itüdergisi/***d***
mühendislik
Cilt: 10, Sayı: 3, 27-36
Haziran 2011

Eylem çıkarımı ve varlık tanıma için ontoloji tabanlı bilgi çıkarımı ve belge yapı analizinin tümleştirilmesi

Şerif ADALI[*1], A. Coşkun SÖNMEZ[*2]

[1]*İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, 34469, Ayazağa, İstanbul*
[2]*Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, 34349,Yıldız, İstanbul*

Özet

Bu çalışmada, Türkçe belgelerin otomatik olarak işlenmesi ve bu belgelerden bilgi çıkarımı için ontoloji tabanlı bilgi çıkarımı ve belge yapı analizi teknikleri bir arada kullanılmıştır. Geleneksel bilgi çıkarımı sistemleri giriş metnini sıralı kelimeler olarak ele almakta iken, önerilen mimari belge yapı şablonlarının ve belge modellerinin sağladığı bilgilerden faydalanmaktadır. Bu özelliklere ek olarak, belgenin doğruluğunu sınamak için, belgede yer alan varlıklar arasındaki ilişkiler sınanmakta ve çıkarımı yapılmış varlıklar ile gerçek veriler karşılaştırılmaktadır. (Örnek: Müşteri veritabanı). Önerilen yaklaşım, Türkçe için biçimbirimsel analiz modülü, belge yapı analiz modülü ve çıkarım ontolojisi içermektedir. Yüksek miktarda dilbilimsel şablona dayalı çalışan bilgi çıkarım sistemlerinin aksine, çıkarım ontolojisi kullanılarak bilgi çıkarımı için sadece alan kavramı tanımları yeterli olmaktadır. Türkçe'de öğeler cümlenin anlamını bozmadan serbestçe yer değiştirebilmektedir. Bu nedenle kullanılan ontoloji tabanlı ayrıştırıcı ile çıkarımı yapılması istenilen varlıkların cümle içindeki pozisyonundan bağımsız olarak bulunması hedeflenmiştir. Test belgeleri yazılı bankacılık talimatlarını, sık uçanlar e-postlarını ve öğrenci dilekçelerini içermektedir. Bu belgeler serbest metin, tablolu, listeli ve maddesel yapıda veriler içermektedir. Deneysel sonuçlar önerilen mimarinin kısıtlı belge alanları için, belge modeli tanıma, varlıkların ve alan eylemlerinin çıkarımı konularında yüksek başarı elde ettiğini göstermiştir.

Anahtar Kelimeler: *Ontoloji tabanlı bilgi çıkarımı, belge yapı analizi, varlık tanıma, doğal dil anlama.*

[*]Yazışmaların yapılacağı yazar: Şerif ADALI. serifadali@yahoo.com; Tel: (216) 379 58 46.
Bu makale, birinci yazar tarafından İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisiği Bölümü ve Mühendisliği Programında tamamlanmış olan "Türkçe belgelerden bilgi çıkarımı için bir tümleşik mimari " adlı doktora tezinden hazırlanmıştır. Makale metni 25.12.2009 tarihinde dergiye ulaşmış, 05.02.2010 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 31.08.2011 tarihine kadar dergiye gönderilmelidir.
Bu makaleye "Adalı, Ş., Sönmez, A Ç., (2011) 'Eylem çıkarımı ve varlık tanıma için ontoloji tabanlı bilgi çıkarımı ve belge yapı analizinin tümleştirilmesi', İTÜ Dergisi/D Mühendislik, 10: 3, 27-36" şeklinde atıf yapabilirsiniz.

Integrating ontology based information extraction and document structure analysis for event extraction and entity recognition

Extended abstract

This study covers research activity in the field of automatic processing and event extraction from documents in Turkish. Proposed approach benefits valuable hints provided by document structure analysis for extracting information. Approach checks entities and relations of entities across document and verifies them by using relational database integration rules that are defined for each domain event. It contains a morphological analyzer for Turkish, a document structure analyzer and an extraction ontology.

Even though there has been an on-going effort for eliminating free-formatted text documents, certain forms of communication continue to be completely unstructured such as fax and e-mail. Proposed approach benefits extraction ontology, where currently available parsing approaches do not use ontology very effectively, mostly depending on rule-based or statistical parsers. Ontology based IE increases portability and scalability of an IE system. Proposed approach requires only extraction concepts when compared to information extraction systems that rely on large set of linguistic patterns. Proposed architecture is tested on a set of 3000 documents in 3 different domains, including data in tabular, list and itemized form. Proposed architecture has an F-Score 99% for extracting information from financial documents, frequent flyer emails and student written letters. Experimental results indicate that it obtained a high performance for detecting document domain, document model, entities and domain events.

Main IE tasks can be listed as extracting entities, extracting pre-specified events, extracting relations between entities and events. Experimental results indicate that document structure analysis and ontology based IE techniques mutually benefit each other. In proposed architecture, input document is treated as a combination of document model and event concept, where entities within document are cross-related to each other. Approach determines document model by locating document blocks, and using document models it determines the document domain. Approach also verifies document model by detecting the domain event in the input document. Using document structure analysis approach also determines some of the unknown entity types. Experimental results also indicate that approach successfully locates data in tabular, list and itemized form. For detecting data in tabular or list form it depends on a specific set of titles called "Descriptor Titles" which refer to column or row headers in tables or lists. In Turkish, morphological analysis is more complex when compared to languages like English, because it has agglutinative morphology. Due to the fact that Turkish is a free constituent language, proposed approach focuses on locating domain specific concepts in a sentence using the proposed "Concept Zoning" technique.

In case approach detects an unknown triggering verb, using concept similarity calculations, it determines the closest matching event in the extraction ontology. This case is especially useful during system development phase. Benefiting this feature a known domain event with an unknown triggering verb or an unknown domain event made up of known domain concepts can be determined. Approach also detects connected actions. It treats each action as seperate events however, in practice certain events require another event to make sense and its triggering verb contains specific morphological features.

Main contributions involved in this thesis are listed as the following items:

- *Test the effect of proposed "Concept Zoning" technique and document layout analysis on entity recognition and event extraction.*
- *Develop an ontology editor for designing domain concepts and events for extracting domain specific events in different document domains.*
- *Validating extracted information by using both relational database integration and document model validation, and benefit this integration for determining unknown entity types.*

Keywords: *Ontology based information extraction, document structure analysis, entity recognition, natural language understanding.*

Giriş

Bilgi çıkarımı doğal dil metinlerden belirli bilgilerin çıkartılması işlemine verilen isimdir. Bu nedenle bilgi çıkarımı, sınırlı doğal dil anlama olarak da nitelendirilmektedir. Bilgi çıkarımında en önemli problemlerden birisi çıkarımı yapılmış bilgilerin belirsizlik içerebilmesidir. Bu çalışmanın ana amacı Türkçe metin içeren belgelerin tanınması ve bu belgelerden önceden belirlenmiş çeşitli bilgi ve eylemlerin çıkartılması için belge yapı analizi ve ontoloji tabanlı bilgi çıkarımı yöntemlerinin bir arada kullanılmasıdır. Bilgi çıkarımı için ontoloji kullanılmasındaki amaç, en az emek ile bir belge alanına ait eylemlerin tanımlanabilmesini sağlamaktadır. Belgelerden otomatik olmadan yapılan bilgi çıkarımı işlemi hem zaman kaybettiren hem de maliyetli bir iş iken, otomatik belge tanıma ve anlama sistemleri sayesinde hız, tutarlılık arttırılmakta ve işletme maliyetleri azaltılmaktadır.

Otomatik belge anlama ve içerdikleri eylemleri gerçekleştirme üzerine çeşitli akademik ve ticari çalışmalar gerçekleştirilmiştir. Wee vd. (1999) finansal bilgilerden bilgi çıkarımı için genel bir bilgi çıkarımı mimarisi adlı çalışmayı yaptılar. Çalışmalarında alan ontolojisi ve çerçeve tanımlama yöntemleri kullanmışlardır. Bir diğer önemli çalışma ise Lytinen ve Gershman (1993) tarafından geliştirilen ticari bir bilgi çıkarım sistemi olan ATRANS sistemidir. Bu sistem bankalar arası para transferi mesajlarını otomatik olarak işleyen bir sistemdir. Ontoloji kullanarak bilgi çıkarımı konusunda çeşitli çalışmalar yapılmıştır. Örneğin, Yıldız ve Miksch (2007) ontoloji temel alarak bilgi çıkarım kurallarının otomatik olarak üretilmesi konusu üzerinde çalışmışlardır. Embley (1998) çıkarım ontolojisi kullanarak yapısal olmayan belgelerden yapısal bilgilerin çıkartılması konusunda bir çalışma yapmıştır. Çalışmasında ontolojiyi kullanarak belgelerde geçen sabit ve anahtar kelimelerin çıkarımını yapacak olan kuralları oluşturmaya çalışmıştır. Temizsoy ve Çiçekçi (1998) ise ontoloji kullanarak Türkçe cümlelerin ayrıştırılması üzerine çalışmışlardır ve özellikle biçimbirimsel özelliklerden faydalanmışlardır. Bununla birlikte Tür, istatiksel yöntemler kullanarak konu sınırı olmadan, temel bilgi çıkarımı konuları üzerine bir doktora tezi sunmuştur. Tür (2000) çalışmasında sesli harf düzeltme, cümle ayrıştırma, konu sınıflandırma ve varlık çıkarımı konuları üzerine çalışmıştır. Bir belgeden belirli bir eylemin çıkartılması işleminde, bilgi çıkarımı yöntemlerine ek olarak belgenin yapısal olarak incelenmesi gerekmektedir. Belgenin yapı analizi kullanılarak, belgenin tanımlanabilir yapısal blokları ile serbest metin blokları birbirinden ayrılmaktadır. Buna ek olarak, belgenin bloklarında belirlenen anahtar değerler kullanılarak çıkarımı yapılmış bilgiler kontrol edilmekte ve ayrıca tipi belirlenememiş varlıkların tipleri bulunmaktadır. Bu yöntem ile ontolojiye yeni keşfedilmiş varlıklar eklenebilmektedir. Önerilen yaklaşım el ile ya da otomatik olarak olası tüm çıkarım şablonlarını yaratmak yerine, çıkarım ontolojisinde tanımlı olan alan kavramlarını üzerinde çalışılan metin bloğunun içerisinde pozisyonlarından bağımsız olarak işaretlemeye çalışmaktadır. Çalışmada kullanılan yaklaşım kural tabanlı bir yaklaşım olup sınırlı sayıda belge alanı ve alan eylemi için tasarlanmıştır.

Hui vd. (1997) geliştirdikleri sistem ile faks mesajlarının içerdiği bilgileri mantıksal ve yapısal sınıflarına göre tanıyarak, belgeleri depolamak için gerekli olan depolama alanını azaltmaya çalışmışlardır. Sahin ve Sawyer (1989) "Akıllı Bankacılık Sistemi" adını verdikleri kural tabanlı bir sistem geliştirmişlerdir. Bu sistem bir teleks mesajını ortalama 30 saniyede işlemektedir. Soderland (1997) alana ait doğal dil işleme için metin analiz kurallarını öğrenme üzerine çalışmıştır. “Wall Street Journal” makalelerinde yer alan üst düzey yöneticilerin görev değişikliklerini içeren belgeleri inceleyip bu belgelerden yer değiştiren yönetici bilgilerini çıkarmaya çalışmıştır. Bir diğer test alanı olarak hastahane taburcu raporlarından teşhis ve tedavi bilgilerinin çıkarılması üzerine çalışmıştır. Vargas-Vera ve Celjuska (2008) hikâyelerden eylemlerin çıkarılması üzerine bir sistem üzerinde çalışmışlardır. Sistem hem hikâyeleri sınıflandırmakta hem de çıkarımı yapılan bilgiler ile ontolojiyi zenginleştirmektedir. Wu vd. (2003) alan ontolojisi kullanarak, alan eylemlerini tanımlama ve çıkarma konusu üzerine çalışmıştır. Çalışmalarının deneysel sonuçları, otomatik eylem çıkarımı ve ontoloji geliştirmenin metin sınıflandırma ve

detaylı bilgi işleme için kullanılabileceğini göstermiştir.

Yangarber (2001) ise taşınabilirlik konusu üzerine yoğunlaşmıştır. Öncelikle elle çıkarım şablonlarının oluşturulması için bir dizi tasarım aracı geliştirmiştir. Kullanıcı bu araçları kullanarak eğitim belgelerinden bir grup örnek cümleyi seçmekte ve bu cümleler ile alan eylemleri arasındaki bağlantıları kurmaktadır. İkinci bir çalışmada ise, sadece en temel şablonları içeren küçük bir şablon kümesini temel alarak kademeli öğrenme yöntemleri kullanarak sistem yeni şablonları, sınıfları ve ilişkili terimleri öğrenmektedir.

Türkçe

Türkçe Avrupa ve Asya'da geniş bir coğrafyaya yayılmış, dünyada en çok konuşulan altıncı dildir. Ural-Altay dilleri ailesine bağlı olan Altay kolunun bir üyesi olan Türkçe bitişken bir dil olup, sözcüklerin sonlarına ard arda çekim ve türetim ekleri eklenerek yüzlerce farklı sözcük oluşturmak mümkündür (Oflazer, 1995). Bu yapısı nedeniyle Türkçe'nin biçimbirimsel analizi diğer dillerle karşılaştırıldığında çok daha karmaşıktır. Örneğin "hesabıma" kelimesi İngilizce'de 3 ayrı kelime ("to my account") ile ifade edilmektedir. Türkçe'de genel olarak cümleyi oluşturan sözcüklerin dizilişleri gözönüne alındığında çoğunlukla özne, nesne, yüklem kalıbına uymakla birlikte öğeler cümle içerisinde serbestçe yer değiştirebilmektedir. Test belgeleri yüklemin sonda bulunduğu kurallı cümleler içeren çeşitli metin alanlarına ait talimatlar içermektedir. Türkçe'nin kelime sıralamasına örnek bir cümle Tablo 1'de gösterilmiştir.

Tablo 1. Türkçe'de kelime sıralaması

Dil	Metin		
Türkçe:	Çocuk	su	içiyor.
İngilizce:	(The) child	water	is drinking.

Mimari detayları

Yaklaşım çeşitli doğal dil işleme işleminin ard arda gerçekleştirildiği bir işlem hattından oluşmaktadır. Öncelikle giriş belgesi simge olarak adlandırılan en küçük birimlere ayrılır. Bu işlemi takip eden adımlar sözlük taraması, alan ve eylem tanıma, belge yapı analizi, biçimbirimsel analiz, giriş metnini kavramlara bölme, kavram özelliklerini tanıma, listeli, tablolu ve maddesel yapıdaki bilgilerin çıkartılması, belirsiz varlık tiplerinin bulunması ve çıkarımı yapılmış bilgilerin belge modeline ve eylemine tümleşik olarak gerçek veriler ile (Örnek: CRM Veritabanı) karşılaştırılması olarak sıralanmaktadır.

Belgeyi oluşturan bloklarının belirlenmesi belge modeli/alanının, tablo, liste ve serbest metin bölgelerinin belirlenmesi ve çözümlenmesi açısından önem taşımaktadır. Bir metin içinde yapısı daha önceden tanımlanmış belge blok şablonlarına uyan ya da yakın olan metin bölgeleri işaretlenmekte, daha sonra bu bloklardan faydalanılarak belge modeli ve alanı belirlenmektedir. Bu işlem alan süzme olarak adlandırılmaktadır. Örneğin, bir bankacılık belgesinde banka ve şube bilgilerini içeren bir belge bloğu bulunurken, sık uçanlar e-postlarında havayolu firması ya da sık uçanlar program ismi içeren bir belge bloğu bulunmaktadır. Sistem bulabildiği şablonların dışında kalan alanı ise serbest metin bölgesi olarak işaretlemektedir. Bu bölgede bir paragraf, listeli, tablolu yada maddesel yapıda veri bulunabilir. Benzer olarak Klink vd. (2000) bir belgenin yapısal analizinde hem yapı bilgisinden hem de metinsel bazı özelliklerden yaralanmıştır. Ding vd. (1999) ise şablon madenciliği kullanarak makalelerden kaynak bilgilerinin çıkartılması konusunu üzerine çalışmışlardır. Önişleme işlemleri bilgi çıkarımı sistemlerinde önemli bir rol oynamaktadır. Bunlar bazı simgelerin aralarına gerekli boşlukların yerleştirilmesi ("100YTL"), metin olarak yazılmış değerlerin sayısal karşılıkları ile değiştirilmesi örnek olarak gösterilebilir. Örnek olarak "Bugün" simgesinin günün tarihi ile, "Sıfır" simgesinin ise sayısal "0" değeri ile değiştirilmesi gösterilebilir.

Her belge alanı için olası belge blok şablonları kural tabanında saklanmaktadır. Örneğin Tablo 2'de bir banka ismi, bir şube ismi ve şehir isminden oluşan bir belge bloğu tanımı gösterilmiştir. Bu blok, genellikle bankacılık.

Tablo 2. Örnek belge blok şablonu tanımı

Blok Tipi	Alanları
Ana Şube	Banka Adı→Şube Adı→ Şehir İsmi

Talimatlarında, faks başlığı bloğundan hemen sonra yer alan ve işlemin gerçekleştiği ana şubeyi işaret eden belge bloğudur. Tanımlanan bu belge blok şablonlarının çeşitli sıralarda ve varyasyonlarda bir araya gelmesi ile alana özel belge modelleri oluşmaktadır. Tablo 3'de bankacılık belgelerinde kullanılan genel belge modeli gösterilmiştir. Belge, müşteri faks bilgilerini içeren bir faks bloğu ile başlayarak, ardından şube bilgileri, talimatın yer aldığı düz metin bloğu ile devam etmekte ve irtibat bilgilerinin geçtiği bölüm ile son bulmaktadır.

Tablo 3. Örnek belge modeli

Belge Modeli	Belge Blokları
Bankacılık	Faks Başlığı→Ana Şube→ Serbest Metin → Şirket Bilgileri

Bununla birlikte önerilen mimari alana bağımlı ve alandan bağımsız varlık tiplerinin tanımlanabilmesi için bir kural tabanına sahiptir. Bu temel varlık tipleri tarih, saat, hesap numarası, kredi kartı numarası, telefon numarası gibi veri tiplerini içermektedir. Genellikle bu tip varlıkların tanınması için bazı basit metin şablonlarının tanımlanması yeterli olmaktadır. Tablo 4'de gg.aa.yyyy (gün-ay-yıl) yapısındaki tarih tipini tanımlamak için gerekli veri yapısı gösterilmiştir. Yaklaşım öncelikle giriş belgesinde bulunan tüm harfleri "C" ve tüm sayısal değerleri ise "N" olarak işaretleyerek giriş metnine karşılık gelen bir eşleşme dosyası oluşturur. Daha sonra ise sistemde tanımlı olan temel varlık tiplerini eşleşme dosyası içerisinde bulmaya çalışır. Tablo 4'e ek olarak sistem diğer alternatif yapıları da tanıyabilmektedir. (23 OCT 08, OCT. 23 2008, EKİ 23 2008, EKİ-23-2008)

Tablo 4. Temel varlık tipi tanımlama

Tipi	Yapısı	Örnek Veri
Tarih	NN.NN.NNNN	01.03.2009
Tarih	NN/NN/NNNN	16/12/2006
Tarih	NN-NN-NNNN	22-05-2009

Bir sonraki adım ise bulunan bu varlık tipinin Tablo 5'de belirtilmiş olan kısıtlama kurallarına uyup uymadığı kontrol edilir.

Tablo 5. Varlık tipi kısıtlamaları

Endeks	Tipi	Denklem	Denklem Değeri	Değer Aralığı
1-2	Metin	İçinde	İkiHaneliGün	
4-5	Metin	İçinde	İkiHaneliAy	
7-10	Sayısal	Aralık	Yıl	>=1900 <=2100

Temel varlık tiplerinin belirlenmesinden sonra kademeli artan sözlük tarama işlemi mevcut simge ve bu simgenin sağında yer alan komşu simgelerin birleşiminden oluşan katarın alan ontolojisinde taranması ile gerçekleştirilir. Bu işlem katara karşılık en az bir değer döndüğü sürece devam etmektedir. Bazı durumlarda katar birden fazla varlık tipine işaret edebilir. Bu değerlere aday değerler denilmektedir. Örneğin "TL" aynı belge alanında belli durumlarda kur tipi olabilirken, bir diğer durumda banka hesap tipi olabilmektedir. Bu tip belirsizlikler ise çıkarım ontolojisi sayesinde ortadan kaldırılmaktadır. Şekil 2'de örnek olarak "Para Transferi" eylemine ait kavram tanımlama ekranının bir bölümü gösterilmiştir. Eylem temel olarak 3 kavramdan oluşmaktadır. Bunlar kaynak hesap, hedef hesap ve para miktarı kavramlarıdır. Şekil 2'de görüldüğü gibi hedef hesap(" Destionation Account") kavramı hesap numarası, hesap tipi, yer ve sahip (Account Number, Account Type, Location, Owner) gibi çeşitli özelliklere sahiptir. Örneğin Şekil 2'deki tanımlamada, hesabın sahibi bir şirket yada kişi olabilmektedir. Alan eylemlerinin, kavramlarının, ilişkili oldukları örneklerin ve kuralların tek bir kullanıcı arabiriminde düzenlenebilmesi için görsel "kavram tasarım" modülü geliştirilmiştir.

+hasSourceAccount
-hasDestinationAccount
+Property: AccountNumber
+Property: AccountType
+Property: Location
-Property: Owner
+Company
+Person
+hasAmount

Şekil 2. Çıkarım kavram tanımları

Çıkarım kavramları ve kavram özeliklerini doğru şekilde bulabilmek için serbest metin bölümü eylem kavramlarını kapsayan alanlara bölünmektedir. Öncelikle eylem tetikleyici fiilin belirlenmesi gerekmektedir. Belge modelinin belirlenmesi belge alanının belirlenmesinde büyük önem taşımaktadır. Bu yöntem ile sadece ilgili alana ait tetikleyici fiiller ve kavramlar taranmaktadır. Ek olarak aynı tetikleyici fiil'in birden fazla alanda ortak kullanılması durumunda belge analizi kullanılarak alan süzme yapılmaktadır. Ayrıca belge modelinde en az bir bloğunun belge alanına özel bir blok olması nedeniyle alan süzme işleminde başarı artmaktadır. Şekil 3'de bir cümlenin kavramlara ve daha sonra bu kavramları oluşturan özelliklere bölünmesi işlemi görselleştirilmiştir.

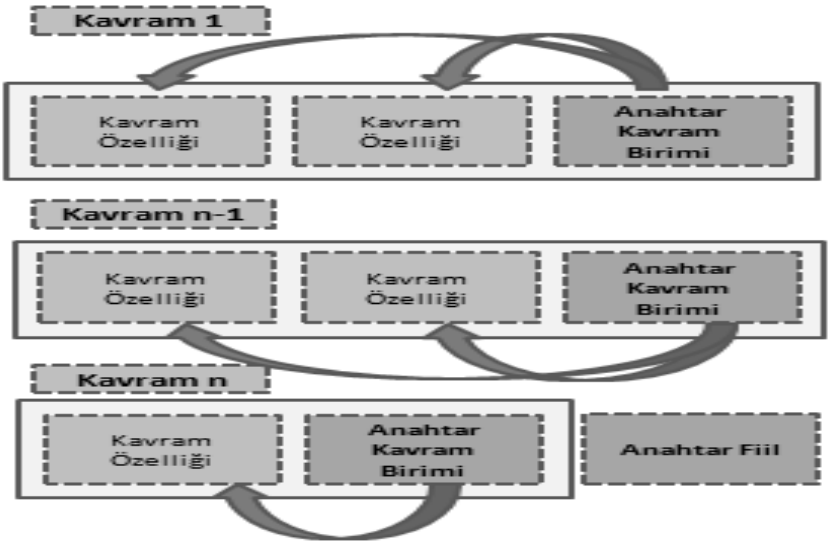
Eylemi tetikleyen fiil'in pozisyonu göz önüne alınarak serbest metin bölgesi eylem kavramlarına bölünür. Ancak sistemde tanımlanmamış bir eylem tetikleyici fiil ile karşılaşılması durumunda, yaklaşım ontoloji'de tanımlı olan anahtar kavram simgelerinden faydalanmaktadır. Her bir eylem için eşleşen maksimum anahtar kavram ve özellik adedi hesaplanmakta ve kullanıcıya onaylatıldıktan sonra bu fiil ilgili eylem ile ilişkilendirilmektedir. Şekil 4'de kavram bölme ile çözümlenmiş bir para transferi talimatı örneği gösterilmiştir. Belgenin tanımlanabilir blokları (Header, Branch, Ending) belirlenmiş, geriye kalan alan ise serbest metin bölgesi (Free) olarak işaretlenmiştir. Çıkarım ontolojisi ise serbest metin içerisinde alan eylemini bularak, bu eylemi oluşturan kavram ve özelliklerini işaretlemektedir.

Bunun için Tablo 6'da gösterilen anahtar kavram birimleri tanımlanmalıdır. Bu işlemde, biçimbirimsel analizi yapılmış tüm simgelerden anahtar kavram özelliği olanlar aşağıda belirtilen kurallara göre seçilir ve işaretlenir.

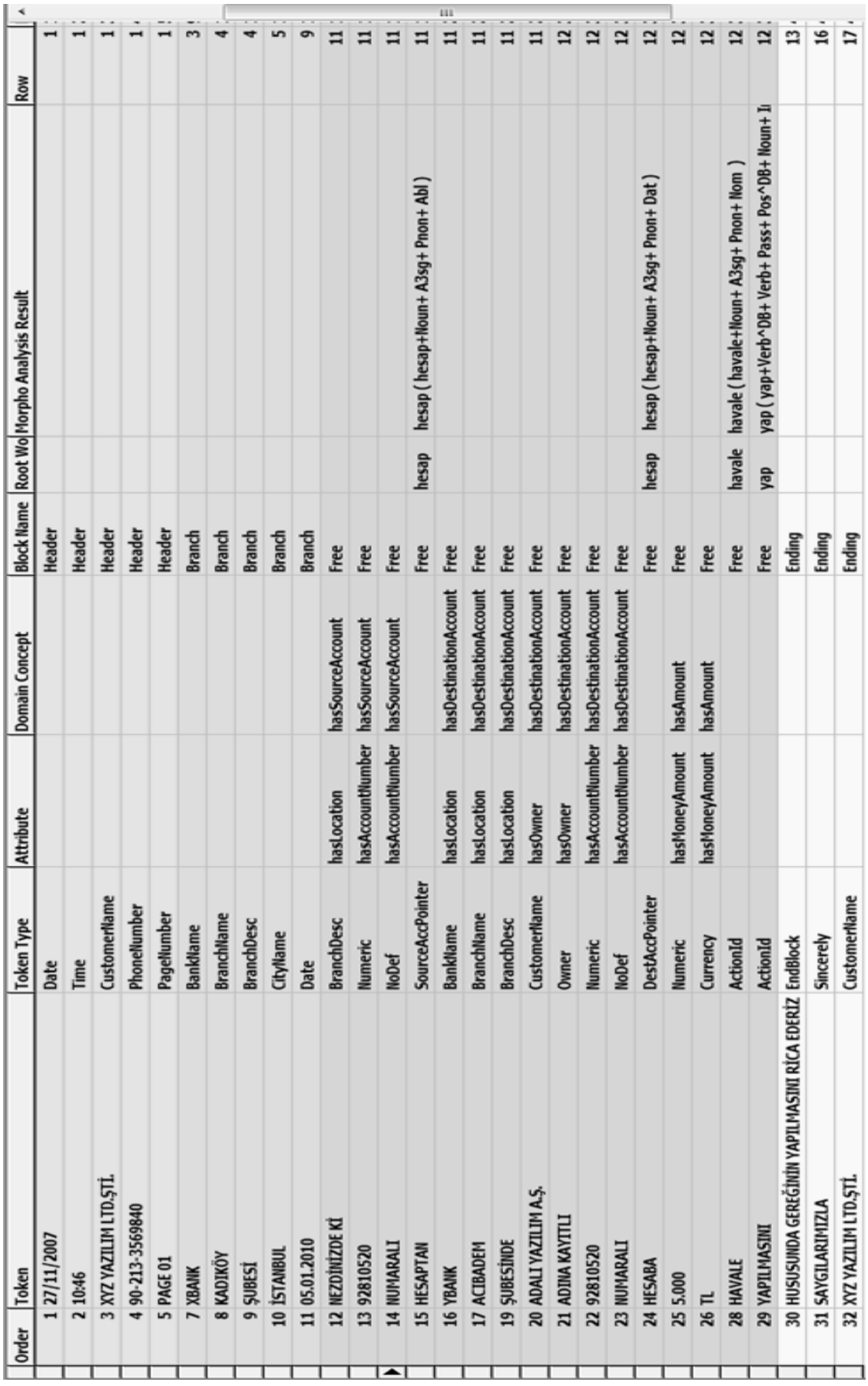
Tablo 6. Kavramlara bölme detayları

Kavram	Tipi	Kök Kelime	Biçim birimsel Detay
Kaynak Hesap	İsim	Hesap	-den hali
Hedef Hesap	İsim	Hesap	-e hali
Miktar	Kur Tipi		

Tablo 6'da yer alan hedef ve kaynak hesap kavramları ontolojide bulunan banka hesabı kavramı ile bağlantılıdır. Tanımlamalarda da görüleceği gibi her iki kavram alanının işaretlenebilmesi için kullanılan kök kelime "hesap" iken, kaynak hesap için anahtar kavram biriminin "-den" hali eki, hedef hesap için ise "-e" hali eki içermesi gerekmektedir. Bu özellik sayesinde kaynak hesap, hedef hesap ve para miktarı kavram alanları serbest olarak yer değiştirseler bile yaklaşım başarılı bir şekilde kavram alanlarını belirlemekte, ve kavram özelliklerini işaretlemektedir.



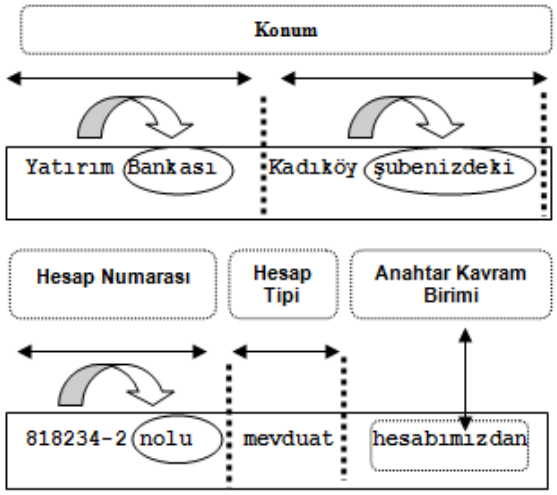
Şekil 3. Çıkarım ontolojisi kullanılarak kavramlara bölme



Order	Token	Token Type	Attribute	Domain Concept	Block Name	Root Wo	Morpho Analysis Result	Row
1	27/11/2007	Date			Header			1
2	10:46	Time			Header			1
3	XYZ YAZILIM LTD.ŞTİ.	CustomerName			Header			1
4	90-213-3569840	PhoneNumber			Header			1
5	PAGE 01	PageNumber			Header			1
7	XBANK	BankName			Branch			3
8	KADIKÖY	BranchName			Branch			4
9	ŞUBESİ	BranchDesc			Branch			4
10	İSTANBUL	CityName			Branch			5
11	05.01.2010	Date			Branch			9
12	NEZDİNİZDE Kİ	BranchDesc	hasLocation	hasSourceAccount	Free			11
13	92810520	Numeric	hasAccountNumber	hasSourceAccount	Free			11
14	NUMARALI	NoDef	hasAccountNumber	hasSourceAccount	Free			11
15	HESAPTAN	SourceAccPointer			Free	hesap	hesap (hesap+Noun+ A3sg+ Pnon+ Abl)	11
16	YBANK	BankName	hasLocation	hasDestinationAccount	Free			11
17	ACIBADEM	BranchName	hasLocation	hasDestinationAccount	Free			11
19	ŞUBESİNDE	BranchDesc	hasLocation	hasDestinationAccount	Free			11
20	ADALI YAZILIM A.Ş.	CustomerName	hasOwner	hasDestinationAccount	Free			11
21	ADINA KAYITLI	Owner	hasOwner	hasDestinationAccount	Free			12
22	92810520	Numeric	hasAccountNumber	hasDestinationAccount	Free			12
23	NUMARALI	NoDef	hasAccountNumber	hasDestinationAccount	Free			12
24	HESABA	DestAccPointer			Free	hesap	hesap (hesap+Noun+ A3sg+ Pnon+ Dat)	12
25	5.000	Numeric	hasMoneyAmount	hasAmount	Free			12
26	TL	Currency	hasMoneyAmount	hasAmount	Free			12
28	HAVALE	ActionId			Free	havale	havale (havale+Noun+ A3sg+ Pnon+ Nom)	12
29	YAPILMASINI	ActionId			Free	yap	yap (yap+Verb^DB+ Verb+ Pass+ Pos^DB+ Noun+ Il	12
30	HUSUSUNDA GEREĞİNİN YAPILMASINI RİCA EDERİZ	EndBlock			Ending			13
31	SAYGILARIMIZLA	Sincerely			Ending			16
32	XYZ YAZILIM LTD.ŞTİ.	CustomerName			Ending			17

Şekil 4. Anahtar birimler ile kavramlara bölme senaryo örneği

Ontoloji tabanlı bilgi çıkarımı sayesinde, çıkarımı yapılan bilgilerin ontolojide hangi kavramlara ve özelliklere karşılık geldiği ve bu özellikleri oluşturan varlıkların detaylı tipleri görselleştirilmiştir. Eğer kavram alanı içerisinde bulunduğu halde ontoloji tarafından çözümlenemeyen bir simge var ise, sistem bu simgelerin tiplerini de kullanıcıya sorarak sisteme tanıtmasını istemektedir (Örnek: hesap tipi).



Şekil 5. Detaylı kavram analizi

Ancak bazı anahtar kavram birimlerini işaretlemek için sözlüksel kaynaklar yeterli olabilmektedir. Örneğin para kavramının bulunduğu alanın işaretlenmesi için kur tipi anahtar kavram birimi olarak kullanılmaktadır (Örnek: 100 YTL). Bazı kavram özellikleri serbest metin, tablo yada listeli yapı dışında diğer belge bloklarında bulunabilir (Örnek: faks başlığı bloğu). Belge yapı analizi sayesinde, yaklaşım tüm varlıklara nesneye yönelik bir yaklaşımla erişmektedir. Örneğin faksı gönderen firma ismi, faks başlığı ("Header") bloğunda bulunan ("Customer Name") değerine eşittir. Bununla birlikte belge yapı analizinden faydalanılarak eylemin parçası olan ancak serbest metin dışındaki bloklarda bulunan verilerin hangi bloklardan alınacağı kullanıcı tarafından tanımlanabilmektedir. Eğer sistem ontoloji ve belge modelini kullanmasına rağmen bazı kavram özelliklerine ulaşamıyorsa o zaman anahtar değer(ler) kullanarak gerçek verilere erişmekte, hem belirlenemeyen varlık tipleri işaretlemekte hem de çıkarımı yapılmış veriler kontrol edilmektedir. Yaklaşım aynı zamanda bağlantılı eylemlerin çıkarımını da sağlamaktadır. Örnek olarak belirli miktarda dövizin başka bir kur'a çevrilerek karşılığının bir hesaba aktarılması eylemi gösterilebilir. Yaklaşım "Kur çevirme-Hesaba Aktarma" işlemini farklı iki eylem olarak ele almaktadır. Benzer olarak "fon bozdurma" eylemi de tek başına değil ancak bir başka eylem ile birlikte kullanıldığında anlam kazanmaktadırlar. (Örnek: Fon Bozdurma-Hesaba Aktarma). Ayrıca bağlantılı eylemlerde, eylem tetikleyici fiiller (Örnek: "Çekilerek", "Çevrilerek", "Bozdurularak") özel biçimbirimsel özelliklere sahiptirler. Sistem bu yaklaşım ile farklı bağlantılı eylemleri (Örnek: Kur bozdurma-Fon satın alma) ek tanımlama yapmaya gerek olmadan çözümleyebilmektedir.

Listeli yapıların çözümlenmesi

Tablo ve listeli yapılı verilere genellikle eylem kavramlarının tamamının ya da bir bölümünün serbest metin alanının dışında bulunduğu durumlarda karşılaşılır. Örnek liste yapısı Şekil 6'da gösterilmiştir.

Bankanız nezdinde bulunan 85853726-164 no.lu ytl hesabımızdan aşağıda dökümü yapılan transferlerin yapılmasını bilgilerinize arz ederim.

ALICI :	ADALI YAZILIM SAN. A.Ş.
BANKA:	X BANK
ŞUBE ADI:	KADIKÖY ŞB.
HESAP NO:	48948778
MİKTARI:	200 TL

Şekil 6. Listeli yapı örneği

Kavram alanlarını bulabilmek için eylemi tetikleyen fiilin endeksi işaretlenir ve bu endeksten önce ve sonra yer alan tüm eylem kavramları belirlenir. Ancak tablo ya da liste yapılarının çözümlenmesinde tanımlayıcı etiketler kullanılmaktadır ("ALICI", "BANKA","ŞUBE", "HESAP NO", "TUTAR"). Sistemde bu etiketler eylem ve belge alanı bazında ilişkilendirilmiştir. Şekil 6'da görüldüğü gibi tanımlayıcı bir etiketi takiben bu etikete karşılık gelen varlık bulunmaktadır, ancak tablolu verilerde ise tanımlayıcı etiket ve varlıklar farklı satırlarda yer

almaktadır. Tablolu yapıda bir satırda tüm tanımlayıcı etiketler sıralanmakta bu etiketler tablonun her bir kolonuna karşılık gelmekte ve bu satırı takip eden satırlarda çözümlenecek olan varlıklar bulunmaktadır. Ancak hem listeli hem de tablolu yapılarda bir tanımlayıcı etikete karşılık birden fazla varlık değeri bulunabilir. Örneğin para miktarı etiketleri olan "MİKTAR:" ve "TUTAR:" dan sonra yaklaşım bir sayısal değer ve kur tipi aramaktadır (Örnek: 500 YTL). Etiketten bağımsız çalışması durumda hatalı bilgi çıkarımları meydana gelmektedir. Örneğin listeli yapıda bir firma isminden önce "GÖNDEREN" veya "ALICI" tanımlayıcı etiketine bakılması, eylemin doğru tanınmasında önem taşımaktadır.

Tablolu yapıların çözümlenmesi

Yöntem metin içerisinde bulunan ilk tanımlayıcı etikete gider, bu etiketi takip eden başka etiketler olup olmadığını kontrol eder ve bulabildiği tüm etiketleri işaretler. Bu işlemin ardından son etiketi takip eden ilk varlığın tipini kayıt eder ve serbest metin bölgesini satırlara bölmek için bu varlık tipini kullanır. Daha sonra ilk tanımlayıcı etiket'e gider ve etiketlere karşılık gelen varlık tiplerini işaretler. Ancak hem tablolu hem de listeli yapıdaki bilgilerin hatasız olarak çıkarılması için olası tüm tanımlayıcı etiketlerin kural tabanında bulunması gerekmektedir. Basit bir tablolu yapı örneği Şekil 7'de gösterilmiştir.

ALICI	BANKA	ŞUBE
SERKAN ADALI	EKONOMİ BANKASI	KADIKÖY ŞUBESİ
ALİ MUTLU	YATIRIM BANKASI	ACIBADEM ŞUBESİ

Şekil 7. Tablolu yapı örneği

Maddeseleştirilmiş eylemler

Bazı test belgeleri maddeselleştirilmiş eylem listeleri içermektedir. Tablolu yapıya benzer olarak her bir eylem için ayrı çıkarım çıktısı oluşturulmaktadır. Şekil 8'de görüldüğü gibi, bir transfer eyleminde kaynak hesap aynı olmak üzere hedef hesap ve para miktarları maddeselleştirilmiş yapıda belirtilmiştir.

Şubeniz nezdinizdeki 839001-2 nolu hesabımızdan aşağıdakı havalelerin yapılmasını rica ederiz.

1. Bereketli bank karşıyaka şubesi'nde bulunan 236795-1 nolu hesabımıza 1000 YTL aktarılmasını

2. Akdeniz bankası suadiye şubesi'nde bulunan 415919-1 nolu hesabımıza 600 YTL aktarılmasını

Şekil 8. Maddeselleştirilmiş liste örneği

Deneysel çalışma sonuçları

Önerilen mimari toplam 3000 belge ile test edilmiştir. Test belgeleri yazılı bankacılık talimatlarını, sık uçanlar müşteri e-postlarını ve öğrenci dilekçelerini içermektedir. Farklı eylem ve belge modellerini içeren test belgeleri düz metin, tablo, liste ve maddesel yapıda veriler içermektedir. Tek bir eylem içeren belgenin bir kişisel bilgisayar üzerinde çözümlenme süresi ortalama 10 saniyedir. F-Ölçütü belge yapı analizi ve ontoloji tabanlı kavram bölme tekniği sayesinde, kısıtlı belge alanı ve eylemi tanıma için %99 olarak belirlenmiştir. Deneysel sonuçlar, başarılı şekilde belgelerin işlenmesi için belge blok tanımlamalarının, ontoloji çıkarım kavramları ve özelliklerinin, varlık tipleri ve örneklerinin (bankalar, şubeler, kurlar, meslekler, hesap tipler, enstitü isimleri, üniversite bölümleri, vs.) sistemde tanımlı olması gerektiğini göstermiştir. Çıkarımı yapılmış tüm verilerin %5'i, tetikleyici fiilin sistem tarafından bilinmediği halde eylemi oluşturan kavramların tanınması nedeniyle başarı ile çözümlenmiştir. Örneğin "EFT" eylemi, "Elektrik fon transferi yapılması", "Eft sistemi ile aktarılması", "E.F.T yapılması" gibi çeşitli fiillerle tetiklenmektedir. Serbest metin alanı dışındaki belge bloklarında bulunan varlık tipleri ise belge blok şablonlarından faydalanılarak başarı ile çözümlenmiştir. Bu yöntem ile çözümlenmiş varlıkların adedi tüm varlıkların %4'ü kadardır. Bu varlıklar genellikle bilinmeyen şirket uzantısı içeren bir şirket ismi, yabancı bir şirket ismi, yabancı bir şahıs ismi, bilinmeyen ad yada soyad, tanımlanmamış/kısaltma kullanılmış bir meslek adı, kısaltma ismi kullanılmış banka ismidir. Ancak testler sırasında sistemin bilgi çıkarımı yapamadığı bazı durum-

lar bulunmaktadır. Örneğin para miktarı kavramına karşılık gelen "100 YTL" değerleri yerine talimat metninde "YTL 100" ifadesinin kullanılması, para transferi eyleminde anahtar kavram birimi olan "Hesap" kelimesi yerine "adrese", "firmaya" kelimelerinin kullanılması yada bazı kavramların bir bölümünün çıkartılmasıdır. Yaklaşımın tablolu ve listeli yapıdaki varlıkları çıkarabilmesi için gerekli tanımlayıcı etiketler tanımlanmıştır. Ancak sayısı az olmakla birlikte bazı talimatlar hiç tanımlayıcı etiket içermemektedir. Etikete bağımlı olarak çalışması sistemin zayıf yönü olarak gösterilebilir. Ancak sistemin tanımlayıcı etiketlere bağlı olmadan çalışması durumunda aynı tipte iki varlığın eylemde hangi kavram özelliğine karşılık geldiğini bulma konusunda sorunlar oluşmaktadır. Örneğin listeli yapıda belirlenmiş iki farklı firma yada hesap numarasının etiketlere bakmadan ("GÖNDEREN FİRMA", "ALICI FİRMA ADI", "GÖNDEREN HESAP NO","ALICI HESAP NO") eylemdeki doğru karşılıkları bulunamamaktadır.

Sonuçlar

Deneysel sonuçlar da göstermiştir ki, kavram bölme yöntemi ve belge yapı analizi yöntemleri bir arada kullanılarak kısıtlı sayıda doküman modeli ve alan eylemi tanıma konusunda yaklaşım başarılı olmuştur.

Kaynaklar

Ding Y., Chowdhury G., Foo, S. (1999). Template Mining for the Extraction of Citation from Digital Documents, *Proceedings of the 2nd Asian Digital Library Conference*, 47–62, Taipei, Taiwan.

Embley D., Douglas D.M., Randy D.S. (1998). Ontology based Extraction and Structuring of Information From Data Rich Unstructured.

Documents Proceedings of the Conference on Information and Knowledge Management 52-59 Washington DC USA.

Hui S.C., Chan K.Y., Leung M.K., Qian G.Y. (1997). A Distributed Fax Messaging System, *Journal of Network and Computer Applications* **20**, 2, 171–190.

Klink S., Andreas D., Kieninger T. (2000). Document Structure Analysis Based on Layout and Text Analysis, *Proceedings of International Workshop on Document Analysis Systems*, 99–111, Rio de Janeiro, Brazil.

Lytinen S., Gershman A. (1993). ATRANS: Automatic processing of money transfer Messages, *Proceedings of the 5th National Conference of the American Association for Artificial Intelligence*, 93–99, Philadelphia, USA.

Oflazer K. (1995). Two-level Description of Turkish Morphology, *Literary and Linguistic Computing* **9**, 2, 137–148.

Sahin, K., Sawyer, R. K. (1989). The Intelligent Banking System, *Innovative applications of artificial intelligence*, MIT Press, 43-50.

Soderland S.G. (1997). Learning text analysis rules for domain-specific natural language processing, *Ph.D. Thesis*, University of Massachusetts, USA.

Temizsoy M.,Cicekci I. (1998). An Ontology-based Approach to Parsing Turkish Sentences, *LNCS, 1529*, 124–135, Springer, Heidelberg.

Tür G. (2000). A Statistical Information Extraction System for Turkish, *Ph.D. Thesis*, Bilkent University, Ankara, Turkey.

Vargas-Vera M.,Hall W., Keynes M. (2008). Ontology-driven Event Recognition on Stories, The Open University, United Kingdom.

Wee L.K.A.,Tong L.C., Tan C.L. (1999). A Generic Information Extraction Architecture for Financial Applications, *International Journal of Expert Systems with Applications* **16,** 4, 343–356.

Wu S., Tsai T., Hsu, W. (2003). Domain Event Extraction and Representation with Domain Ontology, *Proceedings of the IJCAI 2003 Workshop on Information Integration on the Web*, 33–38, Acapulco, Mexico.

Yangarber R. (2001). Scenario Customization for Information Extraction, *Ph.D. Thesis*, New York University, NYC, USA.

Yıldız B. , Miksch S. (2007). Motivating Ontology-driven Information Extraction, *Indian Statistical Institute Platinum Jubilee Conference Series*, 45–53, Kolkata, India.