Türkçe Hayat Bilgisi Veri Tabanının Oluşturulması

M.Fatih Amasyalı[1], Bahar İnak[1], M.Zeki Ersen[1]

[1] Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul

mfatih@ce.yildiz.edu.tr, baharinak@gmail.com, mzekiersen@hotmail.com

Özet: Kullanıcılarının gündelik hayatları hakkında bilgilere sahip bilgisayarlar geliştirmek fikri üzerinde uzun süredir çalışmalar yapılmaktadır. Böyle bilgilere sahip bir sistem, örneğin kullanıcısı kedisinin hasta olduğunu söylediğinde, kullanıcıya en yakındaki veterinerin erişim bilgilerini verebilecektir. Ya da kullanıcısının yarınki evlilik yıldönümü için çiçek siparişi vermesi gerektiğini hatırlatabilecektir. İngilizce başta olmak üzere birçok dilde bu tarz veri tabanları oluşturma çalışmaları sürmektedir. Bu çalışmada ise bir ilk olarak Türkçe gündelik hayat bilgisi veri tabanı tasarlanmış ve içine çeşitli kaynaklarda yer alan bilgiler konarak, büyük miktarda veriyi içermesi sağlanmıştır. Ayrıca içerdiği bilgilere erişim için bir web arayüzü tasarlanmıştır.

Anahtar Sözcükler: Hayat Bilgisi Veri Tabanları, Anlamsal Web, Doğal Dil İşleme.

Construction of Turkish Commonsense Database

Abstract: There are several studies about computer systems have commonsense knowledge. Such a system, when the user says that his/her cat is ill, responses the contact information of the nearest veterinarian. Or it can remind ordering flower for the user's tomorrow wedding anniversary.For English and several languages there are several attemps to construct such systems. In this study, it is the first time that Turkish commonsense database was designed and filled with knowledge from several resources. However, a web user interface was design to access to this database.

Keywords: Commonsense Databases, Semantic Web, Natural Language Processing

1. Giriş

Bilgisayarlar / makineler günümüzde yüzlerimizi, parmak izlerimizi tanıyabilmekte, hava tahminleri yapabilmekteler. Ancak yaşadığımız dünya hakkında, kullanıcılarının gündelik yaşamları hakkında fazla şey bildikleri söylenemez. Örneğin insanların geceleri uyuduklarını, sinemalarda ve tiyatrolarda cep telefonlarının sessize alındığını bilgisayarlar bilselerdi günlük hayatımızı kolaylaştırmada çok daha faydalı olabilirlerdi.

Sağlayacağı faydaların büyük olmasıyla birlikte gündelik hayat bilgilerinin bilgisayarlara aktarılmasında çeşitli problemler bulunmaktadır. İlki gereken bilgi miktarının büyüklüğüdür. Çeşitli çalışmalarda insanların sahip olduğu bu tarz bilgi adedi olarak milyonlarca bilgi parçasından söz edilmektedir [1]. Bu kadar büyük miktarda verinin nasıl toplanacağı, kim tarafından toplanacağı, nasıl bir veri yapısında tutulacağı soruları hala tartışılan sorulardır. Bilginin nasıl toplanacağı ile ilgili 2 temel yaklaşım bulunmaktadır. İlki kısıtlı sayıda bilgi mühendisi tarafından bilgilerin sisteme teker teker özenli bir şekilde girilmesi, ikincisi ise çok sayıda uzman olmayan kişi tarafından rastgele girilmesidir. Her iki yaklaşımında artı ve eksi yönleri

bulunmaktadır. İlk yaklaşımda bilgilerin güvenilirliği artmakta ancak bilgi çeşitliliği ikinci yaklaşıma göre azalmaktadır. Şüphesiz ki bir düzine insanın aklına gelecek şeylerle, söz gelimi 10 bin kişinin aklına gelecek şeylerin varyasyonu çok farklı olacaktır. İkinci yaklaşımda ise girilen bilgilerin çok kontrollü olamayacağı, gereksiz bilgi tekrarlarının, uyuşmazlıkların ortaya çıkacağı şüphesizdir. Literatürde her iki yaklaşım içinde çeşitli çalışmalar yapılmıştır. 2. bölümde bu çalışmalara yer verilmiştir.

Gündelik hayat bilgilerinin bilgisayarlara aktarılmasında karşılaşılan ikinci problem ise, bu kadar çok ve aralarında uyuşmazlıklar bulunan bilgiyle nasıl yeni bilgilerin üretilebileceği, bu bilgilerle ne zaman ve nasıl çıkarım yapılacağıdır. Bu problem içinde literatürde çeşitli çözüm önerileri geliştirilmiştir [2,3].

Günümüzde İngilizce ve birkaç dil için, bu tarz veri tabanları oluşturulmuş ve uygulamaya yönelik çalışmalar ortaya çıkmaya başlamıştır. Ancak Türkçe için bu çalışma ilktir.

Çalışmanın sonraki bölümlerinde sırasıyla mevcut gündelik hayat bilgisi veri tabanlarının tanıtımı, tasarlanan sistemin yapısı, alt parçaları, kullanıcı arayüzü ve gelecekte yapılamsı planlanan çalışmalar anlatılmıştır.

2. Benzeri Çalışmalar

Eksikliklerinin, bilgisayarların aptal olarak nitelendirilmesindeki en büyük etkenlerden biri olması ve olası faydalarının büyük olması gündelik hayat bilgisi veri tabanları oluşturma yönündeki çalışmalara sebep olmuştur. Bu amaçla çeşitli kişi ve gruplarca birçok çalışma yapılmıştır. Bu bölümde bu çalışmalardan en popüler olanları anlatılmıştır.

2.1 Cyc

İçerdiği bilgilerin sınırlı sayıdaki uzman kişi tarafından elle girilmesi görüşünü benimseyen bir çalışmadır [4]. Lenat tarafından 1990 yılında oluşturulmaya başlanmıştır. Günümüzde içerisinde yüzbinlerce kavramin milyonlarca ilişkisinin olduğu söylenmektedir. Geliştirme sürecinde veri tabanının bir kısmı halka açılmış, web kullanıcılarının da veri tabanına katkıda bulunmaları amaçlanmıştır. Sistemin veri tabanına
http://www.cycfoundation.org/concepts
adresinden erişilebilir.

2.2 ThoughtTreasure

Erik T. Mueller tarafından 1994 yılında geliştirilmeye başlanmış olan veri tabanı içerisinde 25 bin kavrama ait 50 bin bilgi parçası içermektedir [5]. Bu projenin veri tabanı da Cyc gibi kısıtlı sayıdaki insan tarafından elle oluşturulmuştur. Bununla birlikte veri tabanında senaryolar olarak adlandırılan insanların gündelik hayatlarında sıklıkla yaptıkları restorana gitmek, sinemaya gitmek gibi olağan durumların içerdiği alt olaylar da yer almaktadır.

2.3 OpenMind

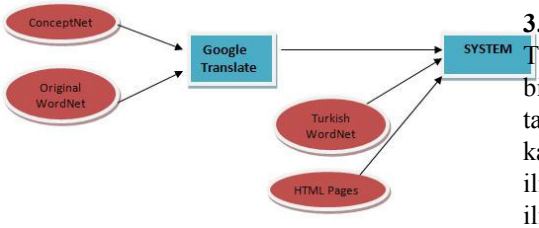
2000 yılındaki başlangıcından itibaren, gereken büyük miktarda bilginin ancak çok sayıda katılımcı ile toplanabileceği fikrinden yola çıkan tasarımcıları ve onlara destek veren binlerce gönüllü web kullanıcısı sayesinde 1 milyon cümle sayısına kısa sürede ulaşmış bir veri tabanıdır [6]. Katılımcıların uzman olmadıkları gerçeğinden yola çıkan tasarımcılar bilgileri her biri farklı türde bilgileri toplayan birçok web arayüzünden cümle formatında almışlardır. Toplanan bilgiler herkesin kullanımına açıktır. Ancak bilgilerin cümle formatında olması işlenmesini, uyumsuzluk ve rastgeleliklerin olması bilgilerin güvenilirliğini azaltmaktadır. Bununla birlikte tasarımcılar bilgilerin güvenilirliğinin tekrar sayılarıyla belirlenebileceğini düşünmüşlerdir.

3. Sistemin Tasarımı

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

3.1 Veri Kaynakları

Sisteme bilgi sağlayan kaynaklar Şekil 1'de görülmektedir.



Şekil 1. Sistemin Kaynakları

Şekil 1'de görüldüğü gibi sistemin 4 temel veri kaynağı bulunmaktadır. Kaynakların 2'si ConceptNet ve orijinal Wordnet ingilizce kaynaklar oldukları için otomatik bir çeviri sisteminden geçirildikten sonra kullanılmışlardır.

3.1.1 ConceptNet

OpenMind projesinde toplanan cümlelerden otomatik olarak oluşturulmuş yaklaşık 200 bin kavram içeren bir anlamsal ağdır [7]. 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.

3.1.2 Wordnet

George A. Miller tarafından oluşturulmaya başlanmış bir veri tabanıdır [8]. 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.

3.1.3 Türkçe Wordnet

Orijinal Wordnet'in Türkçe'sinin oluşturulması için BalkaNet projesi kapsamında hazırlanan bir veri tabanıdır [9]. Veri tabanına www.hlst.sabanciuniv.edu/TL/ adresinden erişilebilir.

3.1.4 HTML sayfaları

Bir web örümceği kullanılarak kaydedilmiş 400 bin adet web sitesinin html kodlarından oluşan bir veri tabanıdır.

3.2 Tasarlanan Veri Tabanı Yapısı

Tasarlanan sistemimizde gündelik hayata ait bilgiler temelde 3 tabloda tutulmuştur. İlk tabloda bir ya da birkaç kelimeden oluşan kavramlar, ikinci tabloda kavramlar arası ilişkilerin türleri, üçüncü tabloda ise ilişkilerin kendileri bulunmaktadır. Ayrıca her tabloda verilerin güvenilirliklerinin hesaplanmasında kullanılan çeşitli parametrelerde yer almaktadır.

Kavramları içeren tabloda ve ilişki türlerini içeren tablolarda herbir kavrama ve ilişki türüne tekil bir id verilmiş ve ilişkiler tablosunda ilşikiler bu id'ler üzerinden tanımlanmıştır.

3.3 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 tablosuna kaydedilmiştir. Bununla birlikte 2 kelime içerem kelime grupları ayrıca isim-isim, sıfat-isim, isim-fiil gibi ilişki türleriyle ilişkiler tablosuna da kaydedilmiştir.

3.4 Veri Tabanına Ait İstatistikler

Sistem 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 Tablo 1'de verilmiştir.

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	20330	0	294	0
nelerdir?				
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 -	0	0	0	25250

Tamlaması				
Toplam ilişki sayısı	253260	534051	42455	259464
Genel Toplam = **1089230**				

Tablo 1. Veri tabanının içerdiği ilişki türleri ve frekansları

Tablo 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.

4. Sistemin Kullanımı

Sistemin içerdiği bilgilere erişim için kullanılan bir arayüzü bulunmaktadır. Kullanıcılar sisteme giriş yaptıktan sonra Şekil 2'de gösterilmiş olan arayüze erişmektedirler.

Şekil 2. Sistemin Arayüzü

Şekil 2'deki arayüzde kullanıcının kavramlarla doldurabileceği iki alan, iki kavram arasındaki ilişki türünü seçebileceği bir çoktan seçmeli liste ve ilişkilerin getirileceği kaynakları seçeceği seçme kutuları yer almaktadır. Bu alanlar kullanarak; şu kavramın hangi kavramlarla hangi ya da şu tür ilişkide olduğu, hangi kavram ikililerinin şu tür ilişkiye sahip olduğu gibi çeşitli sorgular yapılabilmektedir. Bununla birlikte kullanıcının sistemin verdiği cevaplar hakkında puan vermesi de sisteme entegre edilmiştir. Bunun amacı kullanıcılardan gelen geri bildirimlerle bilgilerin güvenilirliklerini arttırmaktır.

4.1 Cevapların Sıralanma Ölçütleri

Kullanıcı sistemde bir sorgulama yaptığında bulunan cevaplar güvenilirlik derecelerine göre sıralanarak kullanıcıya gösterilmektedir. Güvenilirlik değerlerinin hesaplanması sorgu türlerine göre farklılık göstermektedir. Eğer kullanıcı cevabı sadece kavramlardan oluşan bir sorgu (Ör: ağaç ile bütünün üyesi ilişkisine sahip kavramlar nelerdir?) gönderirse cevaplar kavramların frekanslarına göre, cevabı kavram ve ilişki türlerini içeren bir sorgu (Ör: ağaç ile hangi kavramların hangi tür ilişkileri vardır?) gönderirse kavram ve ilişkinin frekansına göre, cevap sadece ilişki türlerini içeren bir sorgu (Ör: ağaç ile kağıt arasında hangi tür bir ilişki vardır?) içinse cevaplar ilişki türünün frekansına göre hesaplanan güvenilirlik katsayılarına göre sıralanarak kullanıcıya gösterilir.

4.2 Sistemin İçerdiği Bilgilere Örnekler

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

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

Tablo 1. Veri tabanının içerdiği bilgilere 6 ilişki türünden örnekler

Tablo 2'de yer alan bilgiler 4.1. bölümde anlatılan sıralama ölçütlerine göre

sıralandıklarında her bir ilişki türü için en yüksek puanlı / en güvenilir bilgilerdir

5. Sonuç

Gündelik hayat bilgisi veri tabanlarının geleceğin bilgisayar sistemlerinin vazgeçilmez parçaları olacağı yönünde birçok görüş bulunmaktadır. Bu nedenle literatürde birçok çalışma yer almaktadır. Bu çalışma da ise Türkçe için ilk gündelik hayat bilgisi veritabanı oluşturulmuş ve erişim için bir web arayüzü tasarlanmıştır.
Sistemin içerdiği yenilikler olarak, Türkçe için bir ilk olması, birçok kaynaktan beslenmesi ve arama seçeneklerinin benzeri sistemlere göre daha gelişmiş olmasıdır.

6. Gelecek Çalışmalar

Gelecekte yapılması planlanan çalışmalar 3 başlıkta toplanmaktadır. İlki sistemin içerdiği bilgi miktarının arttırılması, ikincisi içerdiği bilgilerin kalitesinin arttırılması, üçüncüsü ise bu bilgileri kullanan uygulamaların hayata geçirilmesidir. İlk başlık için cümlelerin öğelerinin kullanımıyla nesne-yer, eylem-yer, özne-eylem gibi ilişki türlerine ait bilgi ikililerinin toplanması düşünülmektedir. İkinci başlık için, Verbosity [11] tarzı oyunlarla kullanıcıların bilgilerin güvenilirliğini arttırmaları sağlanacaktır. Son başlık içinse, akıllı ajanda, akıllı web tarayıcısı, otomatik soru cevaplama uygulamaları düşünülmektedir.

5. Kaynaklar

[1] 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).

[2] 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.

[3] Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N. ve Shepard, B. "Common Sense Reasoning – From Cyc to Intelligent Assistant", In Yang Cai and Julio Abascal (eds.), **Ambient Intelligence in Everyday Life**, LNAI 3864, 1-31, Springer, 2006

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

[5] Mueller, E. T., "Natural language processing with ThoughtTreasure", New York: **Signiform**, 1998.

[6] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins ve Wan Li Zhu, "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.

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

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

[9] 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).

[10] http://code.google.com/p/zemberek/

[11] von Ahn, L., Kedia, M. ve Blum, M., "Verbosity: A Game for Collecting Common-Sense Facts"., **ACM Conference on Human Factors in Computing Systems** (CHI Notes), 2006.