Türkçe Bir Sözlükteki Tanımlardan Kavramlar Arasındaki Üst-Kavram İlişkilerinin Çıkarılması

Onur Güngör, Tunga Güngör
Boğaziçi Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul
onurgu@boun.edu.tr, gungort@boun.edu.tr

Özet: Bu bildiride, bir sözlükteki sözcüklerin arasındaki anlamsal ilişkileri çıkaran ve hiyerarşik bir yapı oluşturan kural tabanlı bir yöntem sunulmaktadır. Yöntemdeki ana kurallar üç gruba ayrılabilir: sözcüğün yüzey biçimini kullanan kurallar, sözcüğün kategorisini kullanan kurallar ve sözcüğün tanımını kullanan kurallar. Oluşturulan hiyerarşinin kök düğümleri, İngilizce WordNet veri tabanından alınmıştır. Üst-kavram çıkarma oranı yaklaşık %94 olarak tespit edilmiştir. Hiyerarşinin içeriği ve eksiklikleri tartışılmış, Türkçe WordNet ile karşılaştırılmıştır.

Anahtar Kelimeler: Anlamsal ilişkiler, bilgisayarla işlenebilir sözlükler.

Extracting Hypernymy Relations Between Concepts From The Definitions in a Turkish Dictionary

Abstract: In this paper, we present a rule-based method in order to extract semantic relations between words in a dictionary and build a hierarchical structure. The main rules used in the method can be divided into three groups: rules that use the surface form of the word, rules that use the category of the word, and rules that use the definition of the word. The root nodes of the hierarchy built were taken from English WordNet. The hypernym extraction ratio was observed around 94%. The contents and the deficiencies of the hierarchy were discussed and it was compared with Turkish WordNet.

Keywords: Semantic relations, machine-readable dictionaries.

1. Giriş

Doğal dil işleme (DDİ) sistemleri, günümüzde çoğunlukla metinleri biçim bilimsel (*morphological*) ve söz dizimsel (*syntactic*) açılardan analiz etmekte, anlam bilimsel (*semantical*) özellikleri dikkate almamaktadır. Anlam bilimsel çıkarımların yapılabilmesi için, diğer kaynaklara ilave olarak, dildeki sözcükler ve kavramlar arasındaki anlamsal ilişkileri tutan bir veri tabanına ihtiyaç duyulmaktadır. Örneğin, bir bilgisayarlı çeviri (*machine translation*) sisteminin, kaynak dildeki bir sözcüğün hedef dildeki iki olası anlamı arasında seçim yaparken, kavramlar arasındaki kısıtlamaları dikkate alması sistemin başarısını arttırabilir. Bu duruma somut bir örnek olarak, bir cümlenin söz dizimsel öğelerine ayrıştırıldığı (*parsing*) ve cümlenin öznesinin canlı bir varlık olduğu düşünülsün. Çevirinin yapılacağı dilde özneye karşılık gelen birden çok kavram mevcutsa, bu kavramlar arasından canlı bir varlığı simgeleyenin seçilmesi gerekmektedir. Sistemin anlamsal ilişkileri kapsayan uygun bir veri tabanı tarafından desteklenmesi durumunda, veri tabanını sorgulayarak öznenin olası anlamları arasından canlı nesne özelliğine sahip olanını seçmek mümkün olacaktır.

Literatürde bu tür veri tabanlarını oluşturmak için çeşitli çalışmalar bulunmaktadır. Bu veri tabanları içerisinde en bilineni, isim, fiil ve sıfat kökenli sözcükler için eş anlam kümeleri (*synonym set – synset*) ve bunlar arasındaki bazı anlamsal ilişkileri içeren WordNet'tir [1].

WordNet'in ilk sürümü 5 yıl süren bir çalışmanın ürünüdür. İlk sürümde yaklaşık 95.600 sözcük biçimi (bunların yaklaşık olarak yarısı iki veya daha çok sözcükten oluşan öbeklerdir) 70.000 eş anlam kümesine ayrıştırılmıştır. Anlaşılacağı üzere, bu tür veri tabanlarını elle geliştirmek oldukça büyük miktarda insan emeği ve zamanı gerektirmektedir.

Bu bildiride, bir sözlükteki sözcükleri otomatik olarak analiz ederek anlamsal bir hiyerarşik yapı oluşturan bir yöntem anlatılmaktadır. Bu hiyerarşideki düğümler (*nodes*) birbirlerine alt-kavram ve üst-kavram ilişkileriyle bağlanırlar. Bu çalışmada, Türk Dil Kurumu (TDK) tarafından yayımlanmış olan güncel Türkçe sözlüğün elektronik sürümü kullanılmıştır [2]. Bu çalışma, Türkçe bir sözlükteki sözcüklerin arasındaki kavramsal ilişkileri kullanarak tamamen otomatik olarak alt-kavram/üst-kavram hiyerarşisi oluşturan ilk çalışmadır. Buna ilave olarak, sözcükler arasındaki eş anlamlılık ilişkileri de çıkarılmaya çalışılmıştır.

Chodorow ve Byrd tarafından İngilizce için yapılan bir çalışmada, isim kökenli ve fiil kökenli sözcükler dikkate alınmış, birincisi için tanımın içindeki isim öbeği, ikincisi için ise tanımın içindeki fiil öbeği çıkarılmıştır [3]. Bu öbeklerdeki ana sözcüğün (*head word*) üst-kavram olduğu varsayılarak, bu sözcüğün tespit edilmesine çalışılmıştır.

Başka bir araştırmada, üst-kavram ilişkilerinin çıkarılmasına yönelik çalışmalarda sadece tek bir sözlükten yararlanmanın yetersiz olduğu öne sürülmüştür [4]. Birden fazla sözlük kullanıldığında ise bu hataların önemli ölçüde azaldığı ifade edilmiştir.

Sözlüklerden anlamsal ilişkilerin otomatik olarak çıkarılması üzerine yapılan araştırmaların zenginleşeceği ve tatmin edici sonuçlar alınacağı yönündeki görüşlere karşın, bazı araştırmacılara göre bu konudaki araştırmalar beklenen niteliğe ulaşamamıştır [4] ve [5]. Bu durumun, büyük oranda sözlüklerdeki eksik tanımlara ve tutarsızlıklara bağlandığı görüşü öne sürülmüştür.

Sözlük tanımlarının çeşitli örüntüler şeklinde temsil edildiği ve bu örüntüler arasında bir hiyerarşinin kurulduğu bir çalışma [6]'de verilmiştir. Bir tanıma karşılık gelen örüntü önce daha özgül örüntüler içerisinde aranmakta, bulunamadığı durumlarda daha genel örüntülerle eşleştirmek mümkün olmaktadır.

2. Yöntem

Bu çalışmada, üst-kavram ve alt-kavram ilişkilerini içeren hiyerarşik bir yapının yaratılması amacıyla iki temel aşama uygulanmıştır. İlk aşamada, sözlükteki bütün isim kökenli sözcüklerin üst-kavramları, bu sözcüklerin sözlük tanımlarına üst-kavram çıkarma algoritması uygulanarak toplanmıştır. Çıkarılan üst-kavramlar ikinci aşamada kullanılmak üzere bir dizinde tutulmaktadır. İkinci aşamada ise, birinci aşamada oluşturulan dizin kullanılarak hiyerarşik yapı elde edilmiştir. Bahsedilen aşamalardan ilki 2.1. bölümde, ikincisi 2.2. bölümde anlatılmaktadır.

2.1. Üst-kavramların Çıkarılması

Sözcüklerin sözlük tanımlarından üst-kavramların çıkarılması için, buluşsal bir yönteme (*heuristics*) dayanan bir algoritma geliştirilmiştir. İlk olarak, analiz edilmekte olan tanım, ayırıcı olarak virgül karakteri kullanılarak parçalara bölünür. Sözlükteki tanımlar, aşağıda düzenli gramer (*regular grammar*) biçiminde belirtilen genel örüntüyü (*pattern*) izlemektedir:

sözcük : (sözcük* üstkavram) (, sözcük* üstkavram)* (, eşanlamlı)*.

Tanım parçalara ayrıldıktan sonra, en son parçadan başlanarak en baştaki parçaya doğru bazı kurallar uygulanarak ilerlenmektedir. Bir üst-kavram bulunduğu zaman bu kuralların uygulanması durmaktadır. Bir sözcüğün eş anlamlılarının her zaman üst-kavramlarından sonra gelmelerinden dolayı, süreci bu noktada

durdurmak eş anlamlıların çıkarılmasında sorun yaratmamaktadır.

Türkçe sondan eklemeli (*agglutinative*) bir dil olduğundan, sözlük tanımlarında üst-kavramlar ve eş anlamlı sözcükler genellikle ekli olarak bulunmaktadır. Bu özellikten dolayı, olası üst-kavramlar ve eş anlamlılar kurallar tarafından tespit edildikten sonra, bunları sözlük tanımlarının yapısını da dikkate alarak biçim bilimsel olarak analiz eden üst-kavram seçme kriterleri (ÜSK) uygulanır. Bahsedilen kriterlerin detayları 2.1.2. bölümde verilmiştir.

Aşağıda, "dörtgen" sözcüğüne ait sözlükte yer alan bilgiler gösterilmiştir. Geliştirilen yöntem tarafından tanım analiz edildiğinde, ilk olarak tanım iki parçaya ayrılır: "dört kenarlı çokgen" ve "dörtkenar". Olası üst-kavram olarak "çokgen" ve olası eş anlam olarak "dörtkenar" sözcükleri bulunduktan sonra, ilkine ÜSK uygulanır ve üst-kavram olarak "çokgen" sözcüğü elde edilir. Ayrıca, 2.1.1. bölümde tanımlanacak Kural 10 kullanılarak "dörtkenar" sözcüğü "dörtgen" sözcüğünün eş anlamlısı olarak kaydedilir.

Sözcük: Dörtgen
Sözlüksel kategori: İsim, geometri
Tanım: Dört kenarlı çokgen, dörtkenar.

2.1.1. İlişkilerin Tespitinde Kullanılan Kurallar

Sözlükteki isim kökenli sözcüklerin tanımlarının dikkatlice incelenmesi sonucunda, sözcükler arasındaki anlamsal ilişkileri çıkarmak üzere çeşitli kurallar belirlenmiştir. Bu kurallar üç gruba bölünebilir:

a) İsmin yüzey biçimine (*surface form*) göre üst-kavramı belirleyen kurallar,
b) İsmin sözlükte belirtilen kategorisine göre üst-kavramı belirleyen kurallar,
c) İsmin sözlükteki tanımına göre üst-kavramı belirleyen kurallar.

Buna göre 11 adet kural belirlenmiştir. Birinci ve ikinci gruplara ait sadece birer kural vardır; diğer dokuz kural üçüncü grubu oluşturmaktadır. Üçüncü grubun en önemli kurallarından Kural 10 Tablo 1'de gösterilmiştir.

Tablo 1 Kural 10

Kural 10: İncelenen parça önceki kuralların aradığı sözcük öbekleriyle bitmiyorsa, öncelikle bütün parçanın sözlükte olup olmadığı kontrol edilir. Sözlükte bulunuyorsa, bu parçayı oluşturan sözcük veya sözcükler işlenen sözcüğün eş anlamlıları olarak kabul edilir. Aksi takdirde, parçanın son sözcüğüne ÜSK uygulanıp üst-kavram belirlenir. *Örnek:* Sözcük: Satım Sözlüksel kategori: İsim, ticaret Tanım: Satma işi, satış. "Satış" sözcüğü (son parçanın tümü) sözlükte yer aldığı için, "satım" sözcüğünün eş anlamlısı olarak belirlenir.

2.1.2 Üst-kavram Seçme Kriterleri

Bir sözcüğe ait olası üst-kavramlar önceki bölümde verilen kurallar ile çıkarıldıktan sonra, üst-kavram seçme kriterleri (ÜSK) olarak adlandırılan bir analizden daha geçirilmektedir. ÜSK, temel olarak, sözcüğün bir üst-kavram olup olamayacağını ve eğer oluyorsa sözcüğün üst-kavram olarak kullanılması gereken biçimini belirler.

Tanımdan elde edilen sözcük, biçim bilimsel bir analizden geçirilmektedir. Bu çalışmada, biçim bilimsel analiz programı olarak Zemberek sistemi kullanılmıştır [7]. Zemberek, Türkçe'nin doğal dil işleme yöntemleri vasıtasıyla işlenmesi sırasında ortaya çıkan bilişsel problemlerin çözümünü kolaylaştırmayı amaçlayan bir program kütüphanesinden ve uygulama programlarından oluşmaktadır. Çalışmamızda Zemberek'in çıktı olarak verdiği sözcüğün olası ayrıştırmalarından sadece ilki kullanılmaktadır.

Biçim bilimsel analiz programından elde edilen ayrıştırma, Grup 1, Grup 2 ve Grup 3 adı verilen kriterler ile karşılaştırılmaktadır. Bu kriterlerden birine uyması durumunda, karşılık

gelen sözcük, incelenmekte olan sözcüğün üst-kavramı olarak hiyerarşik yapıya eklenmektedir. Aksi halde, sözcüğün üst-kavram olma özelliği taşımadığı sonucuna varılmaktadır. Tablo 2'de Grup 2 kriterlerinin ayrıntılı bir incelemesi bulunmaktadır.

Tablo 2 ÜSK 2. Grup

Analiz 1:
fiil kökü + (bir veya daha fazla ek) + isimfiil (eylemlik) eki
Örnek:
(...) eşleme
eşle + -me
Üst-kavram: eşleme

Analiz 2:
isim kökü + yapım eki (bulunma eki)
Örnek:
(...) kitaplık
kitap + -lık
Üst-kavram: kitaplık

Analiz 3:
isim kökü + yapım eki (durum eki)
Örnek:
(...) iyilik
iyi + -lık
Üst-kavram: iyilik

Analiz 4:
isim kökü + yapım eki
Örnek:
(...) kitapçı
kitap + -çı
Üst-kavram: kitapçı

İşlem:
Eğer sözcük yukarıdaki analizlerden birine uyuyorsa, üst-kavram biçimi olarak sözcüğün tümü belirlenir.

2.1.3 Eş anlamlı Sözcüklerin Çıkarılması

Bir sözcüğün tanımı içinde yer alan eş anlamlı sözcükler, 2.1.1. bölümde bahsedilen ve Tablo 1'de gösterilen kural yardımıyla tespit edilmektedir. Eş anlamlı sözcükler, eş anlamlılık kümelerinde (*synonym set*) toplanmaktadır: bir küme, birbirlerine eş anlamlılık ilişkisi ile bağlı bütün sözcükleri kapsamaktadır. Fakat bir kümeye kümedeki genel kavramla ilintili olmayan sözcüklerin eklendiği durumlar da olmaktadır. Örneğin, algoritma tarafından elde edilen eş anlamlılık kümelerinden biri aşağıda gösterilmiştir:

{tertip, düzenleme, kura, ...}

Sözlükte yer alan tanımlara göre, "tertip" ile "düzenleme" sözcükleri arasında ve "tertip" ile "kura" sözcükleri arasında eş anlamlılık ilişkileri mevcuttur. Fakat "tertip" sözcüğünün farklı anlamları diğer iki sözcük ile eş anlamlıdır. Buna göre, "düzenleme" ve "kura" sözcükleri arasında bu tür bir ilişki mevcut değildir. Buradaki hata, yukarıda bahsedildiği gibi, bir sözcüğe ait anlamların bir bütün olarak ele alınmasından kaynaklanmaktadır.

2.2 Hiyerarşik Yapının Kurulması

2.1. bölümde anlatıldığı şekilde, algoritmanın ilk aşamasında sözlük tanımlarından çıkarılan üst-kavramlar bir indeks yapısında tutulmaktadır. Bu indeks yapısının her bir kaydında, bir sözcük ve sözcüğün üst-kavramı yer almaktadır. Algoritmanın ikinci aşamasında ise, bu indeks yapısı taranarak kavramsal ilişkilerin hiyerarşik yapısı oluşturulacaktır.

Hiyerarşinin en üst noktasındaki düğümlerin, İngilizce WordNet'teki en üst düğümlerin (kavramların) Türkçe çevirimleri olmasına karar verilmiştir. Bunun sebebi, İngilizce WordNet'in oldukça titiz çalışmalar sonucunda ortaya çıkarılmış tutarlı ve hemen hemen hatasız bir veri tabanı olması ve en üstte yer alan kavramların dildeki diğer bütün kavramları kapsayacak şekilde seçilmiş olmasıdır. İngilizce WordNet'te bu özelliği taşıyan 25 kavram bulunmaktadır [8]. İki dil arasındaki farklılıklardan dolayı, İngilizce bazı kavramların Türkçe sözlükte karşılıkları bulunmamaktadır. Buna göre, Türkçe'de oluşturulan hiyerarşik ilişkiler yapısında en üstte 21 düğüm yer almaktadır.

Hiyerarşik yapı, tabloda belirtilen kavramlardan başlanarak, hazırlanmış olan indeks yapısının derinliğine arama (*depth first search*) metodu ile taranması ile oluşturulmuştur. Tarama işlemi sırasında, daha önce ziyaret edilmiş bir

düğüm tekrar ziyaret edilmemiş ve böylece hiyerarşide çevrimler (*cycle*) oluşması önlenmiştir. Ayrıca belirtilmesi gereken bir husus, hiyerarşinin en üst noktasında birden fazla kavram olduğundan dolayı, sonuçta elde edilen yapının tek bir ağaç değil, pek çok ağaçtan oluşan bir orman (*forest*) olduğudur.

3. Sonuçlar ve Tartışma

Uygulanan yöntem sonucunda oluşan üst-kavram ilişkilerini gösteren hiyerarşideki kavramlar tek bir sözcükten oluşabildiği gibi, birden çok sözcüğü içeren ifadelerden de oluşabilmektedir (Şekil-4). Hiyerarşik yapının içerisinde ifadelerin yer alabilmesi özelliği, bu çalışmayı, hiyerarşideki elemanların sözcüklerle sınırlı tutulduğu literatürdeki diğer çalışmalardan ayırmaktadır. Üst-kavram ilişkilerini çıkaran algoritma tarafından taranan yaklaşık 83.000 kavramdan 78.000 tanesi için en az bir üst-kavram bulunmuştur. Diğer bir deyişle, üst-kavram çıkarma oranı %94 olmuştur. Üst-kavram olarak bir kereden fazla geçen kavramlar teke indirildiğinde, 60.000 farklı üst-kavram olduğu tespit edilmiştir. Sözlüklerdeki eksik tanımlardan ve tutarsızlıklardan dolayı 2.1. bölümde çıkarılan üst-kavramların tümünün hiyerarşik yapıda bulunmadığı gözlenmiştir.

Hiyerarşi 72 seviyeden oluşmaktadır. Seviye sayısının fazla oluşunun nedeni, bir sözcüğün sözlük tanımından o sözcüğün üst-kavramı bulunduğunda, hiyerarşik yapıda sözcüğün üst-kavramın o bağlamda taşıdığı anlamı yerine üst-kavramın bütün sözlük anlamlarına bağlanmasıdır. Bu da hiyerarşide gerçekte olmaması gereken bağlar oluşturduğundan dolayı seviye sayısını arttırmaktadır. Başka bir neden ise sözlükte yer alan tanımların belirli bir standarda sahip olmaması ve dolayısıyla üst-kavramların aynı olması gereken durumlarda farklı üst-kavramlar tespit edilerek hiyerarşik yapıya eklenmesidir.

Hiyerarşik yapıda en fazla alt-kavramı olan sözcük, yaklaşık olarak 7.700 alt-kavrama sahip olan "iş" sözcüğüdür. Bu sözcüğün yüksek sayıda alt-kavrama sahip olmasının nedeni, TDK sözlüğünün hemen her fiil için, fiilden oluşan ve "işi" sözcüğü ile tanımlanan isim kökenli bir sözcüğü de içermesidir (örneğin, "okuma: okumak işi"). Bu tür üst-kavramlar algoritmada Grup 1 ÜSK ile çıkarılmaktadır.

Üst-kavram ilişkilerini içeren hiyerarşik yapının hatasız olarak kurulabilmesi için, bir sözcüğün üst-kavramının doğru olarak tespit edilmesine ek olarak, bu üst-kavramın sözlükteki hangi anlamının ilgili tanımdaki kullanıma karşılık geldiği de belirlenmelidir.

Üst-kavramların anlamları tespit edilmeden üst-kavram/alt-kavram ilişkileri hiyerarşik yapıya eklenirse, bir düğümün altında, gerçekte o düğümde ifade edilen sözcüğün farklı anlamlarının alt-kavramları olan sözcüklerin hepsi görünecektir. Bu durum, hiyerarşideki seviye sayısının artmasına yol açacağı gibi, yanlış üst-kavram/alt-kavram ilişkilerinin ortaya çıkmasına da neden olacaktır. Aşağıda bazı sözcüklerin sözlük tanımları ve Şekil-1'de de algoritma tarafından oluşturulan yapı verilmiştir:

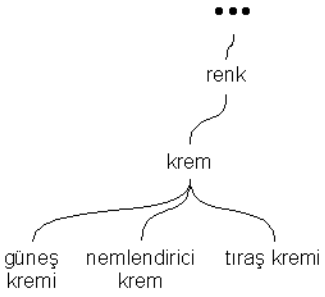
krem: 1. Tene yumuşaklık vermek veya güneş, yağmur vb. dış etkilerden korunmak için sürülen koyu kıvamlı madde. 2. Açık saman rengi.

güneş kremi: Güneşlenme sırasında cildin kurumasını, aşırı yanmasını ve çatlamasını önleyen bir tür özel krem.

"Krem" sözcüğünün birinci anlamının üst-kavramı "madde", ikinci anlamının üst-kavramı ise "renk" olarak bulunur; "güneş kremi" sözcüğünün üst-kavramı da "krem" olarak tespit edilir. Buna göre, güneş kremi bir çeşit kremdir, fakat sözlükte "krem" sözcüğünün hangi anlamına bağlanması gerektiği açık olarak belirtilmemiştir. Algoritma, anlam muğlaklıklarını çözmeden bulunan ilişkileri hiyerarşik yapıya yansıttığında, Şekil-1'de görülen durum oluşur: "krem" sözcüğünün her iki anlamının

alt-kavramları da tek bir düğüm altında toplanmıştır. Bu yapıdaki bağlantıları takip ederek, güneş kreminin bir renk çeşidi olduğu şeklindeki hatalı çıkarıma varmak olasıdır.

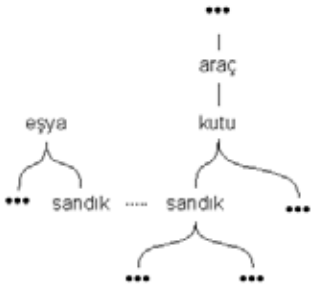
Sözlük tanımlarında bir sözcüğün veya sözcük grubunun birden fazla üst-kavramının olması olasıdır. Bu tür durumlarda, bulunan üst-kavram/alt-kavram bağlantılarının olduğu gibi hiyerarşik yapıya yansıtılması, bu yapının ağaç olma özelliğini bozacak ve onu bir çizge (*graph*) şekline dönüştürecektir. Veri yapıları ile ilgili konular üzerinde çalışan kişiler tarafından bilindiği gibi, arama (*search*) ve dolaşma (*traversal*) algoritmalarının performansı açısından, ağaç yapısının çizge yapısına göre oldukça önemli üstünlükleri vardır. Bu nedenle, bu çalışmada, hiyerarşik yapının ağaç özelliğinin korunması tercih edilmiştir.



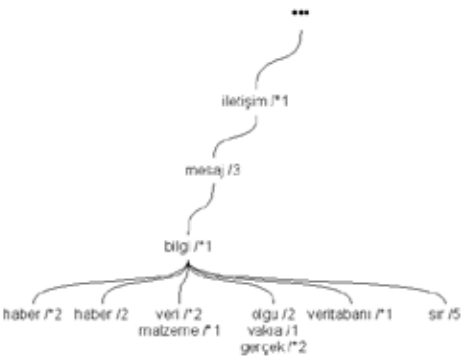
Şekil-1 'Güneş kremi' ve 'krem'in hiyerarşideki yerleri

Bir sözcüğün birden fazla üst-kavramı olduğu durumda, sözcüğe ait düğümün üst-kavramlara karşılık gelen birden fazla üst düğüme bağlanması yerine, sözcük için üst-kavram sayısı kadar düğüm yaratılmakta ve her bir düğüm ayrı bir üst-kavram üst düğümüne bağlanmaktadır. Bununla beraber, yapıda tekrarlamalara yol açmamak için, bu sözcüğün alt-kavramları, sözcüğe ait düğümlerden sadece bir tanesinin altında listelenmektedir. Şekil-2'de bir örnek verilmiştir. "Sandık" sözcüğünün iki üst-kavramı bulunmaktadır: "eşya" ve "kutu". Bu nedenle, "sandık" sözcüğü hiyerarşik yapıda iki düğüm ile simgelenir ve her biri üst-kavramlardan birine bağlanır. "Sandık" sözcüğünün alt-kavramları ise bu iki düğümden sadece birine bağlanır ve diğer düğüm için tekrarlanmaz.

Bildirinin önceki bölümlerinde değinildiği gibi, Türkçe WordNet, Türkçe sözcükler arasındaki çeşitli anlam bilimsel ilişkileri (eş anlamlılık, üst-kavram, alt-kavram vs.) içeren geniş bir veri tabanıdır [9]. Bir dil bilimsel ilişki içerisinde yer alan sözcüklerin anlam muğlaklıkları elle düzeltilmiş ve insan kontrolü altında hiyerarşik yapı hazırlanmıştır. Bu nedenle, Türkçe WordNet'in büyük ölçüde doğru olduğu kabul edilebilir ve benzeri çalışmaların kıyaslanması açısından iyi bir referans olarak düşünülebilir.

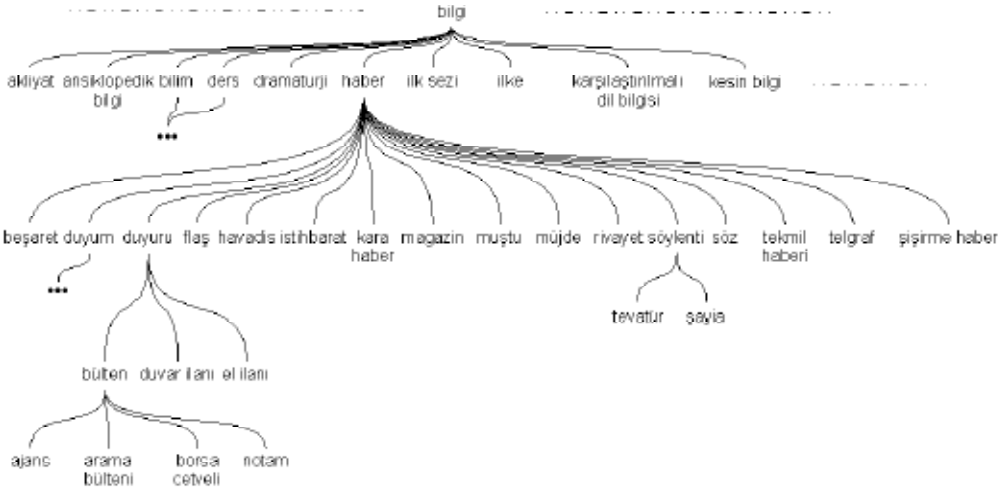


Şekil-2 Dolaşık Hiyerarşi



Şekil-3 Türkçe WordNet

Türkçe WordNet veri tabanının üst-kavram/alt-kavram ilişkilerine ait bölümünden alınan bir örnek Şekil-3'te gösterilmiştir. Sözcüklerin yanında görülen rakamlar, sözcüğün sözlükteki kaçıncı anlamı olduğunu ifade etmektedir. Buna göre, Türkçe WordNet'te bir sözcük tek bir düğüm ile gösterilmemekte, sözcüğün anlam sa-



Şekil-4 Hiyerarşiden örnek bir bölüm

yısı kadar düğüm yer almaktadır. Bu makalede anlatılan yöntem sonucunda oluşan hiyerarşik yapıdaki ilgili kısım da Şekil-4'te verilmiştir.

İki yapı karşılaştırıldığında dikkati çeken ilk nokta, Türkçe WordNet'in daha az ve özlü bilgi içerdiğidir. Tam olarak üst-kavram/alt-kavram ilişkisi içerisinde görünmeyen kavramlara yer verilmemiştir. Diğer yapıda ise bir ölçüde bu tür bir ilişki içine sokulabilecek bütün kavramlar birbirlerine bağlanmıştır; bu durum kullanılan sözlüğün özelliklerinden kaynaklanmaktadır. Örneğin, dersin bir anlamda bilgi sağlayan bir kavram olduğu, duyurunun haber iletimi amacıyla kullanıldığı ve bültenin bir çeşit duyuru aracı olduğu çıkarımlarını yapmak mümkündür. Türkçe WordNet'te ise "haber" sözcüğü yaprak düğümdür (*leaf node*) ve alt-kavramları bulunmamaktadır. Hiyerarşik yapıların diğer kısımlarında da benzer bir durum söz konusudur. Buna göre, bu makalede bahsedilen yöntem sonucu elde edilen hiyerarşik yapının daha kapsayıcı olduğu ve kavramlar arasındaki anlam bilimsel bağları bulma gereksinimi olan doğal dil çalışmalarında çok daha fazla ilişkinin ortaya çıkarılmasına yarayacağı düşünülebilir.

Bununla ilintili olarak değinilmesi gereken diğer bir nokta, Türkçe WordNet'teki hiyerarşik yapının üst-kavram/alt-kavram ilişkileri açısından hemen hemen hatasız oluşu, oysa diğer yapıda çeşitli hataların bulunmasıdır. Daha önce açıklandığı üzere, Türkçe WordNet'te yer alan kavramların anlam muğlaklıklarının elle giderilmiş olması çalışmayı oldukça zahmetli bir hale getirmektedir. Bu makalede bahsedilen çalışma ise tamamen otomatik olarak işlemektedir. Geliştirilmiş olan algoritmaya sözcük anlamlarındaki muğlaklıkların giderilmesi amacıyla uygun bir modül eklenmesi durumunda, hata oranının önemli ölçüde düşeceği beklenebilir. Bu konu, şu anda üzerinde çalışmakta olduğumuz bir konudur.

4. Sonuç

Bu makalede, Türkçe diline yönelik olarak, Türk Dil Kurumu'nun (TDK) elektronik sözlüğü kullanılarak, sözcükler arasındaki üst-kavram, alt-kavram ve eş anlamlılık ilişkilerinin tamamen otomatik olarak tespit edilmesi amacıyla geliştirilmiş olan bir yöntem ve bu yöntemin uygulanması anlatılmıştır.

Sözlükteki yapılar ve kavramlar ayrıntılı olarak incelenerek, kullanılan sözlüğe özgü özellikler tespit edilmiştir. Bu özellikler, 11 adet kural yardımıyla ve birtakım biçim bilimsel kriterler

kullanılarak temsil edilmiştir. Hiyerarşinin ana kavramları olarak İngilizce WordNet'ten alınan 21 grup kullanılmıştır. Oluşturulan kural tabanlı yöntem uygulanarak, Türkçe için bütün sözlükteki anlam bilimsel ilişkiler çıkarılmıştır.

Bu çalışmanın sonuçları, sözlükteki tanımlardan önemli miktarda anlamsal bilginin çıkarılabileceğini göstermiştir. Çalışmanın başlıca eksikliği, daha önce de bahsedildiği üzere, elde edilen kavramlar arasındaki anlam muğlaklığının giderilmemiş olmasıdır. Bu konu, tarafımızdan şu anda üzerinde çalışılmakta olan bir konudur. Bununla ilgili olarak literatürde oldukça detaylı ve başarılı araştırmalar mevcuttur. Bu metotların elektronik sözlüğün özellikleri de dikkate alınarak uyarlanması üzerinde çalışılmaktadır. Sözcük anlamlarının belirlenmesi durumunda, çıkarılan anlamsal ilişkilerin oldukça yüksek bir doğruluk oranına erişeceği düşünülmektedir.

Gelecekte hedeflenen bir diğer gelişme, sözlük tanımlarını inceleyerek ve bazı istatistiksel gözlemler yaparak kullanılan kuralları öğrenebilecek bir algoritmanın oluşturulmasıdır. Bu makalenin kapsamında yer alan anlamsal ilişkilerin yanı sıra, diğer türdeki ilişkilerin (parça-bütün, karşıtlık, vb.) çıkarılması da başka bir araştırma konusudur. Son olarak, farklı kök düğüm kümeleri kullanarak, oluşacak hiyerarşilerin kapsam ve gösterim kaliteleri açısından karşılaştırılması da gelecekteki konular arasında düşünülebilir.

Kaynakça

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

[2] Türk Dil Kurumu Ana Sayfası, http://www.tdk.gov.tr/, *Türk Dil Kurumu*.

[3] Chodorow, M. S., Byrd, R. J., "Extracting semantic hierarchies from a large on-line dictionary", Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, 1985, University of Chicago, Chicago, Illinois, 299-304.

[4] Ide, N. and Véronis, J., "Refining taxonomies extracted from machine-readable dictionaries". In Hockey, S., Ide, N. *Research in Humanities Computing 2*, Oxford University Press, 1993.

[5] Ide, N. and Veronis, J., "Machine Readable Dictionaries: What have we learned, Where do we go?", in: Calzolari and C. Guo (eds) Proceedings of the COLING94 International Workshop on Directions of Lexical Research, 1994, Beijing, 137-146.

[6] Alshawi, ve Hiyan, "Processing dictionary definitions with phrasal pattern hierarchies", *American Journal of Computational Linguistics*, Vol. 13(3), 1987, 195-202.

[7] Zemberek Projesi Geliştirme Sayfaları, https://zemberek.dev.java.net/, *Zemberek Projesi*.

[8] Miller, G. A., "Nouns in WordNet: A Lexical Inheritance System", 1993.

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