Tümce Öğelerinden Hayat Bilgisi Çıkarımı

Commonsense Knowledge Acquisition by Sentence Analysis

Cihan Özköse, Mehmet Fatih Amasyalı
Yıldız Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü
cihanozkose@gmail.com; mfatih@ce.yildiz.edu.tr

Özetçe

Yaşadığımız dünya ve bizim hakkımızda bilgiye (insanlar neyi sever/sevmez, ister/istemez?, insanlar/nesneler nerelerde bulunur?, hangi eylemler nerede/hangi amaçla yapılır? vb.) sahip bilgisayarlar, hayatlarımızı daha fazla kolaylaştıracaklardır. Örneğin böyle sisteme, kullanıcısı kedisinin hasta olduğunu söylediğinde sistem ona en yakındaki veterinerin telefonunu verecek, bilgisayarlara yapacakları her şeyi en ince ayrıntılarına kadar anlatma gereksinimimiz ortadan kalkacaktır. Ancak böyle uygulamaları mümkün kılacak anlamsal veritabanlarının elle oluşturulması oldukça zor ve zaman alıcı bir süreçtir. Bu tür bilgilerin Türkçe toplanması için başlattığımız projede birçok metin koleksiyonundan ve birçok otomatik bilgi çıkarımı yönteminden yararlanılması düşünülmektedir. Bu çalışmada, bu tür bilgilerin tümcelerin öğe ikililerinden üretilebileceği fikri ve bunun uygulaması sunulmuştur. Çok sayıda tümcenin öğe çözümlemelerinden bu tür bilgiler (hayat bilgisi) otomatik olarak üretilmiştir. Örneğin özne - dolaylı tümleç öğe ikililerinden bir şeyin nerelerde bulunduğu, özne-yüklem öğe ikililerinden bir şeyin yapabildiği şeylerin neler olduğu bilgileri çıkarılmıştır. Yapılan çalışma sayesinde Türkçe için oluşturduğumuz hayat bilgisi veri tabanı için 13 farklı ilişki türüne ait 26.877 adet kavram ikilisi bulunmuştur.

Anahtar sözcükler: Otomatik Bilgi Çıkarımı, Hayat Bilgisi Veri Tabanları, Tümce Çözümleme, Tümcelerin Öğelerıne Ayrılması

Abstract

Computers having commonsense knowledge (What do people like/dislike, want /not want?, Where do you find people/objects?, What are the causes of our action? etc.) facilitate our lives more. Such a system, when the user says that his/her cat is ill, responses the contact information of the nearest veterinarian. However, the manually creation of such semantic databases is very difficult and time consuming process. We initiated a project for collecting Turkish commonsense knowledge from several corpus by several automatic information extraction methods. In this study, the idea of such information can be extracted from the pairs of sentence elements is presented and implemented. For example, "subject-indirect object" pairs says a place where the subject can be found. "subject-verb" pairs says what can a subject do. A commonsense knowledge base were constructed consist of 26.877 knowledge triple of 13 relation types.

Keywords: Commonsene Knowledge Acquisition, Commonsense Databases, Sentence Analysis, Automatic Information Extraction

1. Giriş

Hepimiz biliyoruz: Balıkların denizde yaşadıklarını, insanların geceleri uyuduklarını, gökyüzünün mavi olduğunu, futbolun topla oynandığını. Peki ya bilgisayarlar da bunları bilirlerse?

Bu soruyla başlayan, bilgisayarlar için hayat bilgisi veri tabanı oluşturma çabaları uzun süredir devam etmektedir. Bu çalışmalara örnek olarak İngilizce dili için yapılan Cyc [1] ve Openmind [2], Türkçe

için ise şu an geliştirme aşamasında olan CSdb [6] verilebilir.

Bu tür veri tabanlarının oluşturulması, ulaşmaları düşünülen boyut (milyonlarca bilgi) nedeniyle ya yoğun bir insan emeği sürecini ya da otomatik bilgi çıkarımı yöntemlerinin kullanılmasını gerektirmektedir. Bu çalışmada Türkçe için oluşturmaya başladığımız veri tabanına, Türkçe tümcelerin öğelerini kullanan bir otomatik bilgi çıkarımı metoduyla katkı sağlanması amaçlanmıştır.

Literatürde Türkçe dilinde anlamsal ağ oluşturmak için yapılan ilk çalışmalardan birisi Balkanet projesi dahilinde geliştirilen Türkçe Wordnet'tir [3]. Bu çalışma sonucunda 11.628 eşküme ve bunlar arası 17.550 ilişki içeren bir anlamsal veritabanı oluşturulmuştur. Literatürde Türkçe dilinde geliştirilmiş otomatik bilgi çıkarımı çalışmaları için ilk olarak Tunga Güngör'ün [4] çalışmasından bahsedilebilir. Türkçe bir sözlükte yer alan sözcük tanımlarından, kavramlar arası alt - üst kavram ve eşanlamlılık ilişkilerini otomatik olarak çıkaran şekilsel kurallar tanımlamıştır. Bu konudaki bir diğer çalışma da daha önce yapmış olduğumuz [5] sözcüklerin türleri ve aldıkları eklerden üretilen şablonları kullanan "Türkçe Wordnet'in Otomatik Oluşturulması" adlı çalışmadır.

Türkçe tümceleri otomatik olarak öğelerine ayıran, yaygın kullanılan, bir yazılım olmadığı için; görüldüğü gibi literatürde bilgi çıkarımı için tümcelerin öğelerinin kullanıldığı bir çalışma bulunmamaktadır. Çalışmamız bu yönüyle Türkçe için bir ilktir.

Raporun ikinci bölümünde geliştirme aşamasında olan CSdb'den kısaca bahsedilmiş, sonraki bölümde tümcelerin öğelere ayrılması anlatılmıştır. 4.bölümde bilgilerin genelleştirilmesi için yapılan filtreleme işlemi, 5. bölümde, elde edilen yeni bilgilerin CSdb'ye aktarımı anlatılmıştır. Son bölümde ise sonuçlar ve gelecekte yapılması planlanan çalışmalar yer almaktadır.

2. Mevcut Sistem

Türkçe hayat bilgisi veri tabanı (CSdb) oluşturma çalışmalarımız [6] sonucunda 40 farklı ilişki türüne sahip 475.407 adet kavram veritabanımıza kaydedilmişti. Veri tabanımızdaki bu kavramlara ait 1.089.230 adet ilişki bulunmaktadır. Kavramlar CSdb veri tabanımızda "*kavram1 - ilişki türü - kavram2"* üçlüleri şeklinde tutulmaktadır. Bu çalışma ile yeni ilişki türleri ve bunlara ait kavramların bulunması ve veri tabanına kaydedilmesi; bu sayede daha kapsamlı bir hayat bilgisi veri tabanının oluşturulması amaçlanmıştır.

3. Tümcelerin Öğelerinin Bulunması

Bu bölümde; önce bilgi çıkarımında kullanılacak tümcelerin elde edilmesi, daha sonra tümcelerin öğelerine ayrılması çalışmaları anlatılmıştır.

3.1. Tümce Veri Tabanının Oluşturulması

Otomatik bilgi çıkarımı uygulamalarında elde edilen bilginin güvenilirliğinin yüksek olması için büyük miktarda bilginin kullanılması gerekmektedir. Bu nedenle; öğelerinden bilgi çıkarılacak tümcelerin sayısının büyük olması için "Bilkent Üniversitesi Bilgiye Erişim Grubu" tarafından haber metinlerinden derlenmiş olan Bilcol [7] derlemi kullanılmıştır. Üzerinde işlem yapılacak birimler tümceler olduğundan, derlemin tümceler veri tabanına dönüştürülmesi işlemi 3 adımda yapılmıştır.

1. XML formatındaki derlemde haber metinleri <Text> </Text> etiketleri içinde yer almaktadır. Yazılan basit bir ayrıştırıcı ile derlem her satırda bir tümce içeren bir metin dosyasına dönüştürülmüştür. Bu işlem sonunda 2.187.072 adet tümce içeren bir dosya elde edilmiştir. Bu tümcelerde toplam 30.816.387 adet sözcük bulunmaktadır. Diğer bir deyişle tümcelerimiz ortalama 14,1 sözcük içermektedir.

2. Her bir tümcenin içerdiği tüm sözcüklerin Zemberek [8] çözümleyicisi ile morfolojik çözümlenmesi yapılmıştır. Zemberek tarafından çözümlenemeyen sözcükler tümcelerden silinmiştir. Ayrıca içerdiği sözcüklerden en az biri "FIIL_KOK" içermeyen tümceler silinmiştir. Bu işlem sonunda dosyada içinde en az bir adet fiil olan ve tüm sözcükleri Zemberek tarafından çözümlenebilen 1.902.230 adet tümce kalmıştır. Bu tümcelerdeki ortalama sözcük sayısı ise 13,9'dur.

3. Tek sözcükten oluşan veya on beşten fazla sözcük

içeren tümceler silinmiştir. Bu tümceler başarım oranını düşüreceği ve çözümleme işini zorlaştıracağı için elenmiş ve dosyada 1.218.335 tümce kalmıştır. Üzerinde öğeleme işleminin yapılacağı bu 1.218.335 adet tümcede toplam 10.613.476 sözcük bulunmaktadır. Ortalama sözcük sayısı 8,7' dir. Tablo 1'de tümcelerimizde en yüksek frekansa sahip sözcükler, frekansları ve Zemberek tarafından verilen ilk çözümlemeleri (doğru kabul ettiğimiz) verilmiştir.

Tablo 1: Veri Kümemizde En Fazla Yer Alan İlk 15 Kelime, Frekansları ve Zemberek Çözümlemeleri

Frekans	Kelime	Çözümlemesi
182.759	ve	[ISIM_KOK]
68.193	için	[ISIM_KOK]
35.944	çok	[ISIM_KOK]
31.603	daha	[ISIM_KOK]
27.364	bin	[FIIL_KOK]
26.418	yüzde	[ISIM_KOK]
25.605	en	[ISIM_KOK]
22.072	ilk	[ISIM_KOK]
18.726	büyük	[ISIM_KOK]
17.478	yıl	[FIIL_KOK]
16.083	yeni	[ISIM_KOK]
15.776	son	[SAYI_KOK]
15.696	devam	[ISIM_KOK]
14857	iyi	[ISIM_KOK]
14557	her	[ISIM_KOK]

3.2. Tümcelerin Öğelerinin Bulunması

Bir önceki adımda oluşturulan (1.218.335 adet) tümcelerin öğelerinin bulunması işlemi 12 adımda gerçekleştirilmiştir.

1. Tümcelerdeki her sözcük Zemberek ile çözümlenmiş ve tümcelerin içerikten bağımsız (context-free) çözümlemeleri cozumlemeler.txt adlı bir dosyaya kaydedilmiştir. Bu işlem sonunda 1.218.335 adet tümceden 1.140.422 adet farklı tümce çözümlemesi elde edilmiştir. Tablo 2'de frekansı en yüksek olan 15 tümce çözümlemesi verilmiştir. Tablodaki tümce çözümlemelerinde sözcükler "[ ]" içerisinde verilmiştir.

Tablo 2: Veri Kümemizde Frekansı En Yüksek Olan 15 Tümce Çözümlemesi ve Frekansları

Frekansı	Tümce Çözümlemesi
1486	[ISIM_KOK][FIIL_KOK, FIIL_ZAMAN]
1000	[ISIM_KOK][FIIL_KOK, FIIL_ZAMAN, FIIL_KISI]
874	[ISIM_KOK][ISIM_KOK][FIIL_KOK, FIIL_ZAMAN]
518	[ISIM_KOK][FIIL_KOK]
492	[ISIM_KOK][ISIM_KOK][FIIL_KOK, FIIL_ZAMAN, FIIL_KISI]
462	[ISIM_KOK, ISIM_TAMLAMA_I] [FIIL_KOK, FIIL_ZAMAN]
383	[OZEL_KOK][ISIM_KOK][FIIL_KOK, FIIL_ZAMAN]
373	[ISIM_KOK][ISIM_KOK][ISIM_KOK] [FIIL_KOK, FIIL_ZAMAN]
359	[FIIL_KOK][ISIM_KOK]
337	[ISIM_KOK][FIIL_KOK, FIIL_GEN,FIIL_ZAMAN]
327	[ISIM_KOK][ISIM_KOK, ISIM_TAMLAMA_I][FIIL_KOK, FIIL_ZAMAN]
302	[OZEL_KOK][FIIL_KOK, FIIL_ZAMAN]
249	[OZEL_KOK][FIIL_KOK]
245	[ISIM_KOK][FIIL_KOK, FIIL_SIFAT]
235	[ISIM_KOK, ISIM_TAMLAMA_I] [FIIL_KOK, FIIL_GEN, FIIL_ZAMAN]

Tablo 2 incelendiğinde en yüksek frekansa sahip tümce çözümlemelerinin 2-3 sözcükten oluşan tümceler oldukları görülmektedir. 1.218.335 adet tümce çözümlemesinde, en yüksek frekansın 1486 ve farklı tümce çözümlemesi sayısının 1.140.422 olması Türkçede tümce yapılarının ne kadar değişken olduğunu göstermektedir.

2. cozumlemeler.txt dosyasında yer alan tüm sözcük çözümlemelerinin frekansları bulunmuş ve frekanslar.txt adlı bir dosyaya kaydedilmiştir. Zemberek ile veri kümemizde 10.063 adet farklı sözcük çözümlemesi bulunmuştur. Bunlardan frekansı en yüksek olanlar frekanslarıyla birlikte Tablo 3'te verilmiştir.

Tablo 3: Veri Kümemizde En Fazla Yer Alan İlk 15 Kelime Çözümlemesi ve Frekansları

Frekans	Çözümleme
2.924.409	[ISIM_KOK]
551.421	[OZEL_KOK]
496.564	[ISIM_KOK, ISIM_TAMLAMA_I]
352.132	[FIIL_KOK]
344.408	[FIIL_KOK, FIIL_GECMISZAMAN_DI]
277.460	[SAYI_KOK]
234.010	[ISIM_KOK, ISIM_KALMA_DE]
210.670	[ISIM_KOK, ISIM_YONELME_E]
204.476	[ZAMIR_KOK]
190.810	[FIIL_KOK, FIIL_DONUSUM_EN]
175.209	[ISIM_KOK, ISIM_TAMLAMA_IN]
142.200	[EDAT_KOK]
133.023	[ISIM_KOK, ISIM_BULUNMA_LI]
132.807	[ZAMAN_KOK]
126.009	[ISIM_KOK, ISIM_COGUL_LER, ISIM_BELIRTME_I]

Tablo 3 incelendiğinde veri kümemizde en fazla isim türünden sözcüklerin, daha sonra fiil türünden sözcüklerin yer aldığı görülmektedir.

3. Daha genel öğelere ayırma kuralları elde edebilmek için, sözcük çözümlemelerinde çeşitli genelleştirmeler yapılmıştır. Örneğin tüm fiil zaman ve kişi ekleri birer genel ek haline getirilmiştir. Bu genelleştirme sonucunda 10.063 adet olan farklı sözcük çözümlemesi sayısı 5.249'a indirilmiştir. Tablo 4'te bu işleme ait örnekler verilmiştir.

Tablo 4: Kelime Çözümlemelerinin Genelleştirilmesi Örnekleri

Zemberek Ek Etiketi	Genelleştirilmiş Hali
FIIL_GECMISZAMAN_DI	FIIL_ZAMAN
FIIL_GECMISZAMAN_MIS	
FIIL_GELECEKZAMAN_ECEK	
FIIL_GENISZAMAN_IR	
FIIL_SIMDIKIZAMAN_IYOR	
FIIL_KISI_BEN	FIIL_KISI
FIIL_KISI_BIZ	
FIIL_KISI_ONLAR	
FIIL_KISI_SEN	
FIIL_KISI_SIZ	

4. Çözümlemeler incelenmiş ve bulunabilecek 15 öğe türü (ÖZNE, NESNE, NE_İLE, NEYE, NEYİ, NASIL, NE_ZAMAN, NEREDE, NEREDEN, NE_KADAR, SEBEP, ŞART, SORU, YÜKLEM, NEXT) belirlenmiştir. Türkçede öğeler birden fazla sözcükten de oluşabilmektedir. Bunun için "next" etiketi oluşturulmuştur ve "bu sözcüğe şimdilik bir öğe türü atama, bir sonraki sözcüğe geç" anlamındadır. Örneğin, sıfatlar "next" etiketiyle işaretlenen sözcüklerdir.

5. 3. adımda bulunan 5.249 adet tekil çözümlemelerden frekansı 1000'den yüksek olan çözümlemelerin (340 adet) karşılık geleceği öğe türleri elle işaretlenmiş ve ogeturleri.txt adlı bir dosyaya kaydedilmiştir. Tablo 5'te frekansı en yüksek olan çözümlemelerden bir kesit verilmiştir.

Tablo 5: Yüksek Frekanslı Çözümlemelerin Öğe Türü Karşılıkları

Frekans	Kelime Çözümlemesi	Örnekler	Öğe
2.920.147	[ISIM_KOK]	entegre	özne
549.121	[OZEL_KOK]	Haydar	özne
509.831	[FIIL_KOK, FIIL_ZAMAN]	ulaşır	yüklem
495.712	[ISIM_KOK, ISIM_TAMLAMA_I]	ufku	özne
351.620	[FIIL_KOK]	yakın	yüklem
277.266	[SAYI_KOK]	bir	next
251.107	[FIIL_KOK, FIIL_SIFAT]	olan /kalacak	next
233.537	[ISIM_KOK, ISIM_KALMA_DE]	yasada	nerede
210.106	[ISIM_KOK, ISIM_YONELME_E]	düşmana	neye
204.758	[ZAMIR_KOK]	bu	özne
174.974	[ISIM_KOK, ISIM_TAMLAMA_IN]	padişahın	next
132.659	[ZAMAN_KOK]	gün/ gece/ kadar	zaman
125.792	[ISIM_KOK, ISIM_COGUL_LER, ISIM_BELIRTME_I]	krizleri/ sonuçları	neyi
112.129	[ISIM_KOK, ISIM_COGUL_LER]	öğrenciler	özne
111.999	[ISIM_KOK, ISIM_TAMLAMA_IN, ISIM_KALMA_DE]	takviminde	nerede

6. ogetürleri.txt dosyasındaki sözcük çözümlemeleri öğe türlerine göre gruplandırılmış ve her bir öğe türü için çeşitli kurallar elle oluşturulmuştur. Aşağıda bu kurallara ait örnekler verilmiştir.

Nasıl öğesi için;
Sözcük çözümlemesinin sonu aşağıda verildiği gibi bitiyorsa.

…FIIL_SUREKLILIK_EREK]
…ISIM_TARAFINDAN_CE]
…FIIL_OLUMSUZLUK_DEN]
…ISIM_GIBI_CE]
…ISIM_KUCULTME]
…FIIL_IMSI_IP]

Ne zaman öğesi için;

Sözcük çözümlemesinin sonu aşağıda verildiği gibi bitiyorsa.

…IMEK_ZAMAN_KEN]
…FIIL_DEVAMLILIK_DIKCE]
…FIIL_ZAMAN_INCE]
…FIIL_MASTER_CE]

VEYA Sözcük çözümlemesinin başı aşağıda verildiği gibi başlıyorsa.
[ZAMAN_KOK…

Next öğesi için;

Sözcük çözümlemesinin sonu aşağıda verildiği gibi bitiyorsa.

…FIIL_SIFAT]
…ISIM_TAMLAMA_IN]
…ISIM_BULUNMA_KI]
…FIIL_GEN]
…SAYI_SIRA_INCI]
…ISIM_ILISKI_SEL]
…ISIM_YOKLUK_SIZ]

7. Bilindiği gibi Türkçede bir sözcüğün birden fazla morfolojik çözümü vardır. Zemberek bu çözümlerin tümünü vermektedir. Bu çalışmada; bu çözümlerden ilki kullanıldığından bazı sözcükler için doğru çözümlemeler kullanılamamıştır. Bu hatayı gidermek için frekansı yüksek olan sözcüklerin Zemberek çözümlemeleri incelenmiş; yanlış çözümlemelerin doğruları, sözcükler ile birlikte bir dosyaya yazılmış ve tümcelerin öğelenmesi sırasında bu sözcüklerden biri geldiğinde, 6.adımda oluşturulan öğelere ayırma kuralları uygulanmadan direkt öğe türü belirlenmiştir. Tablo 6'da Zembereğin çözümlemeleri ve bu sözcükler için tarafımızca belirlenen öğe türleri verilmiştir.

Tablo 6: Veri Kümemizde En Fazla Yer Alan İlk 15 sözcük İçin Sözcüklere Atanan Öğe Türleri

Frekans	Sözcük	Çözümlemesi	Öğe Türü
182.759	Ve	[ISIM_KOK]	next
68.193	İçin	[ISIM_KOK]	sebep
35.944	Çok	[ISIM_KOK]	next
31.603	daha	[ISIM_KOK]	next
27.364	bin	[FIIL_KOK]	next
26.418	yüzde	[ISIM_KOK]	next
25.605	en	[ISIM_KOK]	next
22.072	ilk	[ISIM_KOK]	next
18.726	büyük	[ISIM_KOK]	next
17.478	yıl	[FIIL_KOK]	zaman
16.083	yeni	[ISIM_KOK]	next
15.776	son	[SAYI_KOK]	next
15.696	devam	[ISIM_KOK]	next
14.857	iyi	[ISIM_KOK]	next
14.557	her	[ISIM_KOK]	next

8. tümceler.txt dosyasındaki tüm tümcelerdeki sözcüklere, 6. ve 7. adımda oluşturulan kurallar ile öğe türleri atanmıştır.

9. "next" öğe türüne atanan sözcükler, "next" olmayan ilk sözcüğe kadar birleştirilmişlerdir. Ayrıca sondaki next türündeki sözcükler silinmiştir. Aşağıdaki örnekte ilk çözümlemede her sözcüğe karşılık bir öğe türü atanmışken, next'lerin birleştirilmesiyle NEYİ öğesine 3, YÜKLEM öğesine 2 kelime atanmıştır.

   ÖR:
   ÖZNE NEXT NEXT NEYİ NEXT YÜKLEM
   ÖZNE NEYİ YÜKLEM

10. İlk ÖZNE'den sonraki ÖZNE'ler NESNE olarak etiketlenmiştir.

    ÖR:
    ÖZNE NEYİ ÖZNE NASIL YÜKLEM
    ÖZNE NEYİ NESNE NASIL YÜKLEM

11. Aynı öğe türüne sahip kelimeler birleştirilmiştir.

    ÖR:
    ÖZNE NESNE NEYE NASIL NASIL NEYİ YÜKLEM
    ÖZNE NESNE NEYE NASIL YÜKLEM

Aşağıda örnek bir tümcenin işlenmesinin adımları verilmiştir.

Tümce: Bunu diğer liderler ifade etti.
Çözümleme:

Bunu [ZAMIR_KOK,ISIM_SAHIPLIK, ISIM_BELIRTME_I] NEYİ
diğer [ISIM_KOK] NEXT
liderler [ISIM_KOK,ISIM_COGUL_LER] ÖZNE
ifade [ISIM_KOK] ÖZNE
etti [FIIL_KOK,FIIL_ZAMAN]: YÜKLEM

NEXT'in bir sonraki kelimeyle birleşmesi sonucu:
NEYİ ÖZNE ÖZNE YÜKLEM

İkinci ÖZNE'nin NESNE olarak atanması sonucu:
NEYİ ÖZNE NESNE YÜKLEM

Sonuç Çözümleme:
NEYİ: Bunu
ÖZNE: diğer liderler
NESNE: ifade
YÜKLEM: etti

12. 8, 9, 10 ve 11.adımlarda gerçekleştirilen kelime gruplarını ve öğe türlerini belirleme işlemlerinin sonucu iki tabloya sahip ilişkisel bir veri tabanına kaydedilmiştir. İlk tabloda kelime ya da kelime dizilerine karşılık gelen id'ler tutulmuştur. İkinci tablonun her bir satırı bir tümceye, kolonları ise öğe türlerine karşılık gelmektedir. Tablo 7 ve 8'de "Çocuk okula bu sabah gitti.", "Çocuk patatesleri kızarttı." ve "Kız İstanbul'a döndü." tümceleri için oluşturulan hayali tablolar verilmiştir. Öğeler veri tabanına kaydedilirken daha genel sonuçlar elde edilebilmesi için çekim eklerinden arındırılarak kökleriyle veri tabanına kaydedilmişlerdir.

Tablo 7: Kavramlar Tablosu

İd	Kavram
1	Çocuk
2	Okul
3	bu sabah
4	Git
5	Patates
6	Kızart
7	kız
8	İstanbul
9	Dön

Tablo 8: Öğeler Tablosu

Özne	Neyi	Neye	Ne Zaman	Yüklem
1		2	3	4
1	5			5
7		8		9

Tablo 9'da öğe türlerinin 1.218.335 tümcenin kaçında yer aldıkları ve buna bağlı yüzde oranları verilmiştir.

Tablo 9: Öğe Türlerinin Tümcelerde Yer alma Sayı ve Yüzdeleri

Öğe Türü	Kaç Tümcede Yer Aldığı	Yüzdesi
ÖZNE	1.137.495	93,3
NESNE	934.502	76,7
NE_İLE	134.989	11
NEYE	449.488	36,9
NEYİ	499.267	41
NASIL	168.210	13,8
NE_ZAMAN	283.321	23,3
NEREDE	467.848	38,4
NEREDEN	200.945	16,5
NE_KADAR	2.796	0,2
SEBEP	82.374	6,7
ŞART	34.530	2,8
SORU	14.658	1,2
YÜKLEM	1.010.734	83

Tablo 9 incelendiğinde tümcelerde en fazla yer alan öğe türünün ÖZNE, en az yer alanın ise NE_KADAR olduğu görülmektedir. Ancak; niceliği oldukça büyük olan veri kümesinden elde edilmiş olsa da bu tablonun, Türkçe tümcelerin genel yapısını yansıttığını söylemek yanlış olabilir. Bunun bir sebebi yapılan öğelerine ayrılma işleminin doğruluk oranının %100 olmayışıdır. 5. bölümde bunun sebepleri anlatılmıştır. Bir diğer sebep ise burada kullanılan tümcelerin sadece haber makalelerinden alınmış olmasıdır. Farklı türde bir metin kümesi kullanılırsa bu oranlar değişebilir.

4. Elde Edilen Bilgilerin Filtrelenmesi

Hayat bilgisi veri tabanına eklenmesi için büyük miktarda aday öğe bilgisi 3.bölümün sonunda elde edilmiştir. Bunların içinde çok miktarda gürültü de bulunmaktadır. Daha kaliteli bilgilerin bulunması için, oluşturulan Öğeler tablosunda, öğe ikililerinin frekanslarına bağlı olarak bir filtreleme yapılmıştır. Öncelikle 1000'den fazla tekrar eden kavramlar filtrelenmiştir. Daha sonra kavramlar tablosuna "Kavram-1 ve Kavram-2 kaç tümcede sırasıyla Öğe Türü-1 ve Öğe Türü-2 görevlerinde kullanılmıştır?" soruları sorulmuş ve frekansı 3'den yüksek olan öğe ikilileri belirlenmiştir. Bu işlem birçok öğe türü ikilisi için tekrarlanmıştır. Bu işlem sonucu toplamda 26.877 adet öğe ikilisi elde edilmiştir.

Öğeler tablosundan hayat bilgisi veri tabanımıza eklenebilecek öğe ikilileri, ikililerin mevcut sistemdeki ilişkilere karşılıkları ve ilişkilere ait örnekler Tablo 10'da verilmiştir.

Tablo-10: Öğrencilerden CSdb'ye

Öğe ikilisi	CSdb karşılığı	İlişki Örnekleri
ÖZNE - YÜKLEM	Bu ne yapabilir ?	yargıtay-boz, hazine-borçlan, mahkeme-reddet, itfaiye-söndür, saldırgan-kaç, endeks-gerile, mahkeme-karar ertele, sıcaklık-in, hasan-şaş, cem-uzan, bomba-patla , alarm-çal, halk-seç, her şey-bit.
NESNE – YÜKLEM	Buna ne yapılır?	gurur - duy, forma-giy, sorun - çöz, özür - dile, namaz - kıl, soru - sor, anlaşma -imzala, rekor-kır, yangın-söndür, suç-işle, su-iç, sigara-iç
NEYİ – YÜKLEM	Buna ne yapılır?	sorun-çöz, bayram-kutla, soru-cevapla, borç-öde, soru-sor, saldırı-kına, kol-sıva, para-öde, maç-oyna, göz-yum, rakip-yen, vatandaş-uyar, ses-duy, anayasa-onayla, el-uzat
NEYE – YÜKLEM	Bu ne(re)ye yapılır?	duvar-çarp, uçak-bin, otobüs-bin, manşet-taşı, defter-yaz, refüj-çarp, mezar-defnet, düğme-bas, top-vur, kamuoyu-duyur, masa-otur, sokak-dök
NASIL – YÜKLEM	Bu nasıl yapılır?	vur-öldür, yan-öl, çarp-dur, don-öl, uzan-çel, kes-öldür, düş-yarala, boğ-öl, bıçakla-yarala, düş-öl
NEİLE – YÜKLEM	Bu ne ile yapılır? Bunun için ne kullanılır ?	tabanca-öldür, tabanca-yarala, kafa-vur, tüfek-öldür, sabır-bekle, göz-bak, alkış-karşıla, coşku-kutla, fünye-patla
SEBEP – YÜKLEM	Bu neden yapılır?	kazan-mutlu, son-bekle, kavuş-çok mutlu, başar-mutlu
ŞART – YÜKLEM	Bu hangi şartla yapılır?	Bak-gör, kız-söyle, gör-inan, söyle-inan, sor-söyle
NEZAMAN –YÜKLEM	Bu ne zaman yapılır?	Her yıl-katıl mayıs-kutla, önce gün-öl, geçen hafta-kazan, hiçbir zaman-unut
NEREDE – YÜKLEM	Bu nerede yapılır?	gazete-yayımla, mezar-defnet, makam-görüş, kurul-görüş, yangın-yan, film-canlan, meydan-topla, kapı-karşıla, mahkeme-yargıla, üniversite-oku, tribün-izle, fabrika-üret, fuar-sergile
ÖZNE – NEREDE	Bu nerede bulunur?	cumhurbaşkanı-köşk, cenaze-mezar, film-festival, imza-tören, yağ-karadeniz, sakat-antreman, rol-film, çocuk-yuva, gol-kale, ödül-tören, yolcu-uçak, yatak-oda, şehit-mezar, kurt-vadi, yedek-klübe
NESNE – NEREDE	Bu nerede bulunur?	oy-seçim, gol-kale, sevgi-gösteri, saldırı-Bağdat, oy-kurultay, dolar-piyasa, öğrenci-sınav, deniz-çanakkale, tedavi-servis
NEYİ – NEREDE	Bu nerede bulunur?	kanun-komisyon, senet-borsa, hazırlık-tesis, ödül-festival, avukat-mahkeme, saldırı-Irak, tedavi-hastane, hastalık-hastane, çocuk-okul, takım-stat

Tablo 10 incelendiğinde oldukça başarılı öğe ikililerinin bulunabildiği görülmektedir. Bununla birlikte üretilen sonuçların güvenilirliği bir şekilde ölçülmelidir. Ancak üretilen ilişkilerin doğruluklarını kontrol edebileceğimiz bir kaynak olmadığından bu işlem için insan yargısına başvurulmuş ve her ilişki türü için bir ilişkiler alt kümesi doğru ya da yanlış olarak işaretlenmiştir. Tablo 11'de ilişki türlerinin doğruluk oranları verilmiştir.

Tablo 11: İlişkilerin Doğruluk Oranları

İlişki ismi	Frekansı 3'den büyük olan ilişki sayısı	İncelenen ilişki sayısı	Doğru ilişki sayısı	Doğru ilişki yüzdesi
ÖZNE – YÜKLEM	4006	241	38	0,16
NESNE – YÜKLEM	1896	102	40	0,39
NEİLE – YÜKLEM	1635	163	49	0,30
NEYE – YÜKLEM	3242	301	146	0,49
NEYİ – YÜKLEM	2918	283	172	0,61
NASIL – YÜKLEM	1441	156	44	0,28
NE ZAMAN – YÜKLEM	1432	149	25	0,17
NEREDE – YÜKLEM	2594	165	66	0,40
SEBEP – YÜKLEM	135	135	5	0,04
ŞART – YÜKLEM	345	152	6	0,04
ÖZNE – NEREDE	4292	282	57	0,20
NESNE – NEREDE	1152	249	34	0,14
NEYİ – NEREDE	1789	297	72	0,24

Tablo 11 incelendiğinde en yüksek doğruluk oranlarına sahip öğe ikililerinin sırasıyla NEYİ – YÜKLEM, NEYE – YÜKLEM, NEREDE – YÜKLEM ve NESNE – YÜKLEM olduğu görülmektedir. Tüm ikililer şekilsel kurallar ile bulunduğundan, düşük doğruluk oranına sahip ikililer için şekilsel kuralların yeterli olmadığı, bu tür öğe ikililerini belirlemede kelime anlamlarının da etkin olduğu sonucuna ulaşılabilir. Doğruluk oranının düşük olma sebepleri bir sonraki bölümde detaylandırılmıştır.

5. Sistemin Eksiklikleri

Gerçeklenen öğelerine ayırma sisteminin mevcut eksiklikleri ve olası çözüm yolları bu bölümde anlatılmaktadır.

1. Kelime anlamlarının çözümlemeye etkisi: Tümcelerin öğelerinin bulunmasında kelimelerin çözümlemeleri yeterli değildir. Örneğin "Çocuk düştü." ve "Pilav yedi." tümcelerindeki kelime çözümlemeleri tamamen aynı olmalarına ([ISIM_KOK] [FIIL_KOK, FIIL_ZAMAN]) karşın, ilk tümcenin öğe dizilişi "ÖZNE YÜKLEM" iken ikincininki "NESNE YÜKLEM" şeklindedir. Bu ayrım sadece kelimelerin anlamları kullanılarak yapılabilir. Geliştirilen sistem sadece kelime çözümlemelerini kullandığı için, örnekteki her iki tümceyi de aynı şekilde öğelerine ayırmaktadır. Şüphesiz; öğe türlerinin doğru belirlenebilmesi için kelime anlamlarına duyulan gereklilik sadece ÖZNE YÜKLEM ve NESNE YÜKLEM öğe ikililerini değil, tüm öğe ikililerini etkilemektedir. Bu problemin çözümünde, Türkçe için bu kapsamda geliştirilmiş anlamsal ağlar (Türkçe Wordnet [6], CSdb [3]) kullanılabilir.

2. Basit kural tabanlı bir sistemin kullanımı: Öğelerine ayırma işleminde sezgisel olarak elle belirlenen kurallar yerine istatistiksel bir modelin (Gizli Markov Süreçleri vb.) kullanımı, öğelerin doğru belirlenme oranını arttırabilir. Ancak, bunun için elle etiketlenmiş tümce öğelerine ihtiyaç bulunmaktadır. Türkçe'de bu amaçla oluşturulan ODTÜ-SABANCI ağaç derlemi [9] bulunmaktadır.

3. Aynı kelimenin birden fazla şekilde çözümlenebilmesi*:* Zemberek kütüphanesinde kelime anlamı durulaştırma işlevi olmadığından, birden fazla şekilde çözümlenebilecek kelimelerde Zembereğin ilk ürettiği çözümleme doğru kabul edilmiştir. Buna örnek olarak Tablo 3'te yer alan "yeme" kelimesi verilebilir. Bu kelime içinde bulunduğu tümceye göre aşağıdaki 3 şekilde çözümlenebilir.

1-yem[isim_kok]-e[isim_yonelme_e]
2-ye[fiil_kok]-me[fiil_olumsuzluk_me]
3-ye[fiil_kok]-me[fiil_dönüşüm_me]

Ancak Zemberek kelimeleri tek başına değerlendirdiğinden, ilk olarak ürettiği 1. çözüm doğru olarak kabul edilmiştir. Bu gibi problemlerin

çözülebilmesi için ya Zemberek'e kelime anlamı durulaştırma modülü eklenmelidir ya da Zemberek yerine bu özelliği olan bir araç [10] kullanılmalıdır.

4. Kompleks tümceler: Sistem, sadece 15'ten az sayıda kelime içeren tümceler üzerinde çalıştırılmıştır. Bunun sebebi çok fazla öğe içeren tümcelerin öğelerine ayrılma işleminin tek fiil içerenlere göre oldukça zor olmasıdır. Bunun için tümcedeki tümcecikler ve bu tümceciklerin tümcedeki görevleri de belirlenmelidir. Ancak sistemin geliştirilme amacı her türlü tümceyi öğelerine ayıran bir yazılım geliştirmekten ziyade, hayat bilgisi veri tabanımıza veri sağlamak olduğundan, büyük miktarda veri üzerinde basit kurallarla hızlı çalışan bir sistem tasarlanmış ve gerçekleştirilmiştir.

6. Sonuçlar

Chris Riesbeck, yapay zekanın tanımını "bilgisayarların neden bu kadar aptal olduklarını bulmak" olarak vermektedir. Bu soruya verilebilecek bir cevap bilgisayarların bulundukları dünya ve insanların yaşantısı hakkında bilgilerinin olmamasıdır. Tüm insanların yaşayarak öğrendikleri, her zaman kullandıkları ve içerdiği bilgi miktarı milyonlarla ifade edilen bu bilgi bütününe ortak hayat bilgisi (commonsense knowledge) denmektedir. İnsanların kullandıkları çıkarsama mekanizmaları formel olarak ifade edilebilmelerine rağmen bilgisayarlar, insanların sahip oldukları hayat bilgisi veri tabanına sahip olmadıkları için onların kullanıcıları olan bizlerin isteklerimizi, anlamamakta ve her şeyi bizden beklemektedirler. Örneğin; bir arkadaşımıza kedimizin hasta olduğunu söylediğimizde, bize hemen tanıdığı bir veterinerin telefonunu verebilir. Arkadaşımız sahip olduğu hayat bilgisi sayesinde bu çıkarımı kolaylıkla yapabilmektedir. Bu çıkarım işleminin şekilsel yapısı oldukça basittir ve sonuç olarak bilgisayarlarımızın bu çıkarımı yapamıyor olmasının sebebi olarak geriye bir tek onların hayat bilgisinden yoksun olmaları kalmaktadır. Bu fikirden hareketle İngilizce için çeşitli hayat bilgisi veri tabanları oluşturulmaya başlanmıştır. Türkçe için de benzer çalışmalar mevcuttur. Bu amaçla başlattığımız Türkçe hayat bilgisi veri tabanımıza (CSdb) yeni bilgiler eklemek amacıyla yapılan bu çalışmada Bilcol [4] derleminden elde edilen 1.218.335 adet tümce üzerinde öğelere ayırma işlemi gerçekleştirilmiş, sıkça kullanılan 26.877 adet öğe ikilisi bulunmuştur. Sistemin ara yüzüne www.kemikoyun.yildiz.edu.tr/commonsense adresinden erişilebilir. Sistemin ara yüzünde veri tabanındaki bilgilerin doğruluklarını arttırmak için bir oyun (CSoyun) bulunmaktadır. Kullanıcılar sistemdeki bilgilere kendilerince doğruluklarına göre puan vermektedir. Bir bilginin güvenilirliği o bilgiye verilen puanların ortalaması ile belirlenmektedir. Bu mekanizma sayesinde sistemdeki bilgilerin güvenilirlikleri oyun oynandıkça artmaktadır. Bu mekanizma sayesinde sisteme eklenen bilgilerin hepsinin doğru olmasına gerek yoktur. Yanlış olanların güvenilirlikleri oyun oynandıkça düşecektir. Kullanıma açılması planlanan hayat bilgisi veri tabanında sadece güvenilirlik değeri yüksek bilgilere yer verilecektir.

Sonuç olarak; bu çalışma ile, çeşitli kaynaklarla beslemeye devam ettiğimiz CSdb projesinde öğe ikililerinden bilgi kaynağı olarak faydalanabileceği görülmüştür. CSdb projesinin Türkçe doğal dil işleme ile ilgilenen tüm araştırmacılara metin anlama, otomatik metin çevirisi, kelime anlamı durulaştırma, metin sınıflandırma vb. konularda hizmet etmesi hedeflenmektedir.

7. Kaynakça

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

[2] **Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T. ve Zhu, W.L.,** 2002. "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.

[3] **Bilgin O., Çetinoğlu Ö. ve Oflazer K.,** 2004. "Building a Wordnet for Turkish", *Romanian Journal of Information Science and Technology*, cilt 7, 1-2.

[4] **Güngör, O. ve Güngör, T.,** 2005. " Türkçe için Bilgisayarla İşlenebilir Sözlük Kullanarak Kavramlar Arasındaki İlişkilerin Belirlenmesi", *Akademik Bilişim Konferansı*, 2007.

[5] **Amasyalı M.F.,** "Türkçe Wordnet'in Otomatik Olarak Oluşturulması", *Sinyal İşleme ve Uygulamaları Sempozyumu*.

[6] **Amasyalı, M.F., İnak, B. ve Ersen, M.Z.,** 2010. "Türkçe Hayat Bilgisi Veri Tabanının

Oluşturulması", *Akademik Bilişim Sempozyumu*.

[7] **Can, F., Koçberber, S., Bağlıoğlu, O., Kardaş, S., Öcalan, H.C., Uyar, E.,** 2009. "Türkçe haberlerde yeni olay bulma ve izleme: Bir deney derleminin oluşturulması", *Akademik Bilişim Sempozyumu*.

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

[9] **Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, Gökhan Tür**, 2003. Building a Turkish Treebank, Invited chapter in Building and Exploiting Syntactically-annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers.

[10] **Haşim Sak, Tunga Güngör, Murat Saraçlar,** 2007. Morphological Disambiguation of Turkish Text with Perceptron Algorithm, CICLing 2007, LNCS 4394, 107–118.