Journal of Engineering and Natural Sciences
Mühendislik ve Fen Bilimleri Dergisi

Sigma 32, 116-127, 2014

Research Article / Araştırma Makalesi

TURKISH COMMONSENSE DATABASE (CSDB) AND CSOYUN (A GAME WITH A PURPOSE)

Serkan ÖZCAN[1], M. Fatih AMASYALI[2*]

[1]*BTC-AG Bilişim Hizmetleri, İSTANBUL*
[2]*Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Müh. Bölümü, Davutpaşa-İSTANBUL*

Received/Geliş: 17.07.2013 Revised/Düzeltme: 19.11.2013 Accepted/Kabul: 20.01.2014

ABSTRACT

It is obvious that the computer systems having commonsense knowledge is more usefull. When the user types "my cat is sick" into the virtual asistant with commonsense knowledge, the system serves the list of veterinarians addresses in the user's area. For English and several languages there are several attempts to construct such systems. In this study, the first Turkish commonsense database (CSdb) and CSoyun (a game with a purpose) is presented. CSoyun is designed to draw upon the internet community in order to add new assertions and rate the existing assertions in the CSdb and thereby improve the quality of the database.
Keywords: Commonsense Databases, semantic web, natural language processing, automated reasoning, non-monotonic reasoning, games with a purpose, gamification, knowledge acquisition, knowledge extraction, feedback mechanisms, artificial intelligence.

TÜRKÇE HAYAT BİLGİSİ VERİTABANI VE CSOYUN

ÖZET

Kullanıcısının yaşadığı dünyayı, amaçlarını, planlarını bilen uygulamaların çok daha kullanışlı olacağı açıktır. Bu tür hayat bilgilerine sahip bir sanal asistana, kullanıcısı "kedim hasta" dediğinde ona en yakın veterinerin iletişim bilgilerini verebilecektir. İngilizce ve diğer diller için bu tür hayat bilgilerinin yer aldığı very tabanlarının oluşturulması için çeşitli çalımalar yapılmıştır. Bu çalışmada ilk Türkçe Hayat Bilgisi veri tabanı olan CSdb ve bu sistemin geliştirilmesi için tasarlanmış CSoyun uygulaması tanıtılmıştır. Kullanıcılarına oyun oynatarak veritabanındaki verilerin doğruluğunu belirleyen ve bilgi miktarını arttıran sistemin 3 senelik kullanım tarihi de özetlenmiştir. CSoyun sayesinde, CSdb kendi kendine büyüyen ve iyileşen bir yapıya kavuşturulmuştur.
Anahtar Sözcükler: Hayat Bilgisi veritabanları, anlamsal ağlar, doğal dil işleme, otomatik çıkarım, monoton olmayan çıkarım, amaçlı oyunlar, oyunlaştırma, bilgi çıkarımı, bilgi edinimi, geri besleme mekanizmaları, yapay zeka.

1. GİRİŞ

Dünya satranç şampiyonlarını yenen bilgisayar programlarımız olmasına rağmen, bir fotoğrafta neler olduğunu bir insan kadar doğru söyleyebilen, ya da verilen bir hikaye hakkındaki sorulara cevap verebilen yazılımlar geliştiremiyoruz. Bilgisayarlara bir şeyi yaptırmak için her şeyi

[*]Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: mfatih@ce.yildiz.edu.tr, tel: (212) 383 57 30

söylemek gerekir. Söylediklerimizden söylemediklerimizi anlayamazlar / tahmin edemezler. Aptal makine teriminin sebebi budur.

Chris Riesbeck, yapay zeka çalışmalarını "bilgisayarların neden bu kadar aptal olduklarını anlama arayışı" olarak tanımlamıştır [1]. Bilgisayarların bu eksiklerinin altında bizlerin yaşadığı dünya hakkında bir şey bilmemeleri diğer bir ifadeyle hayat bilgisinden yoksun olmaları yatmaktadır. Günlük hayatımız hakkında çıkarım yapabilen, bize yardımcı olabilen sanal asistanlar / bilgisayarlar yapay zeka çalışmalarının uzun zamandan beri odak noktalarındandır.

"Kullanıcına, gelen telefon çağrısını bildir" amacına sahip bir telefon, kullanıcısının o an bulunduğu ortamın ses seviyesini ölçerek, saate bakarak, kullanıcısının takvimine bakarak en uygun yolu (yüksek sesle çal, düşük sesle çal, titreşim yap, vb.) seçebilir. Bu seçimi yapabilmesi için "insanların yüksek sesli ortamlarda düşük sesleri duymasının zor olduğu" gibi bilgilere de sahip olması gerekir. Bu sayede günlük yaşamımızda telefonumuzun ses seviyeleriyle oynamamıza gerek kalmaz. Bu tür bilgilerle donanmış sistemlerin bizlere daha faydalı olabileceği açıktır. Verilen bilgilerden çıkarım yapma konusunda çok sınırlı olan günümüz sistemleri hayat bilgisi adını verilen bu tür bilgilere sahip olduklarında "Ali lokantada yemek yedi" cümlesinden "Ali yemek için ödeme yaptı" bilgisini çıkarabileceklerdir. Böyle bir kabiliyete sahip sistemlerin potansiyelleri, kullanım alanları neredeyse sonsuzdur. Kullanıcılarının yaşadıkları dünya, amaçları, kısıtları, bir şeyleri yapmanın çeşitli yolları hakkında bilgi sahibi olan sistemler günümüz mobil cihazlarında yer alan algılayıcı (kamera, mikrofon, ivme sensörü, GPS vb.) ve kullanıcı bilgileriyle (posta kutusu, takvim, sosyal medya iletileri vb.) etkileşime geçerek bağlam tabanlı hesaplama (context-aware computing) alanını oluşturmaya başlamıştır.

Bu tür sistemlerin geliştirilebilmesinin önünde çözülmesi gereken 2 büyük problem vardır. İlki bu bilgilerin bilgisayarlara verilmesi, ikincisi ise bu bilgilerle çıkarım yapabilme kapasitesinin bilgisayarlara verilebilmesidir.

İlk problem, bilgisayarlara hayat bilgisini vermek. Bu bilgilere İnsanların ortak sahip olduğu bilgi, kurallar (hayat bilgisi) sayısı çok fazla. İnsanlar bu bilgiyi yaşamları boyunca öğrenirler. Bu bilgiler öyle bilinen şeylerdir ki, diyaloglarda, yazılı metinlerde çok sıklıkla yer almazlar. Örneğin:

Herkes anne ve babasından gençtir. İnsanların boyları ortalama 1.70 m'dir. Süt beyaz renktedir. İnsanlar kandırılmaktan hoşlanmazlar. Eğer bir bıçağı keskin tarafından tutarsan elini kesebilir. Bir kağıt ateşe düşerse, alev alır. İnsanlar günde 3 kez yemek yerler. İnsanlar yemek yedikten birkaç saat sonra acıkmaya başlarlar. İnsanlar yeni insanlarla tanışmak için partilere giderler. İnsanlar genellikle geceleri uyurlar. Bunlar ve benzeri milyonlarca bilgi ve kuralı biliyoruz.

İkinci problem, bilgisayarlara bu bilgilerle çıkarım yapma kabiliyetini vermektir. Bu bilgileri hergün defalarca kullanarak karşılaştığımız büyük ya da küçük problemleri çözüyoruz. Aşağıda bu tür bilgi işleyen kurallara örnekler verilmiştir:

- Bir problemin varsa, ona benzer bir problemi çözdüğün geçmiş bir tecrübeni hatırla.
- Eğer bir şey yaparsan, ardından ne olacağını tahmin et
- Eğer bir işte başarısız olursan, neyi farklı yapabileceğini / neyin buna sebep olduğunu düşün
- Bir olay gördüğünde, buna neyin sebep olmuş olabileceğini düşün
- Eğer birisi bir şey yaparsa, o kişinin bu şeyi yapmaktaki amacının ne olabileceğini kendine sor.

Yukarıdaki bilgiler ise dış dünyaya ait olmayan, aksine iç dünyamızda diğer bilgilerileri nasıl işleyeceğimiz / onları nasıl kullanacağımız hakkındaki bilgilerdir. Sayıca, ilk problemdeki bilgilere göre çok daha az olan bu tür bilgiler, üzerlerinde çalışacakları bilgiler yoksa kullanılamazlar. Bu sebeple bu çalışmada ilk problemdeki bilgi türleri üzerine yoğunlaşılmıştır.

Bu tür bilgilerin bilgisayarlara aktarımı için çeşitli dillerde çalışmalar yapılmıştır. Çalışmaların en yoğun olduğu dil İngilizce'dir. Türkçe için ise Amasyali, İnak ve Ersen'nin yapmış oldukları CSdb veri tabanı mevcuttur [2].

Bölüm 2'de hayat bilgisi veri tabanları oluşturmadaki çeşitli çalışmalar anlatılmıştır. 3. bölümde oyunlaştırmanın bu tür projelerde kullanım örnekleri verilmiştir. CSdb'nin oluşturulmasının detayları Bölüm 4'de sunumuştur. CSdb'nin veri tabanındaki bilgilerin miktarını ve güvenilirliğini arttırmak için tasarlanan CSoyun ve ayrıntıları Bölüm 5'te yer almaktadır. Oynanma istatistikleri Bölüm 6'da verilmiş, Bölüm 7'de ise çalışmamızdan elde ettiğimiz sonuçlar tartışılmış ve gelecek planlarımız verilmiştir.

2. HAYAT BİLGİSİ VERİTABANLARI NASIL OLUŞTURULUR?

En basit bir kelime sorgusuna bile binlerce sonuç dönen Google arama motoruna "insanlar geceleri uyurlar" sorgusu gönderildiğinde sadece 29 sonuç gelmektedir. Hayat bilgisi dediğimiz bu kadar bariz bilgiler, yazılı olarak nadiren yazılmaktadır, karşılıklı konuşmalarda nadiren geçmektedir. Birbirimizi anlamakta, çıkarım yapmakta çok sıkça kullanılmalarına rağmen bu bilgileri bilgisayarlara vermek için bir araya getirmek bu sebeple zordur. Bir araya getirilmesi düşünülen bu tür bilgilerin miktarı da göz önüne alındığında hayat bilgisi veri tabanlarının oluşturulmasının zorluğu daha da ortaya çıkmaktadır. Bu zorluklara rağmen literatürde yapılan birçok çalışma mevcuttur. Çalışmaları veri toplama yöntemlerine göre sınıflandırırsak temel olarak 4 yaklaşım öne çıkmaktadır:

1. Uzman Kişilerin Elle Bilgi Girişi: Verilerin benzer formatta olmaları ve doğruluklarının yüksek olması için bu konuda eğitilmiş bir grup insanın yıllar boyunca veri girişi yapmasıdır. Bu yöntemin başlıca savunucusu Doug Lenat ve takımı 20 yıl boyunca bilgi girişi yaparak 1.5 milyon bilgi boyutuna erişmişlerdir. Cyc [3, 4] adını verdikleri veri tabanı verilen büyük emek sayesinde yüksek doğruluğa erişmiştir.

2. Otomatik Bilgi Çıkarımı: Google'da "ve diğer benzeri" sorgusu yapıldığında gelen sonuçlarda sorgunun sağ ve solunda yer alan ikililer incelenirse genelde soldaki ifadenin sağdaki ifadenin alt kavramı olduğu görülmektedir. Bu ve benzeri metin şablonları ile kavramlar arası ilişkilerin otomatik olarak toplanabileceği düşüncesi üzerine yapılmış birçok çalışma mevcuttur [5].

3. Gönüllü Kullanıcıların Elle Bilgi Girişi: "Hayat bilgisini herkes bildiğine göre sadece uzman kişilerden bilgi girişi almaya gerek yok, bir site yapalım, gönüllü herkes bilgi girebilsin" yaklaşımıdır. Openmind [6, 7] bu yaklaşımı benimsemiş ve 8000 gönüllü kullanıcıdan kısa süre içerisinde 1 milyona yakın veri elde etmiştir. Bu bilgi sayısına hızla erişilmesi yalaşımın iyi yönünü, bilgilerin düşük kalitesi ise kötü yönüdür. Ayrıca uzun soluklu çalışacak çok sayıda gönüllü kişi bulmakta zordur.

4. Oyunlaştırma: Bilgisayarda ya da cep telefonunda oyun oynamak insanların yapmaktan sıkılmadıkları bir aktivitedir. Bir şeyler kazanma, başarma ve diğerlerine başarılarını gösterme duyguları bu aktivitelerin sürmesindeki temel motivasyonlardır. Gönüllü kullanıcı bulma ve uzun süre bilgi girişi yapmalarını sağlamak için bu bilgi girişi sistemlerini düz web sayfası formları olarak tasarlamak yerine bir oyun sitesi olarak tasarlanmaktadır. Bu oyunlarda kişiler hem yeni bilgi girişi yapmakta hem de yukarıdaki yaklaşımlarla oluşturulmuş veri tabanlarındaki verilerin doğruluğunu denetlemektedirler. Bu yaklaşıma örnek olarak FACTory [8] ve bu çalışmada anlatılan CSoyun verilebilir. Sonraki bölümde bu yaklaşımın örnekleri detaylandırılacaktır.

3. OYUNLARLA BİLGİ TOPLAMAK ÜZERİNE BENZER ÇALIŞMALAR

2. bölümde hayat bilgisi toplamak için önerilen 3. ve 4. yaklaşımların ortak noktası çok sayıda uzman olmayan kullanıcının internet üzerinden veri tabanına katkıda bulunmasını sağlamaktır.

Bu fikrin temel dayanağı, toplanması planlanan bilgiye herkesin sahip olması ve dolayısıyla herkes tarafından veri tabanına katkıda bulunulabilmesidir. Ancak gönüllü kullanıcıları bu tür veri tabanlarına bilgi girişi yapmaları konusunda ikna etmek oldukça zordur.

Bu zorluğun üstesinden gelmek için bilgi girişini kullanıcıların zevk alacağı bir oyun haline getirmek bir çözümdür. Bu tür oyunların en çok bilineni Google'ın kendi resim arama motorunu iyileştirmek için kullanıma açtığı Resim Etiketleme uygulamasıdır. Bu oyunda kullanıcıya bir resim verilmekte ve belli bir süre içinde resimde neler olduğunu yazması beklenmektedir. Kullanıcıların yazdıkları ifadeler o resmin arama etiketlerine eklenmektedir. Böylece Google resimleri emek harcamadan etiketletmekte, kullanıcılarda puanlarıyla diğer kullanıcılar arasındaki yerlerini görmektedirler.

Verbosity [9] hayat bilgisi toplamak için tasarlanmış 2 kişilik bir oyundur. Oyunculardan biri aklından bir kavram tutmakta, diğeri de çeşitli tahminlerle o kavramı bulmaya çalışmaktadır. Kavramı aklından tutan, diğer oyuncunun her tahmininden önce bir ipucu vermektedir. Sistem ipuçları ve tahminleri kullanarak veri tabanını geliştirmektedir. Örneğin, "gagası var" ipucuna karşılık "martı" tahmin edilirse, "martı" doğru cevap olmasa bile "martının gagası olduğu" bilgisi veri tabanına eklenmektedir. Toplanan verilerden rastgele seçilen bir kısmı üzerinde yapılan testlere göre doğru bilgi giriş oranı %85'tir.

Common Consensus [10] oyunu, "100 kişiye sorduk" yarışma programının veri toplamak için oluşturulmuş bir versiyonudur. "Bir çalışma ofisinde bulunabilecek 5 şeyi 100 kişiye sorduk. Aldığımız cevaplar nelerdir?" şeklindeki sorularla verilen cevaplarla veri tabanı hem büyütülmekte hem de cevapların frekansı kullanılarak var olan bilgilerin güvenilirlik değerleri güncellenmektedir. Toplanan verilerden rastgele seçilen bir kısmı üzerinde yapılan testlere göre girilen bilgilerin %33'ü doğrudur.

Cyc veri tabanının sahibi firma tarafından tasarlanan FACTory [8] oyununda Cyc veri tabanından rastgele seçilen bilgiler kullanıcılara sorulmakta ve cevap olarak doğru, yanlış, bilmiyorum, anlamsız seçeneklerinden birini seçmesi istenmektedir. Bir bilgiye verilen cevapların çoğunluğu sistemin o bilgiye dair güvenilirliğini oluşturmaktadır. Toplanan verilerin doğruluğuna dair bir ölçüme erişilememiştir.

TeachRose (www.teachrose.com), kelimeler arası ilişkileri kişilerden öğrenmek için oluşturulmuş bir sistemdir. Herkesten küçük bir şey öğrenerek, büyük bir veri tabanı oluşturulabileceği varsayımına dayanmaktadır. Kişiler sisteme 4 farklı şekilde bilgi girişi yapabilmektedir. Bunlar; şu kelimeyi görünce aklına gelen 3 kelimeyi söyle, bu kelime ile şu kelime ilişkili midir, benimle konuş, şu kelimeyi görünce aklına gelen bir soru-cevap ikilisini söyle aktiviteleridir. Sistem sadece kelimelerin birbirleriyle ilişkililerinin büyüklüğünü öğrenmekte, kelimeler arası ilişkinin türü hakkında bir fikri bulunmamaktadır. Mevcut veri tabanı, 37 bin kelime, 77 bin ilişki içermektedir. Toplanan verilerin doğruluğuna dair bir ölçüme erişilememiştir.

Oyun olmasa bile kullanıcılardan veri toplamak için geliştirilen bir diğer uygulama ReCAPCHTA'dır. Bir web sistemine giriş yapmak isteyeni gerçek bir insan olup olmadığını anlamak için sistemin girişine içinde biraz bozulmuş bir metin olan resim konması çokça rastlanan bir uygulamadır. Bu resimlerdeki metinler, insanların kolaylıkla okuyabildiği ancak bir OCR (optik karakter tanıma) programının çok zorlanacağı resimlerdir. Bu metinlerin gerçekte ne oldukları bilindiğinden kullanıcı insansa sisteme kolaylıkla giriş yapabilmekte ancak bir bilgisayar programı ise sisteme girememektedir. ReCAPCHTA'da, bu yaklaşım OCR programları tarafından okunması zor metinlerin insanlara okutulması için geliştirilmiştir. Kullanıcının önüne biri içeriği bilinen, diğeri bilinmeyen 2 metin resmi çıkmaktadır. İçeriği bilineni doğru girdiyse, içeriği bilinmeyeni de doğru girdiği varsayılarak metin resimlerinin otomatik okunması gerçekleştirilmektedir.

4. TÜRKÇE HAYAT BİLGİSİ VERİTABANI (CSDB)

Bu bölümde gündelik hayat bilgilerini tutmak için tasarlanan veri tabanının (CSdb) yapısı, veri tabanını doldurmak için kullanılan kaynakların tanıtımı yer almaktadır.

4.1. Sistemin Veri Kaynakları

Sistemin ilk versiyonunda ConceptNet, orijinal Wordnet, Türkçe Wordnet ve 400 bin web sitesi olmak üzere 4 temel veri kaynağı bulunmaktadır. Bu kaynakların ikisi (ConceptNet ve orijinal Wordnet) ingilizce kaynaklar oldukları için otomatik bir çeviri sisteminden geçirildikten sonra kullanılmışlardır. ConceptNet, OpenMind projesinde toplanan cümlelerden otomatik olarak oluşturulmuş yaklaşık 200 bin kavram içeren bir anlamsal ağdır [11]. Kavramlar arası ilişkiler ve bu ilişkilerin işlenmemiş OpenMind veri tabanındaki frekanslarından elde edilmiş güvenilirlik ölçümleri ConceptNet veri tabanında yer almaktadır. Veri tabanına http://web.media.mit.edu/~hugo/conceptnet/ adresinden erişilmektedir. Orijinal Wordnet George A. Miller tarafından oluşturulmaya başlanmış bir veri tabanıdır [12]. Aynı anlama sahip kelime gruplarından oluşan eşkümeler (synset) ve bu eşkümeler arasındaki çeşitli ilişkiler ağından oluşur. Veri tabanına http://wordnetweb.princeton.edu/perl/webwn adresinden erişilebilir. Türkçe Wordnet ise, orijinal Wordnet'in Türkçe'sinin oluşturulması için BalkaNet projesi kapsamında hazırlanan bir veri tabanıdır [13]. Veri tabanına www.hlst.sabanciuniv.edu/TL/ adresinden erişilebilir. Sistemin bir diğer kaynağı bir web örümceği kullanılarak kaydedilmiş 400 bin adet web sitesinin html kodlarından oluşan bir veri tabanıdır.

Sistemin 2010 yılında oluşturulan ilk versiyonunda yukarıdaki 4 veri kaynağı yer almaktadır. 2012 yılında Yazıcı ve Amasyalı tarafından yapılan çalışma [14] ile Türk Dil Kurumu sözlüğü (www.tdk.gov.tr) ve Vikisözlük (http://tr.wiktionary.org/wiki/Ana_Sayfa) kaynaklarından şablon tabanlı otomatik bilgi çıkarımın yöntemiyle toplanan yeni veriler ve ilişkiler veri tabanına dahil edilmiştir.

4.2. Tasarlanan Veri Tabanı Yapısı

Tasarlanan sistemimizde bilgiler temelde 3 çizelgede tutulmuştur. İlk çizelgede bir ya da birkaç kelimeden oluşan kavramlar, ikinci çizelgede kavramlar arası ilişkilerin türleri, üçüncü çizelgede ise ilişkilerin kendileri bulunmaktadır. Kavramları içeren çizelgede ve ilişki türlerini içeren çizelgelerde herbir kavrama ve ilişki türüne tekil bir id verilmiş ve ilişkiler çizelgesinde ilşikiler bu id'ler üzerinden tanımlanmıştır.

4.3. Önişlemler

Tasarlanan veri tabanının doldurulmasında kullanılan kaynaklarda veriler bizim tasarladığımız ortak veri tabanından farklı formatlarda tutulmaktadır. Bu nedenle içerdikleri bilgilerin veri tabanına aktarılmadan önce bir önişlemden geçirilmiştir.

ConceptNet'te bilgiler, kavramları ve ilişkili oldukları kavramları içeren tek bir metin formatındadır. Metin dosyası incelenmiş ve formatı anlaşıldıktan sonra kavramları ve aralarındaki ilişkileri veri tabanımıza kaydeden programlar yazılmıştır.

Wordnet'te bilgiler her bir ilişki türüne ait farklı metin dosyalarında tutulmaktadır. Eğer iki eşküme arasında bir ilişki varsa ilk eşküme içindeki her bir kelimeyle diğer eşküme içindeki her bir kelime arasında o ilişki vardır şeklinde yorumlanmış ve veri tabanımıza bu şekilde kaydedilmiştir. Her bir metin dosyası için aynı metot uygulanmış sadece veri tabanına eklenirken ilişki isimleri değiştirilmiştir.

Türkçe Wordnet'te ise bilgiler xml formatında tutulmaktadır. Ancak xml'ni temel yapısı orijinal Wordnet'le aynıdır (eşkümeler ve eşkümeler arası ilişkiler). Bu nedenle verilere erişmek

ve kendi veri tabanımıza kaydetmek için orijinal Wordnet'te kullanılan yaklaşım izlenmiştir.

Web sayfalarının önişlemlerinde, sayfalar öncelikle HTML kodlarından arındırılmıştır. Daha sonra Zemberek [10] kelime çözümleyicisi kullanılarak tüm kelimeler çözümlenmiş ve frekansı belli bir eşik değerinin üzerinde yer alan kelime ve kelime grupları kavramlar çizelgesine kaydedilmiştir. Bununla birlikte 2 kelime içerem kelime grupları ayrıca isim-isim, sıfat-isim, isim-fiil gibi ilişki türleriyle ilişkiler çizelgesine de kaydedilmiştir.

TDK sözlüğü ve Vikisözlükten bilgi çıkarımı işleminin ayrıntıları için Yazıcı ve Amasyalı'nın çalışması [14] incelenmelidir.

4.4. Veri Tabanına ait İstatistikler

Sistem ilk oluşturulduğunda 4 farklı kaynaktan alınan 475407 adet kavram ve bunlar arasında 40 farklı ilişki türüne ait 1089230 adet ilişki içermektedir. İlişki türleri ve bu ilişkiye sahip kavram sayıları Çizelge 1'de verilmiştir.

Çizelge 1. Veri tabanının içerdiği ilişki türleri ve frekansları

İlişki Türü	Concept Net	Orijinal Wordnet	Türkçe Wordnet	Web
Ne için kullanılır?	36864	0	0	0
Bu ne yapabilir?	51549	0	0	0
Nerede bulunur?	30778	0	0	0
Ne arzu eder?	5989	0	0	0
Bunun için ne gerekir?	17822	0	0	0
Bunun ne özellikleri var?	11214	0	0	0
Neyden yapılmış?	1000	0	0	0
Neyin bir parçası?	8105	0	0	0
İçerdiği olaylar nelerdir?	20330	0	294	0
Bunun tanımı nedir?	2721	0	0	0
Neye sebep olur?	13010	907	237	0
Neyi istetir?	7777	0	0	0
Hangi hedef için bu yapılır?	5297	0	0	0
Bunun için ilk önce ne yaparsın?	3147	0	0	0
Bu ne tarafından oluşturulur?	107	0	0	0
Buna neler yapılır/uygulanır?	145	0	0	0
Bu hangi olayla biter?	2839	0	0	0
Eşanlamlı	0	124320	6999	0
Üst Kavramıdır	34566	282137	24141	0
Benzer Fiiller	0	2807	758	0
Alan adı nedir?	0	0	776	0
Yaklaşık Zıtanlamlı	0	0	1678	0
Durumundadır	0	0	1546	0
Bölümün Bütünü	0	27842	2385	0

Üyenin Bütünü	0	57717	2907	0
Benzer Anlam	0	21999	504	0
Parçanın Bütünü	0	0	230	0
Zıtanlamlı	0	3463	0	0
Sıfatın Eylemi	0	115	0	0
Birlikte geçmek	0	433	0	0
Bu neyi gerektirir?	0	1990	0	0
Bunun içeriği nedir?	0	2349	0	0
Sıfatın İsmi	0	1885	0	0
İsim Hali	0	6087	0	0
Fiil - Fiil	0	0	0	10255
İsim - Fiil	0	0	0	200542
İsim Tamlaması	0	0	0	3370
Sıfat - Fiil	0	0	0	16312
Sıfat - Sıfat	0	0	0	3735
Sıfat - Tamlaması	0	0	0	25250
Toplam ilişki sayısı	253260	534051	42455	259464
Genel Toplam = 1089230				

Çizelge 1 incelendiğinde, farklı kaynaklarda yer alan aynı ilişki türlerinin olmasına rağmen temelde ilişki türlerinin birbirlerinden ayrık olduğu ve tasarladığımız veri tabanının bu açıdan bütünleştirici bir içeriğe sahip olduğu söylenebilir. Veri tabanına 2012 yılında, TDK sözlük ve Vikisözlükten bulunan veriler de eklendiğinde ilişki sayısı 1.21 milyona çıkmıştır.

4.5. Sistemin İçerdiği Bilgilere Örnekler

Sistemin içerdiği çeşitli ilişki türlerinden 6'sına ait çeşitli bilgi ikilileri sistemin içeriği hakkında bilgi vermesi amacıyla Çizelge 2'de verilmiştir.

Sistemin içerdiği kavram sayısı, ilişki sayısı, ilişki türü çeşitliği yüksek olmasına rağmen, bilgilerin güvenilirliği çok düşüktür. Bunun ana sebebi kaynaklardan alınan bilgiler üzerinde otomatik bir çeviri işleminin uygulanmış olmasıdır. Otomatik çeviri sistemi olarak hali hazırdaki en iyisi (Google Translate) kullanılmış olmasına rağmen yine de oldukça problemlidir. Bu sebeple bilgilerin çeşitli uygulamalarda kullanılabilir bir hale getirilmesi için güvenilirliğinin arttırılması gerekmektedir. Bunun için CSoyun uygulaması geliştirilmiştir. 5. Bölümde bu uygulamanın detayları verilmiştir.

Çizelge 2. Veri tabanının içerdiği bilgilere 6 ilişki türünden örnekler

Bunun için ne gerekir?	*Neye sebep olur?*	*Bundan neler yapılır?*
yazmak-araştırmak	Öldürmek-ceza	taş-köprü
denemek-para	Doğurmak-hayat	çelik-makine
uyumak-yatmak	Sevmek-umut	su-bulut
seyahat etmek-enerji	Sevmek-acı	kağıt-gazete
öğrenmek-okumak	ateş-acı	yün-kumaş
yaşam-yiyecek	Öldürmek-üzüntü	kumaş-gömlek
Ne için kullanılır?	*Bu ne yapabilir?*	*Nerede bulunur?*
asker-savaş	kuş-uçmak	oda-bina
çatal-yemek	Kişi-yürümek	kişi-oda
top-oynamak	Bilgisayar-düşünmek	elbise-mağaza
ördek-yemek	çocuk-düşmek	kemik-kişi
hastalık-öldürmek	bıçak-kesmek	asker-savaş
baş-düşünmek	gemi-batmak	öğrenci-okul

5. TASARLANAN OYUN SİTESİ (CSOYUN)

Oluşturulmuş veri tabanındaki bilgilerin doğruluğunun belirlenmesi ve eksikliklerinin kapatılması için gönüllü kullanıcıların katılımının en kolay sağlanabileceği uygulama oyunlardır. İnsanlar oyun siteleri üzerinde oldukça fazla zaman/emek harcamaktadır. Bu zaman bir amaca yöneltilebilirse, gerçekten çok büyük miktarda veriler toplanabilecektir.

Boş bir veri kümesinin doldurulması konusunda kullanıcıların çekinceler yaşayacakları, ancak içeride birşeyler olduğunu gördükçe, katılım yapmaya daha istekli olacakları düşünüldüğünden 4. Bölümde oluşturma adımları anlatılan CSdb CSoyun'a dahil edilmiştir.

Amaç, bilgilerin doğruluğunu öğrenmek olunca, kullanıcıların girdikleri bilgilerin doğruluğunun ölçümü zorlaşmaktadır. Yeni bilgi girişinde de aynı sorun bulunmaktadır. Bu sorunun çözümü için ilişki verilerine güvenilirlik değeri eklenmiş ve kaç kişi tarafından oylandığı, ortalama doğruluk değeri tutulmuştur. Bu sayede veri tabanından güvenilirliği yüksek bilgilerin çekilmesi mümkün olmaktadır. Kullanıcı sayısının mümkün olduğunca fazla olması sistemin bu güvenlik yapısının daha iyi çalışmasını sağlamaktadır.

Kullanıcılar isterlerse sisteme kayıt olarak puanlarının kaydedilmesini sağlayabilmektedir. Oyunlardaki tüm aktiviteler önceden belirlenmiş puanlarla ödüllendirilmektedir. Sistemde en yüksek puanlı kullanıcılar listelenerek oyuncular arası rekabet oluşturulmaktadır. Belli puan seviyelerinde kullanıcının rütbesi yükseltilerek oyuna devam etmesi sağlanmaya çalışılmıştır.

Tasarlanan web sitesinde kullanıcılar 5 farklı oyun oynayabilmektedir. Sisteme www.kemikoyun.yildiz.edu.tr/commonsense adresinden erişilebilir.

BİLDİKLERİM OYUNU: Kullanıcı öncelikle bir ilişki türü seçer. Sistem CSdb'den o ilişkiye sahip kavram ikililerini rastgele kullanıcıya gösterir. Kullanıcı bu bilgilere oy verir. Oylar 0 (en düşük) ile 5 (en yüksek) arasındadır. Verilen her oyla o bilginin güvenilirlik değeri güncellenmiş olur.

ARAŞTIR OYUNU: Kullanıcı öncelikle bir kavram girer ve bir ilişki türü seçer. Sistem CSdb'den o kavramla o ilişkiye sahip kavramları rastgele kullanıcıya gösterir. Kullanıcı bu bilgilere oy verir.

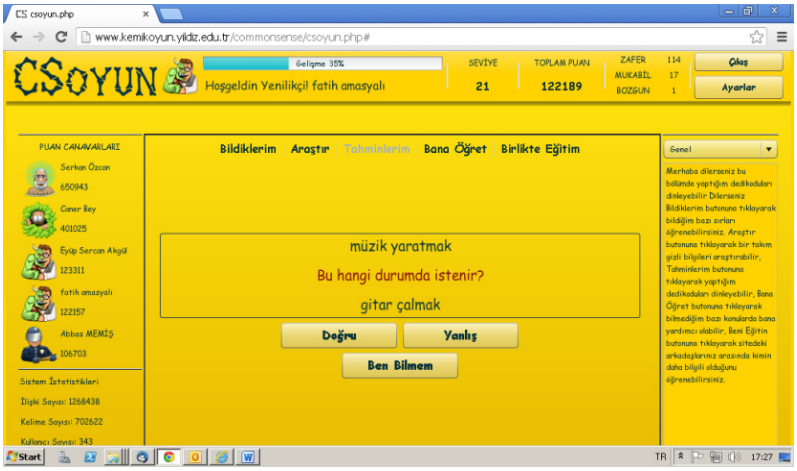
TAHMİNLERİM OYUNU: Bu oyun Csoyun'u benzer sistemlerden/oyunlardan ayıran en önemli özelliğidir. Sadece varolan bilgilerin oylatılması yerine, varolan bilgilerden monoton olmayan (non-monotonic) çıkarım yapılarak doğru olabilecek hipotezler üretilir ve hipotezler çizelgesine kaydedilir.

Monoton çıkarımlar öncülleri doğru ise sonuç çıkarımınında doğru olacağı kesin mekanizmalardır. Tümdengelim çıkarımlar bu türden mekanizmalardır. Örneğin "Ahmet öğrencidir", "Öğrenciler okula gider" öncülleri doğru ise "Ahmet okula gider" çıkarımı kesinlikle doğrudur. Ancak bu tür bir çıkarımın CSdb'ye dahil edilmesinin bir anlamı yoktur. Zaten içindeki bilgilerden istenildiği an bu çıkarımlar yapılabilir. Bu tür çıkarımların yapılıp veri tabanına eklenmesi veri tabanının anlamsız yere büyütülmesidir. Bunun yerine doğru olması muhtemel çıkarımlar yapıp, bunlar kullanıcılar tarafından onaylanırsa veri tabanına dahil etmek çok daha geliştirici bir yaklaşımdır. Bu sayede sistem kendi kendine doğru olması muhtemel yeni hipotezler üretmekte ve kullanıcı etkileşimi ile yeni bilgiler öğrenebilmektedir. Sistemde hipotezlerin 4 farklı üretilme çeşidi vardır. Çizelge 3'de bu türler ve örnekleri verilmiştir.

Çizelge 3. Sistemdeki Çıkarım Türleri

Çıkarım türü	Örnek çıkarımlar
(X-R1-Y) and (Y-R2-Z) → (X-R3-Z)	(martı-altkavramıdır-kuş) and (martı-bu ne yapabilir-uçmak) → (kuş-bu ne yapabilir-uçmak)
(X-R1-Y) → (X-R2-Y)	X içerdiği olaylar nelerdir Y → X bunun için ilk önce ne yaparsın Y
(X-R1-T) and (Z-R1-T) and (X-R2-Y) → (Z-R2-Y)	(Kedi-bir_tür-memeli) and (köpek-bir_tür-memeli) and (kedi-sahiptir-kürk) → (köpek-sahiptir-kürk)
(X-R1-Y) and (X-R2-Z) → (Z-R3-Y)	X nerede bulunur Y, X ne için kullanılır Z → Z nerede bulunur Y

CSdb'de tüm bilgiler (kavram-ilişki türü-kavram) üçlüleri şeklinde tutulmaktadır. Çizelge 4'teki R'ler ilişki türlerini, X, Y, Z ve T'ler kavramları göstermektedir. Kullanıcıya bu oyunda hipotezlerin doğru ya da yanlış olduğu sorulur. Doğru cevabı alınan hipotezler, ilişkiler çizelgesine eklenir. Şekil 1'de Tahminlerin oyununun arayüzü gösterilmiştir.



Şekil 1. Tahminlerin Oyununun Arayüzü

BANA ÖĞRET OYUNU: Kullanıcıya bir kavram ve bir ilişki türü verilir. Bu kavramla bu ilişki türüne sahip kavramlar istenir. Kullanıcının verdiği bilgi eğer CSdb'de bulunuyorsa kullanıcı fazladan puan alır.

BİRLİKTE EĞİTİM OYUNU: Bana öğret oyununun iki kişinin karşılıklı oynadığı versiyonudur. Sistemdeki kullanıcılar ikişerli eşleştirilir. Eşleşen kullanıcılara aynı soru (bu kavramla bu ilişki türüne sahip kavramlar nelerdir?) sorulur. Verilen aynı cevaplar için kullanıcılar fazladan puan alır.

Yukarıda ayrıntıları anlatılan 5 oyunla sistem müdahele gerektirmeyen, kendi kendine gelişebilir bir yapıya kavuşmuştur. Bilgilerin güvenilirlikleri kullanıcı oylarıyla belirlenmekte, yeni ilişkiler tahminlerim, bana öğret ve birlikte eğitim oyunlarıyla üretilmekte, yeni kavramlar bana öğret ve birlikte eğitim oyunlarıyla CSdb'ye girilmektedir. Bu sayede zaman içerisinde (oyunlar oynandıkça) sistemin içerdiği bilgi miktarı ve bilgilerinin güvenilirliği artacaktır.

Sistemde ayrıca oyunların parametrelerinin (kullanılacak ilişki türleri, oyun süreleri, oyun puanları vs.) ayarlanması için ve güvenirliği istenen değerlerin üzerindeki ilişkileri dışa aktarmak için yönetim ekranları da tasarlanmıştır.

6. OYUN OYNAMA İSTATİSTİKLERİ

CSoyun sitesi Ocak 2010 tarihinde erişime açılmıştır. Sistemin ilk haline dair bilgiler Çizelge 1'de verilmiştir. Çizelge 2'de ise sistemin aradan geçen süre boyunca gelişimi gösterilmiştir.

Çizelge 4. Sistemin Oynanma İstatistikleri

	1.Rapor	**2.Rapor**
Rapor Zamanı	Şubat 2012	Temmuz 2013
Oy Sayısı	155 bin	407 bin
Kullanıcı Sayısı	182	343
İlişki sayısı	1.2 milyon	1.27 milyon
Oylanan ilişki sayısı	99 bin	225 bin
Oyların Ortalaması	3.1	3.05
Kavram sayısı	700 bin	702 bin

Çizelge 4'e göre oyunun kullanıma açıldığı 3.5 yıl boyunca veri tabanındaki ilişkilerin yaklaşık %18'i oylanmıştır. Sistemdeki tüm ilişkilerin 57 bin'i CSoyun aktiviteleri ile veri tabanına girmiştir. Geri kalan 1.21 milyonu büyük veri yığınları halinde veri tabanına dahil edilmiştir.

Kullanıcı başına ortalama oy sayısı yaklaşık 1200'dür. Sisteme kayıt yaptıran kişi sayısı oldukça az olmasına rağmen, giriş yapanların verdikleri katkılar yüksektir.

Çizelge 4 incelendiğinde oyların ortalamasının zaman içinde pek değişmediği (3.05 ile 3.1 arasında) görülmektedir. Buradan veri tabanındaki ilişkilerin ortalama yarısının (0 ile 5'in ortası 3'tür) doğru olduğu söylenebilir.

Sistemdeki bilgilerden güvenilirliği yüksek olanların sayısı Çizelge 5'te verilmiştir. En az 10 ve 20 kişi tarafından oylanmış, ortalama güvenilirlik puanı 3, 4 ve 5'ten büyük olan ilişki sayıları listelenmiştir.

Çizelge 5. Güvenilirliklere göre ilişki sayıları

Minimum oy sayısı	Minimum Oy ortalaması	Bu kısıtlara uyan ilişki sayısı
10	3	4.1 bin
10	4	3.7 bin
10	5	1.9 bin
20	3	514
20	4	460
20	5	93

Çizelge 5 incelendiğinde doğruluğundan emin olabileceğimiz (eğer en az 10 kişiden alınan bilgiye güvenirsek) ilişki sayısının toplam ilişki sayısına göre oldukça düşük olduğu görülmektedir.

Verilerine erişilebilen oyun tabanlı diğer sistemlerle CSoyun'un bilgi miktarı, çeşitliliği ve doğruluk oranları karşılaştırıldığında bilgi miktarı ve çeşitliliği açısından CSoyun'un iyi bir seviyede olduğu söylenebilir. CSoyun'un yaklaşık %50 olan doğruluk oranı ise benzerlerine göre orta bir seviyede yer almaktadır. CSoyun veri tabanının (CSdb) ilk oluşturulurken kullanılan otomatik çeviri mevcut sistemin doğruluk oranının düşük olmasının başlıca sebebidir. Bu çeviri yapılmamış olsaydı sistemdeki veriler çok daha doğru ancak miktarı çok daha az olacaktı. Bu durum ise kullanıcılara sürekli aynı soruların, aynı kelimeler hakkındaki soruların sorulmasına sebep olacaktı ve oyunun uzun süre oynanabilirliğini engelleyecekti. Bu sebeple başlangıç olarak küçük ve doğru bir veri tabanı yerine, büyük ve yanlışlar içeren bir veri tabanı tercih edilmiştir.

7. SONUÇ VE GELECEK ÇALIŞMALAR

Gündelik hayat bilgisi veri tabanlarının geleceğin bilgisayar sistemlerinin vazgeçilmez parçaları olacağı yönünde birçok görüş bulunmaktadır. Bu nedenle literatürde birçok çalışma yapılmıştır. Bu çalışma da ise Türkçe için ilk gündelik hayat bilgisi veritabanı oluşturulmuş ve yeni bilgilerin girişi ve var olanların güvenilirliklerinin arttırılması için web üzerinden oynanan bir oyun (CSoyun) tasarlanmıştır.

Tasarlanan sistem müdahele gerektirmeyen, kendi kendine gelişebilir bir yapıdadır. Bilgilerin güvenilirlikleri kullanıcı oylarıyla belirlenmekte, yeni ilişkiler ve yeni kavramlar yine oyunlarla veritabanına girilmektedir. Bu sayede zaman içerisinde (oyunlar oynandıkça) sistemin içerdiği bilgi miktarı ve bilgilerinin güvenilirliği artacaktır.

Oyun istatistikleri incelendiğinde sistemin çalıştığı ancak tüm ilişkilerin oylanmasının çok uzun süre alacağı görülmektedir. Sisteme katkıda bulunan gönüllü sayısının azlığı projenin amacına ulaşmasındaki en büyük engeldir.

Genel olarak sistemdeki verilerin yarısının doğru olabileceği (oyların ortalaması 3), ancak bu ilişkilerin hangilerinin doğru olduğunun belirlenmesinin mevcut CSoyun sistemi ile yapılmasının çok uzun zaman alacağı görülmektedir.

Bu engeli aşmak için oyunun Facebook ve Android ortamlarındaki versiyonlarının yapılması planlanmaktadır. Ayrıca veri tabanına bilgi girişi oyunu yerine başka ve popüler bir oyun oynarken ipuçları istendiğinde CSdb'ye bilgi girişi talep eden farklı oyunların tasarlanması da gündemimizdedir. Bununla birlikte, sistemdeki bilgileri kullanan akıllı ajanda, akıllı web tarayıcısı, otomatik soru cevaplama uygulamalarının geliştirilmesi de düşünülmektedir.

Acknowledgments / Teşekkür

CSdb'nin ilk versiyonu oluşturan Bahar İnak and Zeki Ersen'e, değerleri yorumları için Kemik Doğal Dil İşleme grubu üyelerine (www.kemik.yildiz.edu.tr), veri tabanına değerli katkıları için Yıldız Teknik Üni. Bilgisayar Mühendisliği Bölümü, Yapay Zeka dersi ve Uzman Sistemler dersi

öğrencilerine çok teşekkür ederiz.

REFERENCES / KAYNAKLAR

[1] Riesbeck, C., Northwestern University, Erişim Adresi: http://www.cs.northwestern.edu/~riesbeck/ [Erişim tarihi; 20.01.2014].

[2] Amasyalı, M.F., İnak, B., ve Ersen, M.Z., "Construction of Turkish Commonsense Database", Akademik Bilişim, AB 2010, Muğla, Türkiye

[3] Lenat, D.B., Ramanathan, V.G., Karen, P., Dexter, P., ve Shepherd, M., "CYC: Toward programs with common sense", The Communications of the ACM, 33(8), 31–49, 1990.

[4] Lenat, D. B., "Cyc: A Large-Scale Investment in Knowledge Infrastructure", The Communications of the ACM, 38(11), 33-38, 1995.

[5] Chang, C-H., Kayed, M., Girgis, M.R., Shaalan, K.F., "A Survey of Web Information Extraction Systems", IEEE Trans. on Knowl. and Data Eng., 18(10), 1411-1428, 2006.

[6] Speer R., Havasi C. ve Lieberman H., "AnalogySpace: Reducing the Dimensionality of Commonsense Knowledge", Conference of the Association for the Advancement of Artificial Intelligence (AAAI-08), Chicago, 2008.

[7] Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., ve Zhu, W.L., "Open Mind Common Sense: Knowledge acquisition from the general public", Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, Irvine, CA, 2002.

[8] Cycorp Şirketi Resmi Sitesi, Erişim Adresi: http://game.cyc.com/ [Erişim tarihi; 20.01.2014].
Ahn, M.K., Blum, M., "Verbosity: A game for collecting common-sense knowledge", Proceedings of ACM Conference on Human Factors in Computing Systems, CHI, 2004.

[9] Lieberman, H., Smith, D., ve Teeters, A., "Common consensus: A web-based game for collecting commonsense goals", Workshop on Common Sense for Intelligent Interfaces ACM, 2007.

[10] Liu, H. ve Singh, P., "ConceptNet: A Practical Commonsense Reasoning Toolkit", BT Technology Journal, Volume 22. Kluwer Academic Publishers, 2004.

[11] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. ve Miller, K., "Introduction to WordNet: An On-line Lexical Database", 1993.

[12] Bilgin, O., Çetinoğlu, Ö. ve Oflazer, K., "Building a WordNet for Turkish", Romanian Journal of Information Science and Technology, 7(1-2), 163-172, 2004.

[14] Yazıcı, E., Amasyalı, M.F., "Automatic Extraction of Semantic Relationships Using Turkish Dictionary Definitions", EMO Bilimsel Dergi, 1(1), 1-13, 2011.

Mechanical Engineering Article
/
Makine Mühendisliği Makalesi