Tarihi	0,83	0,89	0,85	0.82
Ölüm Tarihi	0.86	0.60	0.71	0.65

VS3-1 test seti için ölçüm değerleri incelendiğinde sistemin eğitim sırasında kullandığı veri setinin büyüklüğünün sonuçlara yansıdığı görülmektedir.

2. İkinci Grup: Bilgi Kutusunda Sadece Doğum Tarihi Bilgisi Bulunan Metinler

İşlem Adımları:

1. Bilgi Kutusunda sadece doğum tarihi bilgisi dolu olan, kişi bilgi kutusuna sahip 145 adet Wikipedia metni tespit edilmiştir.
2. Kişi bilgi kutusuna sahip, bilgi kutusunda doğum yeri, doğum tarihi ve ölüm tarihi bilgisinden herhangi birisinin tanımlı olduğu ve birinci adımda tespit

edilen 145 adet Wikipedia metni dışında kalan metinler ile eğitim seti sistem tarafından hazırlanmıştır.

3. Oluşturulan eğitim seti ile sistem eğitilmiştir.
4. Eğitim setinden hariç tutulan 145 adet Wikipedia metni için doğum yeri, ölüm tarihi ve doğum tarihi bilgisi çıkarımı test edilmiş ve ölçümlemeler yapılmıştır. Testler sırasında bilgi kutusunda işaretlenmiş bilgiler yerine, Wikipedia metinleri okunarak doğum yeri, ölüm tarihi ve doğum tarihi bilgileri elle etiketlenmiştir. Sistemin bulduğu varlık isimleri ile metinler okunarak yapılan doğum yeri, ölüm tarihi ve doğum tarihi etiketleri karşılaştırılmıştır. Doğum Yeri için yapılan test sonuçlarının hata matrisi Çizelge 5. 11'da, doğum tarihi için yapılan test sonuçlarının hata matrisi Çizelge 5. 12'de, ölüm tarihi için yapılan test sonuçlarının hata matrisi Çizelge 5. 13'de gösterilmiştir. Testler sonucu elde edilen ölçüm değerleri ise Çizelge 5. 14'de gösterilmektedir.

Çizelge 5. 11 VS3-2 Doğum yeri için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	65	6
	Varlık İsmi Değil	21	53

Çizelge 5. 12 VS3-2 Doğum tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	116	7
	Varlık İsmi Değil	9	13

Çizelge 5. 13 VS3-2 Ölüm tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	5	7
	Varlık İsmi Değil	2	131

Çizelge 5. 14 VS3-2 Ölçüm değerleri

Varlık	Tutturma	Bulma	F-ölçüm	Doğruluk
Doğum Yeri	0.91	0.76	0.82	0.81
Doğum Tarihi	0.94	0.94	0.94	0.89
Ölüm Tarihi	0.41	0.71	0.52	0.93

VS3-2 test seti için ölçüm değerleri incelendiğinde sistemin eğitim sırasında kullandığı veri setinin büyüklüğünün sonuçlara yansıdığı görülmektedir. Bunun sonucu olarak doğum yeri bilgisinde sistem başarımı ölüm tarihine göre daha yüksek çıkmıştır. Buna ek olarak doğum yeri bilgisinin metin içerisinde doğum tarihi bilgisine göre daha çeşitli formatlarda geçebildiğinden, doğum yeri bilgisinin sistem tarafından tespit edilme başarımının doğum tarihi bilgisine göre daha düşük olduğu gözlemlenmiştir.

3. Üçüncü Grup: Bilgi Kutusunda Sadece Ölüm Tarihi Bilgisi Bulunan Metinler

İşlem Adımları:

1. Bilgi Kutusunda sadece ölüm tarihi bilgisi dolu olan, kişi bilgi kutusuna sahip 55 adet Wikipedia metni tespit edilmiştir.
2. Kişi bilgi kutusuna sahip, bilgi kutusunda doğum yeri, doğum tarihi ve ölüm tarihi bilgisinden herhangi birisinin tanımlı olduğu ve birinci adımda tespit edilen 55 adet Wikipedia metni dışında kalan metinler ile eğitim seti sistem tarafından hazırlanmıştır.
3. Oluşturulan eğitim seti ile sistem eğitilmiştir.

4. Eğitim setinden hariç tutulan 55 adet Wikipedia metni için doğum tarihi, doğum yeri ve ölüm tarihi bilgisi çıkarımı test edilmiş ve ölçümlemeler yapılmıştır. Testler sırasında bilgi kutusunda işaretlenmiş bilgiler yerine, Wikipedia metinleri okunarak doğum yeri, doğum tarihi ve ölüm tarihi bilgileri elle etiketlenmiştir. Sistemin bulduğu değerler ile metinler okunarak yapılan doğum yeri, doğum tarihi ve ölüm tarihi etiketleri karşılaştırılmıştır. Doğum yeri için yapılan test sonuçlarının hata matrisi Çizelge 5.15'de, doğum yeri için yapılan test sonuçlarının hata matrisi de Çizelge 5.16'da ve ölüm tarihi için yapılan test sonuçlarının hata matrisi de Çizelge 5. 17'de gösterilmiştir. Test grubundaki metinlerin büyük bölümünü sadrazamlar ve hükümdarlara ait Wikipedia metinlerinden oluşmaktadır. Bu metinlerde de doğum bilgilerine yer verilme oranı oldukça azdır. Sistemin başarısı Çizelge 5.18'de gösterilmektedir.

Çizelge 5.15 VS3-3 Doğum yeri için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	11	13
	Varlık İsmi Değil	5	26

Çizelge 5.16 VS3-3 Doğum tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	22	22
	Varlık İsmi Değil	3	8

Çizelge 5. 17 VS3-3 Ölüm tarihi için hata matrisi

		Gerçek	
		Varlık İsmi	**Varlık İsmi Değil**
Tahmin	**Varlık İsmi**	45	0
	Varlık İsmi Değil	10	0

Test grubundaki metinlerin büyük bölümü sadrazamlar ve hükümdarlara ait Wikipedia metinlerinden oluşmaktadır. Bu metinlerde de doğum bilgilerine yer verilme oranı oldukça azdır. Bu durumun sonucu olarak ölçümlemelerde FP ve TP değerleri yüksek çıkmaktadır.

Çizelge 5.18 VS3-3 Ölçüm değerleri

Varlık	Tutturma	Bulma	F-ölçüm	Doğruluk
Doğum Yeri	0.45	0.68	0.55	0.67
Doğum Tarihi	0.50	0.88	0.64	0.55
Ölüm Tarihi	1	0.81	0.89	0.81

VS3-3 test seti ile yapılan testlerde, test için ayrılan Wikipedia metinlerinde çıkartılmak istenen veri büyük oranda bulunmadığından bir önceki veri setleri gibi yüksek başarımlar elde edilememiştir.

BÖLÜM 6

SONUÇ VE ÖNERİLER

İlişki çıkarımı uygulamalarının amacı resmi olan veya olmayan bir dilde yazılmış, belli bir çalışma alanına bağlı veya bağımsız olan tüm dokümanlar içerisinden, dile bağımlı veya bağımsız olarak varlıklar arasındaki ilişkileri ve bu ilişki türlerini bulmaktır.

Semantik arama teknolojileri yaygınlaştıkça ilişki çıkarımı yöntemlerine olan ihtiyaç artmıştır. Semantik arama teknolojilerinin istenilen başarımı sağlayabilmesi için altyapılarında varlıklar arasındaki ilişkilerin depolandığı veri tabanları bulunmaktadır.

İlişki çıkarımı alanında yapılan çalışmalar Türkçe ’de sınırlı sayıda olup, çalışmaların çoğu kural tabanlı olarak gerçekleştirilmiştir. Bu çalışmada resmi olmayan yapılandırılmamış Wikipedia dokümanlarından Şartlı Rastgele Alanlar kullanılarak kişilerin doğum yeri, doğum tarihi, ölüm tarihi bilgilerinin bulunması gerçekleştirilmiştir.

Eğitim verisi ne kadar çok etiketlenmiş varlık içerirse başarı oranı da o ölçüde artmaktadır. Ayrıca, eğitim verisinde bulunan varlıkların nitelikleri de başarıyı etkilemektedir.

Başarımı etkileyen faktörlerden bir tanesi de kelimelerden önce ve sonra gelen kelimelerin özelliklerinin Şartlı Rastgele Alanlar kullanılırken kelime özelliklerine yansıtılmasıdır. Bu işlem sağlandığında kural tabanlı yaklaşımlarda kullanılan birçok adım sağlanmış olmaktadır.

Wikipedia metinleri ne kadar güvenilir olarak nitelendirilse de çalışmanın başarımının arttırılması için bilgi kutusu verileri üzerinde veri temizleme işlemi yapılması başarımı arttırmaktadır.

Deneysel test sonuçlarına bakıldığında en yüksek başarımın doğum tarihi alanında alındığı görülmektedir. Bunun başlıca sebebi bu alandaki verinin belirli formatlarda olmasıdır. İkinci bir etken ise Çizelge 5.1 'de görüleceği gibi, eğitim setlerinde kullanılan bilgi kutuları alanlarında doğum tarihi bilgisinin doluluk oranının yüksek olduğu görülmektedir. Bunun sonucu olarak doğum tarihi için yapılan test sonuçları diğer alanlara oranla daha yüksek çıkmaktadır.

Doğum yeri alanındaki verinin çeşitliliği, metin içerisinde diğer varlık isimlerine oranla farklı formatlarda geçmesinden dolayı, sistemin doğum yeri varlık ismi çıkarımındaki başarımı doğum tarihi alanına göre daha düşük çıkmaktadır.

Ölüm tarihi alanındaki başarımlar incelendiğinde sistemdeki eğitim setinde bulunan etiketli verisinin azlığının sonuçlara olumsuz yansıması görülmektedir. Ancak eğitim setinde artış sağlandığında doğum tarihinde yakalanan başarıma bu varlık ismi için de ulaşılacaktır.

Çalışmanın bir sonraki adımında kişi bilgi kutularında bulunan doğum yeri, doğum tarihi ve ölüm tarihi dışında kalan diğer alanların çıkarımı üzerine çalışılması gerekmektedir. Bu işlem başarı ile tamamlandıktan sonra kişi bilgi kutuları haricinde diğer bilgi kutuları içinde aynı işlem tekrarlanarak Türkçe için ilişkilerin tutulduğu geniş bir veritabanı elde edilebilecektir

KAYNAKLAR

[1] Etzioni,O., Cafarella,M., Downey,D., Popescu, A.-M., Shaked,T., Soderland, S., (2005). "Unsupervised Named-entity Extraction from the Web an Experimental Study", Artificial Intelligence, 165: 91–134.

[2] Banko, M., Cafarella,M. J., Soderland, S., Broadhead, M.,Etzioni, O., (2007). "Open Information Extraction from the Web", Intl. Joint Conf. on Artificial Intelligence.

[3] Liu, Q., Xu, K., Zhang, L., Wang,H., Yu, Y., Pan, Y., (2008). "Catriple: Extracting Triples from Wikipedia Categories", Asian Semantic Web Conf.

[4] Ruiz-Casado, M., Alfonseca,E., Castells, P., (2007). "Automatising the Learning of Lexical Patterns: An Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia", Data & Knowledge Engineering, 61(3): 484–499.

[5] Wang, G., Zhang, H., Wang, H., Yu,Y., (2007). "Enhancing Relation Extraction by Eliciting Selectional Constraint Features from Wikipedia", Intl. Conf. on Applications of Natural Language to Information Systems.

[6] Copestake, A. Herbelot A. (2006). "Acquiring Ontological Relationships from Wikipedia Using RMRS", ISWC 2006 Workshop on Web Content Mining with Human Language Technologies.

[7] Nguyen,D. P. T., Matsuo, Y., Ishizuka, M., (2007). "Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia", IJCAI 2007 Workshop on Text-Mining & Link-Analysis.

[8] Weld, F. Wu and D. S. (2007)."Autonomously Semantifying Wikipedia", Conf. on Information and Knowledge Management.

[9] Lange, D.,Böhm, C.,Naumann, F., (2010). "Extracting Structured Information from Wikipedia Articles to Populate Infoboxes", Intl. Conf. of Information and Knowledge Management.

[10] Oxford Dictionary of English, Oxford University Press, 2003.

[11] BBC News, What is it with Wikipedia, http://news.bbc.co.uk/2/hi/technology/4534712.stm, 20 Kasım 2013.

[12] BBC News, Wikipedia survives research test ,http://news.bbc.co.uk/2/hi/technology/4530930.stm,20 Kasım 2013.

[13] Cucerzan, S., ve Yarowsky,. (1999). "Language Independent Named Entitiy Recognition Combining Morphological and Contextual Evidence", SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[14] Tür, G., Hakkani-Tür, D. Ve Oflazer, (2003)."A Statistical Information Extraction System for Turkish", Natural Language Engineering, (2): 181-210.

[15] Johnson, C. S., (1967)."Hierarchical Clustering Schemes", Psychometrica, 241-254.

[16] Wallach, H.M.,(2004). "Conditional Random Fields: An Introduction", University of Pennsylvania Technical Report.

[17] Anderson, J. R., Michalski, R. S., Carbonell, J. G., & Mitchell, T. M., (1985)."An Artificial Intelligence Approach", Machine Learning, (1).

[18] Manning, C., & Schutze, H.,(1999). "Foundations of Statistical Natural Language Processing.", MIT Press.

[19] Rabiner, L.R. (1989),"A tutorial on hidden Markov models and selected applications in speech recognition", IEEE, 77(2): 257-285.

[20] Wallach, H.M.,(2004), "Conditional Random Fields: An Introduction.", University of Pennsylvania.

[21] Rau, L.F., (1991). "Extracting Companu Names from Text", Artificial Intelligence Applications of IEEE.

[22] MacQueen, J. B., (1967)."Some Methods for classification and Analysis of Mutivariate Observations", Berkeley Symposium on Mathematical Statistics and Probability.

[23] Ekbal, A., Bandyopadhyay, S., (2009). "A Conditional Random Field Approach for Named Entity Recognition in Bengali and Hindi", Linguistic Issues in Language Technology-LiLT.

[24] Nadeau, D.,(2007). "Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision", Ottawa-Carleton Institute for Computer Science School Of Information Technology and Engineering University of Ottawa.

[25] Lafferty, J., McCallum, A. ve Pereira, F., (2001). "Conditional random fields: probabilistic models for segmenting and labeling sequence data", International Conference on Machine Learning.

[26] Cormen, T.H., Leiserson, C.E., ve Rivest, R.L., (1990), "Introduction to Algorithms", MIT Press/MCGraw-Hill.

[27] Bayraktar, Ö. ve Temizel, T.T., (2008)."Person Name Extraction From Turkish Financial News Text Using Local Grammer-Based Approach", ISCIS Computer and Information Sciences.

[28] Ratnaparkhi, A., (1997). "A simple introduction to maximum entropy models for natural language processing",Washington : Institute for Research in Cognitive Science, University of Pennsylvania.

[29] Douthat, A., (1998). "The Message Understanding Conference Scoring Software User's Manual", Message Understanding Conference.

[30] Cheng, Y., Sun, C., Lin, L., Liu, Y., (2010). "A Comparison Study of Conditional Random Fields Toolkits", ICIC 2010, CCIS 93.

[31] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., (1999)."Performance Measures For Information Extraction", DARPA Broadcast News Workshop.

EK-A

KİŞİ BİLGİ KUTUSU ŞABLONLARI

Asker bilgi kutusu

Astronot bilgi kutusu

Basketbolcu bilgi kutusu

Beyzbolcu bilgi kutusu

Bilardo oyuncusu bilgi kutusu

Bilim adamı bilgi kutusu

Bisikletçi bilgi kutusu

Boksör bilgi kutusu

Buz patencisi bilgi kutusu

Casus bilgi kutusu

F1 Sürücü Kaydı

Filozof bilgi kutusu

Futbolcu bilgi kutusu

Gazeteci bilgi kutusu

Go oyuncusu

Golf oyuncusu bilgi kutusu

Güreşçi bilgi kutusu

Hakem bilgi kutusu

Halife bilgi kutusu

Hanedan bilgi kutusu

Hükümdar bilgi kutusu

Jimnastikçi bilgi kutusu

Kişi bilgi kutusu

Kişi künyesi

Komedyen bilgi kutusu

Korsan bilgi kutusu

Kraliyet bilgi kutusu

Makam sahibi bilgi kutusu

Manken bilgi kutusu

Mimar bilgi kutusu

Motosiklet sürücüsü bilgi kutusu

Müslüman bilgin bilgi kutusu

Müzik sanatçısı bilgi kutusu

NBA oyuncusu bilgi kutusu

NHL oyuncusu bilgi kutusu

Oyuncu bilgi kutusu

Porno yıldızı bilgi kutusu

Red Bull Air Race pilotu bilgi kutusu

Sanatçı bilgi kutusu

Satranç oyuncusu bilgi kutusu

Seri katil bilgi kutusu

Snooker Oyuncusu Bilgi Kutusu

Sporcu bilgi kutusu

Sunucu bilgi kutusu

Süpersport sürücüsü

Tarihçi bilgi kutusu

Tenisçi bilgi kutusu

Türkücü bilgi kutusu

Voleybolcu bilgi kutusu

WRC yarışçısı bilgi kutusu

Yazar bilgi kutusu

Yüzücü bilgi kutusu

Çizgi roman yaratıcısı bilgi kutusu

İnternet ünlüsü bilgi kutusu

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı :Canan GİRGİN

Doğum Tarihi ve Yeri :1983iZMİR

Yabancı Dili :İngilizce

E-posta :canankragoz@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	Bilgisayar Mühendisliği	Ege Üniversitesi	2006
Lise	Fen Bilimleri	Uşak ODAL Anadolu Lisesi	2001

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2010	TUBITAK - BTE	Araştırmacı

YAYINLARI

Bildiri

1.Kredi Kartı Başvurularının Değerlendirilmesi için Uzman Sistem Gerçekleştirimi

– Akademik Bilişim 2013

2. Business Model Canvas Perspective on Big Data Applications

–IEEE BigData 2013

3. Şartlı Rastgele Alanlar ile Türkçe Wikipedia Sayfalarından Semantik İlişkilerin Çıkarılması

- SIU 2014

4. Dil Tabanlı Web İçeriklerinin Çekilmesi

- SIU 2014