Türkçe için Bilgisayarla İşlenebilir Sözlük Kullanarak Kavramlar Arasındaki Anlamsal İlişkilerin Belirlenmesi

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

Özetçe

Bu makalede, 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.

Abstract

Extracting Semantic Relations Between Concepts Using a Machine-Readable Dictionary for Turkish

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.

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[1] 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

1 Bunlara örnek olarak, üst-kavramlılık (*hypernymy*), alt-kavramlılık (*hyponymy*), eş anlamlılık (*synonymy*), parça-bütün (*meronymy*), karşıtlık (*antonymy*), vb. ilişkiler verilebilir. Üst-kavram, bir sözcüğün içinde bulunduğu daha genel anlamı temsil eden sözcük olarak tanımlanır. Örneğin, "çiçek" sözcüğü, "gül" sözcüğünün üst-kavramıdır. Alt-kavram ise, üst-kavramın tersidir. Aynı örnekte, "gül" sözcüğü, "çiçek" sözcüğünün alt-kavramıdır.

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 [11]. İsimler, isim eş anlam kümeleri arasındaki üst-kavram (*hypernym*) ve alt-kavram (*hyponym*) ilişkileri kullanılarak hiyerarşik bir yapıya yerleştirilmiştir. Sıfatlar ise büyük oranda eş anlamlılık ve karşıtlık ilişkilerine göre düzenlenmiştir. İsimler ve sıfatlar için kullanılan ilişkiler fiilleri yeterli düzeyde anlamsal ilişkilere ayıramadığından dolayı, bunların uygulanıp uygulanamayacağı veya ne dereceye kadar uygulanabileceği bir tartışma konusu olmuştur. Bu ilişkilere ek olarak, sözlüksel gerektirirlik (*lexical entailment*) çeşitlerine bazı çözümler getiren birkaç ilişki de sunulmaktadır. 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 nedenle, bu veri tabanlarını otomatik olarak hazırlayabilecek algoritmalar üzerinde çalışmak gittikçe önemi artan bir araştırma alanı haline gelmektedir.

Bu makalede, 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 [19]. 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. [11]'de de tartışıldığı gibi, isimler ve fiiller dışındaki diğer kategoriler arasında üst-kavram ve eş anlamlılık ilişkilerinin tanımları belirsiz olduğundan dolayı, çalışmada sadece isimler göz önünde bulundurulmuştur.

Makalenin devamı şu şekilde düzenlenmiştir: 2. bölümde kullanılan yöntem ve algoritma detaylı olarak anlatılmaktadır. Bir sonraki bölümde, kullanılan yöntemin ve kuralların sözlüğe uygulanması sonucu oluşan hiyerarşi yapısı değerlendirilmekte, yöntemin çeşitli eksikliklerine değinilmekte ve Türkçe WordNet'in içerdiği hiyerarşik yapı ile karşılaştırılmaktadır. Bölüm 4'te, araştırmanın konusu ile ilgili literatürde yer alan çalışmalar özetlenmektedir. Son bölüm ise sonuç kısmına ve gelecekte yöntem üzerinde gerçekleştirilebilecek geliştirmelere ayrılmıştır.

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. Bu parçalama işleminin altında yatan fikir, kullanılan sözlüğün yapısına özgü özelliklerin incelenmesi sonucu ortaya çıkmıştır. Sözlükteki tanımlar, aşağıda düzenli gramer (*regular grammar*) biçiminde belirtilen genel örüntüyü (*pattern*) izlemektedir:

sözcük : (s* üst) (, s* üst)* (, eş)*.

Burada, *s* herhangi bir sözcüğü, *üst* sözcüğün olası bir üst-kavramını, *eş* ise sözcüğün olası bir eş

anlamlısını ifade etmektedir. Burada görüldüğü üzere, tanımlar sıklıkla sözcüğün (eğer varsa) eş anlamlılarıyla bitmekte, bunlardan önce ise virgülle ayrılmış bir dizi sözcük gelmektedir. Her dizinin sonundaki sözcük ise olası bir üst-kavramı göstermektedir. Bu yapının belirlenmesi ve tanımın ilk olarak bahsedilen şekilde parçalara bölünmesi, tanımın ilerideki aşamalarda işlenmesini oldukça kolaylaştırmıştır.

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. Bunun nedeni, bu noktada durulmadığında çok sayıda hatalı sözcüğün de üst-kavram olarak çıkarıldığının gözlenmiş olmasıdır. Bir sözcüğün eş anlamlılarının her zaman üst-kavramlarından sonra gelmelerinden dolayı, süreci bu noktada durdurmak eş anlamlıların çıkarılmasında sorun yaratmamaktadır. Ayrıca, eş anlamlı sözcükler tespit edilirken, bir eş anlamlı sözcük bulunduğu zaman kuralların işlemesi durdurulmamakta, böylece tanımın farklı yerlerinde birden fazla eş anlam ortaya çıkarılabilmektedir.

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; bunlar sırasıyla Kural 1 ve Kural 11'dir. Diğer dokuz kural üçüncü grubu oluşturmaktadır. Kurallar Tablo-1 ile Tablo-11 arasındaki tablolarda gösterilmiştir. Her kural için, kuralın uygulanmasının daha rahat anlaşılabilmesi amacıyla, örnek bir sözcük ve kısa bir açıklama verilmiştir.

Tablo-1 : Kural 1

Kural 1: İncelenen sözcük "bilimi" ile bitiyorsa, bu sözcüğün üst-kavramı "bilim"dir. *Örnek:* Sözcük: Kanser bilimi Sözlüksel kategori: İsim, tıp Tanım: Kanser hastalıklarını inceleyen tıp dalı, kanseroloji. Kural 1'e göre, "bilim" üst-kavram olarak belirlenir. Ayrıca, diğer kuralların uygulanmasıyla, "dal" diğer bir üst-kavram ve "kanseroloji" eş anlam olarak belirlenir.

Tablo-2 : Kural 2

Kural 2: İncelenen parça aşağıdaki üç grup sözcük öbeğinden herhangi birisiyle bitiyorsa, dahil olduğu gruba göre belirtilen işlemler uygulanır.

Grup 1:
"olanlardan her biri"
"olanlardan biri"
"olanlardan bazısı"
Sözcük öbeğinden önceki sözcük ÜSK'ya göre işlenip üst-kavram elde edilir.
Örnek:
Sözcük: Kan kardeşi
Sözlüksel kategori: isim
Tanım: Birbirlerinin kanını emerek veya yalayarak ant içmek yoluyla kardeş olanlardan her biri, ant kardeşi, kanka.
Bu örnekte üst-kavram "kardeş" olarak kaydedilir.

Grup 2:
"her biri"
"biri"
"bazısı"
Sözcük öbeğinden önceki sözcükten sırasıyla ismin ayrılma eki ("-dan" eki) ve çoğul eki ("-lar" eki) çıkartılarak üst-kavram elde edilir.
Örnek:
Sözcük: Alet
Sözlüksel kategori: İsim
Tanım: Bir makineyi oluşturan ve işlemesine yardım eden parçalardan her biri.
Sırasıyla ayrılma eki ve çoğul eki "parçalardan" sözcüğünden ayrılır, oluşan "parça" sözcüğü "alet"" sözcüğünün üst-kavramı olarak alınır.

Grup 3:
"S_1 BAĞLAÇ S_2 her biri"
"S_1 BAĞLAÇ S_2 biri"
"S_1 BAĞLAÇ S_2 bazısı"
(S_1 ve S_2 tekil sözcükleri, BAĞLAÇ "ve" veya "veya" sözcüğünü ifade etmektedir.)

S1 ve S2'ye ÜSK uygulanır. Böylelikle iki tane üst-kavram elde edilebilir.
Örnek:
Sözcük: Bakla
Sözlüksel kategori: isim
Tanım: Bir zinciri oluşturan halka veya parçalardan her biri.

Tablo-3 : Kural 3

Kural 3: İncelenen parça "(kimse)" veya "kimse" ile bitiyorsa, "kişi" sözcüğü üst-kavram olarak belirlenir.

Örnek:
Sözcük: Abacı
Sözlüksel kategori: İsim
Tanım: Aba yapan veya satan kimse.

Bu kural uygulanmasaydı, üst-kavram olarak "kimse" zamiri belirlenecekti. Fakat, "kişi" sözcüğünün üst-kavram olarak daha uygun olduğu düşünülebilir.

Tablo-4 : Kural 4

Kural 4: İncelenen parça "işi" sözcüğüyle bitiyorsa, sözcüğün üst-kavramı "iş" olarak belirlenir.

Örnek:
Sözcük: Açılma
Sözlüksel kategori: isim
Tanım: Açılmak işi.

Tablo-5 : Kural 5

Kural 5: İncelenen parça "tümü" sözcüğüyle bitiyorsa, "tümü" sözcüğünden önceki sözcükten sırasıyla sahiplik ve çoğul ekleri ("-ların" ek grubu) çıkarılır. Elde edilen sözcük bu sözcüğün üst-kavramı olarak belirlenir.

Örnek:
Sözcük: Banyo takımı
Sözlüksel kategori: İsim
Tanım: Yıkanmak ve kurulanmak için gerekli olan gereçlerin tümü.

Örneğin yukarıdaki tanımdan ilgili sözcüğün üst-kavramının "gereç" olduğu çıkarılır.

Tablo-6 : Kural 6

Kural 6: İncelenen parça "değil" sözcüğü ile bitiyorsa, incelenmekte olan sözcüğün işlenmesi sonlanır. Bu tür durumlarda, tanımın "değil" sözcüğünden önceki kısımlarında, üst-kavram ve eş anlam bulunmadığı görülmüştür.

Tablo-7 : Kural 7

Kural 7: İncelenen parça "hepsi" sözcüğü ile bitiyorsa, "grup" sözcüğü üst-kavram olarak belirlenir.

Örnek:
Sözcük: Gelin alayı
Sözlüksel kategori: İsim
Tanım: Gelini damat evine götürmek için gelenlerin hepsi.

Tanımdaki gizli üst-kavram "grup" bu kuralla çıkarılabilmektedir.

Tablo-8 : Kural 8

Kural 8: İncelenen parça "$SÖ_1$ ve $SÖ_2$" veya "$SÖ_1$ veya $SÖ_2$" ile bitiyorsa ($SÖ_1$ ve $SÖ_2$ bir ila üç sözcükten oluşan sözcük öbekleridir), bu iki sözcük öbeğinin son sözcüklerine ÜSK uygulanarak üst-kavram belirlenir.
Sözcük öbeklerinin sadece son sözcüklerine ÜSK'nın uygulanmasının nedeni, diğer sözcüklerin genellikle sıfat veya belirteç kökenli olmasıdır.

Örnek:
Sözcük: Açı ölçüm
Sözlüksel kategori: isim, geometri
Tanım: Açı ölçmede söz konusu olan yöntem ve teknik.

Bu örnekte "açı ölçüm" sözcük öbeğinin üst-kavramları "yöntem" ve "teknik" olarak belirlenir.

Tablo-9 : Kural 9

Kural 9: İncelenen parça "vb." ile bitiyorsa, üst-kavram "vb."'den önceki sözcüğe ÜSK uygulanarak bulunur. Eğer "vb." parçanın sonunda değil de başka bir yerinde geçiyorsa, geri kalan parçalar işlenmez ve sözcüğün işlenmesi sonlanır. Bu tür durumlarda, tanımın "vb."den önceki kısımlarında, üst-kavram ve eş anlam içerilmediği görülmüştür.

Örnek:
Sözcük: Acı
Sözlüksel kategori: İsim
Tanım: Ölüm, yangın, deprem vb. olayların yarattığı üzüntü, keder, elem.

Diğer kurallarla, "üzüntü" üst-kavram, "keder" ve "elem" eş anlamlar olarak belirlendikten sonra, parçanın içinde geçen "vb." sözcüğünden dolayı sözcüğün işlenmesi tamamlanır ve başka bir ilişki çıkarılmaz. Görüldüğü üzere, işlenmeyen kısımda yer alan sözcükler ("ölüm", "yangın", "deprem"), "vb."den sonra gelen kavramın ("olay") örnekleridir ve incelenen sözcüğün üst-kavramları ve eş anlamlıları değildir.

Tablo-10 : 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.

Tablo-11 : Kural 11

Kural 11: İncelenen sözcük "botanik" veya "zooloji" kategorilerine aitse, birinci parçanın son sözcüğünden ayrılma eki ("-dan" eki) ayrılır. Parçadaki diğer sözcüklerle ekin ayrılmasıyla oluşan sözcük arka arkaya getirilerek üst-kavram elde edilir. Eğer son parça parantez içinde bir sözcük ile bitiyorsa, bu sözcükten önceki sözcüğe ÜSK uygulanarak üst-kavram elde edilir.
Örnek:
Sözcük: Abdestbozan otu
Sözlüksel kategori: İsim, botanik
Tanım: Gülgillerden, siyah ve yeşil boya çıkarılan bir bitki (Poterium spinosum).
Ayrılma ekinin "gülgillerden" sözcüğünden çıkarılmasından sonra, ilk parçada başka sözcük olmadığından dolayı, sözcüğün üst-kavramı "gülgiller" olarak belirlenir. Bu tip sözcük veya sözcük öbeklerinde ilk üst-kavram bir bitkinin veya hayvanın familyasına karşılık gelmektedir.

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. Burada bahsedilen kriterler, sözlük tanımlarındaki üst-kavram yapılarının incelenmesi sonucu tespit edilmiştir. Tanımlardaki cümleler sözlüğü hazırlayan kişiler tarafından genellikle belirli kalıplara uyan bir dille yazıldıklarından, bazı istisnalar dışında belirlenen kriterlere uyduğu gözlenmiştir.

Tanımdan elde edilen sözcük, biçim bilimsel bir analizden geçirilmektedir. Bu çalışmada, biçim bilimsel analiz programı olarak Zemberek sistemi kullanılmıştır [15]. Zemberek, Türkçe'nin doğal dil işleme yöntemleri vasıtasıyla işlenmesi sırasında ortaya çıkan bilişsel problemlerin çözümünü kolaylaştırmayı amaçlayan bir program kütüphanesinden ve uygulama programlarından oluşmaktadır. Biçim bilimsel analiz programı ile sözcüğün kök hali ve aldığı ekler bulunmaktadır. Dikkat edilmesi gereken bir nokta, sözcüğün içinde bulunduğu bağlamdan (*context*) bağımsız olarak çalışan bütün biçim bilimsel analiz sistemlerinde olduğu gibi, Zemberek programının da hatalı ayrıştırmalar çıkarabilmesidir. Bu hataların önlenebilmesi ancak biçim bilim bazında muğlaklık giderici algoritmaların kullanılması ile mümkündür [20]; bu konu, bu çalışmanın kapsamı dışında tutulmuştur. Çalışmamızda Zemberek'in çıktı olarak verdiği sözcüğün olası ayrıştırmalarından sadece ilki kullanılmaktadır.

Biçim bilimsel analiz programından elde edilen ayrıştırma, Tablo-12, Tablo-13 ve Tablo-14'te gösterilen 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-12 : ÜSK 1. Grup

Analiz 1:
isim kökü + tamlama eki
Örnek:
(... fon) türü
tür + -ü
Üst-kavram: tür

Analiz 2:
isim kökü + üçüncü tekil şahıs iyelik eki
Örnek:
(... çetenin) başı
baş + -ı
Üst-kavram: baş

Analiz 3:
isim kökü
Örnek:
(...) yer
yer
Üst-kavram: yer

İşlem:
Eğer sözcüğün analizi yukarıdaki analizlerden birine uyuyorsa, sözcüğün üst-kavram olabileceği anlaşılır. Sözcüğün üst-kavram olarak kullanılması gereken biçimi olarak, yukarıdaki analizlerde isim kökü olarak ayrıştırılmış sözcük kaydedilir.

Tablo-13 : 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.

Tablo-14 : ÜSK 3. Grup

Analiz 1:
fiil kökü + (bir veya daha fazla ek) + isimfiil (eylemlik) eki + üçüncü şahıs kişi iyelik eki
Örnek:
(...) konuşması
konuş + -ma + -sı
Üst-kavram: konuşma

İşlem:
Eğer sözcük yukarıdaki analize uyuyorsa, üst-kavram biçimi, iyelik ekini sözcükten çıkardıktan sonra kalan sözcük olarak belirlenir.

2.1.3 Eş Anlamlı Sözcüklerin Çıkarılması

Bir sözcüğün tanımı içinde yer alan eş anlamlı sözcükler, 2.1.1. bölümde bahsedilen ve Tablo-10'da 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. 3.3. bölümde daha detaylı olarak bahsedileceği üzere, geliştirilen yöntem sözcük temelli çalıştığından ve sözcüklerin farklı anlamları arasında bir ayrım yapmadığından dolayı, 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.

Tablo-15 : Kök Düğümler

İngilizce WordNet	*Önerilen Yöntem*
{act, action, activity}	{hareket}
{natural object}	*
{animal, fauna}	{hayvan}
{natural phenomenon}	*
{artifact}	*
{person, human being}	{kişi, insan}
{attribute, property}	{özellik}
{plant, flora}	{bitki}
{body, corpus}	{gövde, vücut}
{possession}	{sahip}
{cognition, knowledge}	{kavrama, anlama, bilgi}
{process}	{süreç, işlem}
{communication}	*
{quantity, amount}	{miktar}
{event, happening}	{olay, iş}
{relation}	{ilişki}
{feeling, emotion}	{duygu, his}
{shape}	{şekil}
{food}	{yiyecek, yemek}
{state, condition}	{durum}
{group, collection}	{grup, topluluk, küme}
{substance}	{madde}
{location, place}	{yer}
{time}	{zaman, vakit}
{motive}	{amaç}

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 [10]; bu kavramlar ve bu çalışmada kullanılan karşılıkları Tablo-15'te verilmiştir. Tabloda görüldüğü üzere, iki dil arasındaki farklılıklardan dolayı, İngilizce bazı kavramların Türkçe sözlükte karşılıkları bulunmamaktadır; bu kavramlar tabloda * ile işaretlenmiştir. Buna göre, Türkçe'de oluşturulan hiyerarşik ilişkiler yapısında en üstte 21 düğüm yer almaktadır. Katalanca ve İspanyolca WordNet veri tabanlarının hazırlanmasında da bu yaklaşımdan yola çıkılmıştır [5].

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

Bu bölümde, makalenin önceki kısımlarında detayları verilen yöntemin uygulanması sonucu elde edilen hiyerarşi yapısı gösterilecek ve bu yapıdaki eksikliklere değinilerek yöntemin ne şekilde geliştirilmesi gerektiği açıklanacaktır.

3.1 Hiyerarşik Yapıyla İlgili İstatistikler

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. 4. Bölümdeki literatür özetinde de görüleceği üzere sözlüklerdeki eksik tanımlara ve tutarsızlıklara bağlı olarak, çıkarılan üst-kavramların birebir olarak hiyerarşide temsil edilmesi mümkün olmamaktadır. Bu nedenle 2.1. bölümde çıkarılan üst-kavramların tümünün hiyerarşik yapıda bulunmadığı gözlenmiştir.

Hiyerarşi 72 seviyeden oluşmaktadır. Seviye sayısının fazla oluşunun nedeni, 3.3. kısımda da bahsedileceği üzere, sistemde sözcük anlamlarının muğlaklığını gideren (*word sense disambiguation*) bir modülün bulunmamasıdır. Bu nedenle, bir sözcüğün sözlük tanımından o sözcüğün üst-kavramı bulunduğunda, hiyerarşik yapıda sözcük üst-kavramın bütün sözlük anlamlarına bağlanmaktadır. 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. Bu durumun önlenebilmesi için sözlük hazırlayan uzmanların, sözlüklerin bilgisayarla işlenebileceği olasılığını göz önünde tutarak belirli standartlara uymaları gerekmektedir.

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 yapısal bir özelliğinden kaynaklanmaktadır: hemen her fiil için, fiilden oluşan ve "işi" sözcüğü ile tanımlanan isim kökenli bir sözcük de sözlükte yer almaktadır (örneğin, "okuma: okumak işi"). Bu tür üst-kavramlar algoritmada Grup 1 ÜSK ile çıkarılmaktadır.

3.2 Sözcük Temelli Yaklaşımın Sınırlamaları

2.1.1. bölümde anlatıldığı üzere, üst-kavram ilişkilerini çıkaran algoritma, tanımların içindeki sözcükleri ve sözcük gruplarını dikkate almakta, cümlelerin söz dizimsel yapılarını göz ardı etmektedir. Bu yaklaşımın bazı durumlarda hatalı çıkarımlara yol açacağı açıktır. Aşağıda örnek bir sözlük maddesi ve tanımı verilmiştir:

> Kamu güvenliği: Bir devlette zabıta hizmetleriyle halka sağlanan can ve mal güvenliği

Bu örnek için algoritma Tablo-8'de verilen 8. kuralı uygular ve "kamu güvenliği" maddesi için iki üst-kavram belirler: "güvenlik" ve "can". Bunlardan ikincisinin hatalı bir çıkarım olduğu ve "kamu güvenliği" maddesi ile üst-kavram ilişkisi bulunmadığı görülmektedir. Hatanın sebebi, "can ve mal güvenliği" ifadesinin, çeşitli sözcüklerden oluşan bir isim öbeği olarak ele alınmayıp, birbirlerinden bağımsız sözcükler olarak düşünülmesidir. Bu tür hataların önlenebilmesi, algoritma tarafından, sözcüklerin belirli kurallara göre birleşmesi sonucu oluşan sözcük öbeklerinin de algılanması ile mümkündür. Diğer bir deyişle, sözcük temelli analize ek olarak, bir ölçüde söz dizimsel analiz de devreye sokulmalıdır.

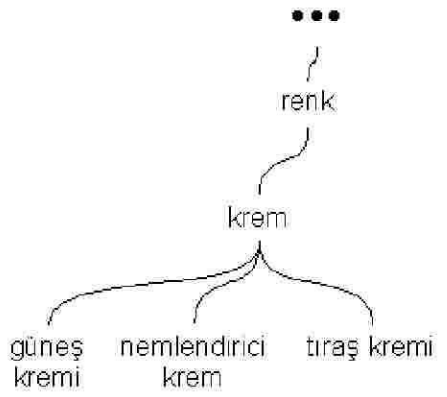
3.3 Sözcüklerdeki Anlam Muğlaklıkları

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. Bu tür bir bilgi, bilgisayarla işlenmeye uygun olarak hazırlanan elektronik sözlüklerin hemen hemen hiçbirinde yer almamaktadır ve sözlüğü analiz eden algoritma tarafından çıkarılmalıdır. Bu problem, doğal dil işleme çalışmalarının ortak bir problemidir ve literatürde sözcük anlamındaki muğlaklığın giderilmesi (*word sense disambiguation*) problemi olarak anılmaktadır.

Bir sözcüğün analizi sonucu bulunan üst-kavramın sözlükte birden çok anlamının olması olağan bir durumdur. 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.

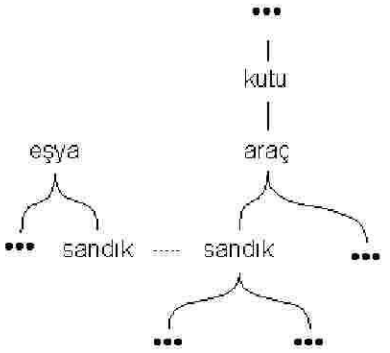


Şekil-1 'Güneş kremi' ve 'krem'in hiyerarşideki yerleri

3.4 Hiyerarşinin Ağaç Yapısının Korunması

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. Bunun sebebi, yapıdaki bir düğümün birden çok üst düğümünün (*parent node*) ve birden çok alt düğümünün (*child node*) bulunabilmesi, dolayısıyla yapıda çevrimlerin (*cycle*) oluşabilmesidir. 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.

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.



Şekil-2 Dolaşık Hiyerarşi

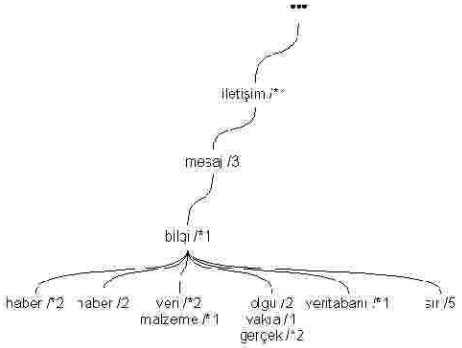
Çizge yerine bahsedilen yapıda bir ağacın yaratılmış olması, hiyerarşik yapı üzerinde arama yapacak olan algoritmalar açısından herhangi bir zorluk teşkil etmeyecektir; basit bir indeksleme mekanizması kurularak aynı sözcüğe ait düğümlerin birinden diğerlerine ulaşmak mümkün olacaktır. Bu durum literatürde "dolaşık hiyerarşi" (*tangled hierarchy*) olarak bilinmektedir [13].

3.5 Türkçe WordNet ile Karşılaştırma

Makalenin 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 [12]. Türkçe WordNet titiz bir çalışma sonucunda yarı otomatik olarak oluşturulmuştur. 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.

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. Örneğin, "haber" sözcüğünün 2. anlamı, "bilgi"

sözcüğünün 1. anlamının alt-kavramıdır. Buna göre, Türkçe WordNet'te bir sözcük tek bir düğüm ile gösterilmemekte, sözcüğün anlam sayısı kadar düğüm yer almaktadır. Bu makalede anlatılan yöntem sonucunda oluşan hiyerarşik yapıdaki ilgili kısım da Şekil-4'te verilmiştir.



Şekil-3 Türkçe WordNet

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.

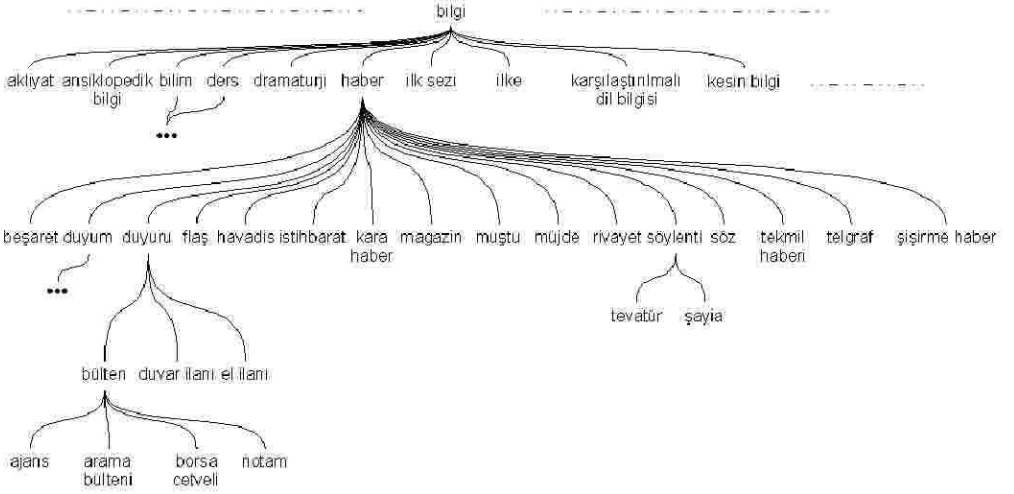
Bununla ilintili olarak değinilmesi gereken diğer bir nokta, Türkçe WordNet'teki hiyerarşik yapının üst-kavram/alt-kavram ilişkileri açısından hemen hemen hatasız oluşu, oysa diğer yapıda çeşitli hataların bulunmasıdır. Daha önce açıklandığı üzere, iki yapı arasındaki fark, Türkçe WordNet'te yer alan kavramların anlam muğlaklıklarının elle giderilmiş olması ve ilişkilerin insanlar tarafından kontrol edilerek hiyerarşiye yansıtılmış olmasıdır. Bu yaklaşım çıktının doğruluk oranını arttırmakta, fakat insan faktörü devreye girdiğinden dolayı çalışmayı oldukça zahmetli bir hale getirmektedir. Bu durum, derlem (*corpus*) işleme amaçlı yürütülen bütün araştırmaların ortak sorunudur. 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. Literatür Özeti

Bu bölümde, bilgisayarla işlenmeye uygun olarak hazırlanan elektronik sözlükler kullanılarak yapılan benzer çalışmalara değinilecektir. Dikkat edilmesi gereken bir nokta, bu çalışmaların hepsinin İngilizce için olduğudur. Chodorow ve Bryd tarafından yapılan çalışmada, isim kökenli ve fiil kökenli sözcükler dikkate alınmış, birincisi için tanımın içindeki isim öbeği, ikincisi için ise tanımın içindeki fiil öbeği çıkarılmıştır [1]. Bu çıkarım birtakım varsayımlara dayandırılmıştır. Örneğin, belirli sözcüklerle ("a", "an", "the", "its", "two", "three", vs.) başlayan ve belirli sözcüklerle ("that", "who" gibi ilgi zamirleri, edatlar, vs.) biten sözcük dizileri isim öbeği olarak kabul edilmiştir. Bu öbeklerdeki ana sözcüğün (*head word*) üst-kavram olduğu varsayılarak, bu sözcüğün tespit edilmesine çalışılmıştır. Öbekteki ilk sözcük genellikle ana sözcük olmaktadır. Örneğin, fiil öbeği için, öbekteki ilk fiilin veya "to" sözcüğünden sonra bir bağlaç geliyorsa bu bağlacı izleyen fiilin üst-kavram olduğu kabul edilmiştir.

Üst-kavramlar tespit edildikten sonra, yaratılmış olan üst-kavram indeks yapısı bizim çalışmamızdakine benzer bir yöntem yardımıyla analiz edilerek hiyerarşi oluşturulmuştur. Fakat iki çalışma arasındaki farklardan birisi, bahsedilen makalede kullanılan yöntemin yarı otomatik olmasıdır. Sözcüklerin anlam muğlaklıkları giderildikten sonra kullanıcıya sözcüğün hiyerarşiye eklenip eklenmeyeceği sorulmuştur. Bu şekilde hiyerarşik yapının oluşması kontrollü bir şekilde sağlanmıştır.

Başka bir araştırmada, üst-kavram ilişkilerinin çıkarılmasına yönelik çalışmalarda sadece tek bir



Şekil-4 Hiyerarşiden örnek bir bölüm

sözlükten yararlanmanın yetersiz olduğu öne sürülmüştür [6]. Bu tür çalışmalarda karşılaşılan güçlüklerden bahsedilmiş ve oluşan hatalar örneklenmiştir. Birden fazla sözlük kullanıldığında ise bu hataların önemli ölçüde azaldığı ifade edilmiştir. Bu öngörünün dayandığı temel fikir, her bir sözlüğün kendine özgü yapısal özelliklerinin bulunduğu, bu özelliklerin kimi durumlarda hatalı çıkarımlara yol açtığı, fakat başka sözlükler de kullanıldığında bu hataların ortadan kaldırılabildiğidir. Bu amaca yönelik başka bir benzer çalışma [14]'te verilmiştir.

Bir maddenin sözlükteki tanımından yola çıkarak o maddenin üst-kavramlarını elde etmek genellikle zor bir işlemdir. Bunun en önemli sebeplerinden birisi, sözlüklerin, üst-kavramların çıkarılmasına yardımcı olacak bir biçimde hazırlanmamış olmasıdır. Dolayısıyla, ortaya çıkan hiyerarşiler eksik ve hatalı olabilmektedir. Bunu önlemek amacıyla, bilgisayarlar tarafından işlenebilen sözlükler için ortak bir biçem önerilmiştir [3]. Bu öneri, sözlük yazarlarının tanımlarda anlamsal ilişkileri açık olarak belirtmelerini istemeye kadar genişletilmiştir.

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 [6,7]. 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 [4]'te verilmiştir. Genel örüntüler hiyerarşinin üst kısımlarında yer almakta, alt kısımlara gidildikçe örüntüler ve dolayısıyla temsil ettikleri söz dizimsel öbekler daha özgül hale gelmektedir. Bu özellik, sistemin dayanıklılığını (*robustness*) arttırmaktadır; bir tanıma karşılık gelen örüntü önce daha özgül örüntüler içerisinde aranmakta, bulunamadığı durumlarda daha genel örüntülerle eşleştirmek mümkün olmaktadır.

Diğer bir çalışmada, genel amaçlı bir ayrıştırma modülü (*parser*) kullanılarak sözcük tanımları söz dizimsel öğelerine ayrılmıştır [2]. Açığa çıkan öğeler arasında çeşitli tipteki ilişkilerin var olup olmadığına bakılmıştır. Bu çalışmada da birden fazla sözlük kullanılmış ve her sözlüğe uygun ayrıştırma ağaçları çıkarılarak birleştirilmiştir.

Buna benzer başka bir çalışmada da cümlelerin sözdizimsel ayrışımlarından bağımlılık yolu (*dependency path*) özelliklerini tespit eden ve bunları üst-kavramları çıkarmakta kullanmak üzere öğrenen bir yöntem üzerine çalışılmıştır [16].

Diğer çalışmalara göre daha basit bir yaklaşım öneren bir araştırmada, isim, fiil ve sıfatlar için önceden belirli kalıplar belirlenmiş ve sözlük tanımlarının bu kalıplara uygunluğu incelenmiştir [9]. Örneğin, isim kökenli bir sözcüğün tanımı "any" sözcüğü ile başlıyorsa, bu sözcüğü izleyen ismin veya isim öbeğinin sözcüğün üst-kavramı olduğu varsayılmıştır. Bu çalışmada, üst-kavram ilişkisine ek olarak, başka ilişkiler ve birtakım anlam bilimsel bilgiler de (parça-bütün, üyelik ilişkileri, fiillerin alt kategorilerine ayrılması, sıfatlar, vs.) çıkarılmaya çalışılmıştır.

Buraya kadar özetlenen çalışmaların hepsinde metin belli bir düzenli ifadeye göre taranıp metine içkin olan üst-kavram ilişkileri çıkarılmaya çalışılmıştır. Bu yaklaşımın yanında derlem içinde sözcüklerin farklı bağlamlarda yan yana gelme oranlarının üst-kavram ilişkisi açısından anlamı olduğu varsayımından yola çıkan istatistiksel yaklaşımdan da söz etmek gerekir.

Bu yaklaşımın önemli çalışmalarından biri olan [17]'da, bir derlemdeki sözcükleri dağılımsal benzerliklerine göre kümelendiren (*cluster*) bir algoritma tanımlanmıştır. Bu çalışmanın eksik yanı bu kümeleri adlandıramamasıdır. Bu kümelerin adlandırılması yolunda yapılan bir çalışmada bu adları kullanarak üst-kavram ilişkilerinin çıkarılabildiği gösterilmiştir [18].

5. 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., 1993. Introduction to WordNet: An On-line Lexical Database.

[2] http://www.tdk.gov.tr/

[3] https://zemberek.dev.java.net/

[4] Sak, H., Güngör, T. and Saraçlar, M., 2007. Morphological Disambiguation of Turkish Text with Perceptron Algorithm, Lecture Notes in Computer Science, Springer-Verlag, (in press).

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

[6] Farreres, X., Rigau, G. and Rodriguez, H., 1998. Using WordNet for Building WordNets, Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada.

[7] Amsler, R. A., 1980. The Structure of the Merriam-Webster Pocket Dictionary, Doctoral Dissertation, TR164, University of Texas, Austin.

[8] Bilgin, O., Çetinoğlu, Ö. and Oflazer, K., 2004. Building a WordNet for Turkish, Romanian Journal of Information Science and Technology, Volume 7, Numbers 1-2, 163-172.
[9] Chodorow, M. S. and Byrd, R. J., 1985. Extracting semantic hierarchies from a large on-line dictionary, Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, University of Chicago, Chicago, Illinois, 299-304.
[10] Ide, N. and Véronis, J., 1993. Refining taxonomies extracted from machine-readable dictionaries. In Hockey, S., Ide, N. Research in Humanities Computing 2, Oxford University Press.
[11] Sanfilippo, A. and Poznanski, V., 1992. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In Proc. of the 3rd Conference on Applied Natural Language Processing (ANLP), 80-87, Trento, Italy.
[12] Calzolari, Nicoletta, 1984. Detecting patterns in a lexical data base, Proceedings of the 10th International Conference on Computational Linguistics, COLING'84, Stanford University, California, 170-173.
[13] Ide, N. and Veronis, J., 1994. 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, Beijing, 137-146.
[14] Alshawi, Hiyan, 1987. Processing dictionary definitions with phrasal pattern hierarchies, American Journal of Computational Linguistics, Vol. 13(3), 195-202.
[15] Dolan, W., Vanderwende, L., and Richardson, S., 1993. Automatically deriving structured knowledge bases from on-line dictionaries, Proceedings of the First Conference of the Pacific Association for Computational Linguistics, Vancouver, Canada, 5-14.
[16] Snow, R., Jurafsky, D. and Ng, A. Y., 2004. Learning Syntactic Patterns for Automatic Hypernym Discovery, Advances in Neural Information Processing Systems, 17.
[17] Markowitz, J., Ahlswede, T. and Evens, M., 1986. Semantically significant patterns in dictionary definitions, Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, New York, 112—119.
[18] Pantel P. and Lin D., 2002. Discovering word senses from text, In Proceedings of ACM SIGKDD02.
[19] Pantel, P., Ravichandran, D., 2004. Automatically Labeling Semantic Classes. Proc. of NAACL.