Hypernym Extraction From Wikipedia and Wiktionary
Vikipedi ve Vikisözlük'ten Hypernym Çıkarma

Emre ŞAŞMAZ, Razieh EHSANİ, Olcay Taner YILDIZ
Bilgisayar Mühendisliği Bölümü
Işık Üniversitesi, İstanbul, Türkiye
emre.sasmaz@isik.edu.tr
{razieh.ehsani, olcaytaner}@isikun.edu.tr

***Abstract*—Doğal dil işleme alanında kullanılan önemli yapılardan bir tanesi WordNet gibi büyük ölçekli sözlüklerdir. WordNet; eşanlamlı, zıt anlamlı gibi anlamsal ilişkileri de içeren kapsamlı bir sözlüktür. Bu bildiride, WordNet'in önemli bir parçası olan Hypernym-Hyponym ilişkisini çıkarmaya çalıştık. Bu amaca ulaşmak için, Vikipedi, Türkçe Sözlük ve Vikisözlük kaynaklarını kullandık. Sonlu Durum Makinelerinden ürettiğimiz kurallarla Hypernym-Hyponym ilişkilerini çıkardık.**

***Abstract*—Large scale dictionaries such as wordnets are one of the important structures used in natural language processing. Wordnet is a comprehensive dictionary containing semantic relations including synonyms, antonyms, etc. In this paper, we try to extract Hypernym-Hyponym relations which are one of the main parts of WordNet. For this aim, we used Wikipedia, Contemporary Dictionary of Turkish and Wiktionary as corpora. We used Finite State Machines for developing several rules for Hypernym-Hyponym extraction.**

***Anahtar Kelimeler*—Doğal Dil İşleme, Anlambilim, Anlamsal İlişki, Hypernym-Hyponym, WordNet**

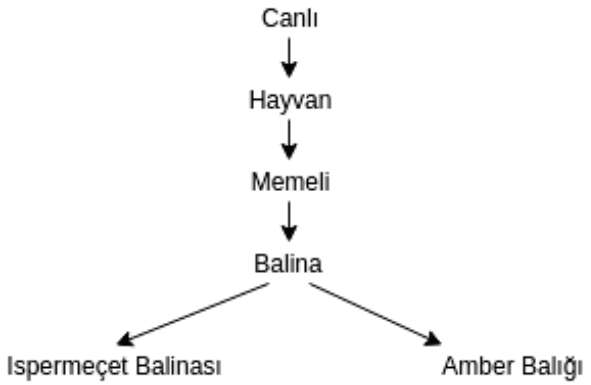
***Keywords*—Natural Language Processing, Semantics, Semantic relation, Hypernym-Hyponym, WordNet**

I. Giriş

İnsan dilini bilgisayarla anlama ve analiz etme Doğal Dil İşleme'nin (DDİ) esas işidir. DDİ'de dilin biçimbilimsel ve gramatik analizi dışında, anlamsal olarak analiz edilmesine anlambilim denir. Anlambilim dalı altında kelimelerin arasındaki bazı anlamsal ilişkileri çıkarmak günümüzün bu alandaki yaygın çalışmalarındandır. Bu anlamsal ilişkileri içeren kaynaklardan en önemlisi WordNet'tir. İngilizce için en kapsamlı WordNet, Princeton WordNet'idir [2]. WordNet eşanlamlı, zıt anlamlı gibi anlamsal ilişkileri içeren kapsamlı bir sözlüktür. Bu yaygın ilişkiler dışında Hypernym ve Hyponym gibi ilişkiler de kapsamlı WordNet'lerde işaretlenmiştir.

Hypernym-Hyponym ilişkisi, iki kelimenin arasında tanımlanan bir anlamsal ilişkidir. Bir kelime diğer kelimenin gerçek hayattaki bir türü veya çeşidi olursa o kelime diğer kelimenin hyponym'i olur. Örnek verecek olursak *amber balığı*, *balina* kelimesinin bir türü; *balina*, *memeli* kelimesinin bir türü; *memeli*, *hayvan* kelimesinin bir türü; *hayvan* ise *canlı* kelimesinin bir türü; veya *kırmızı*, *renk* kelimesinin bir çeşididir (Şekil 1).

978-1-5090-6494-6/17/$31.00 ©~2017 IEEE



Şekil 1: Hypernym ve Hyponym örnekleri

Bu örneklerde *amber balığı*, *balina*'nın; *balina*, *memeli*'nin; *memeli*, *hayvan*'ın; *hayvan*, *canlı*'nın hyponymi ve *renk*, *kırmızı*'nın hypernymidir.

Türkçe için şu ana kadar yapılmış en kapsamlı WordNet çalışması BalkaNet'tir. BalkaNet Türkçe ve Balkan dilleri arasındaki ortak 20000 civarı kelimeyi içermektedir. Daha önce BalkaNet projesinin bir parçası olarak toplam 889 hypernym-hyponym ilişkisi çıkarılmıştır. BalkaNet projesi kapsamında, sözlükteki kelime tanımları üzerine uygulanan kurallar ile bu ilişkiler çıkarılmıştır [1]. Bu kurallar "bir tür", "bir çeşit" gibi ipuçlarını kullanmaktadır. Yıldırım ve Yıldız [4] ise hem kural bazlı hem de istatiksel çalışma ile hypernym-hyponym ilişkilerini büyük bir derlemden çıkarmışlardır.

Bu çalışma, Türkçe için önemli olan Vikipedi ve Vikisözlük verisini kullanarak otomatik hypernym-hyponym ilişkilerini sonlu durum makinelerini kullanarak bulmaktadır.

II. Metodoloji

A. Türkçe Sözlük

Elimizdeki büyük bir veriden hypernym-hyponym ilişkisini çıkarmak için uygun veri bulmak önemlidir. Bu çalışmada ilk başta Milliyet haber derlemini [3] kullandık. Bu veri üstünde "bir tür", "bir çeşit", "gibi", "ve benzeri", "ve diğer" içeren kalıpları sonlu durum makinalar kullanarak çıkarmaya çalıştık. Aldığımız sonuçlara göre, Milliyet derlemi sadece haber

yazılarından oluştuğu ve herhangi bir kelime tanımı içermediği için, bu kurallarla herhangi bir gelişme kaydedilemedi. Bu deneyimin ardından, daha çok tanım ve ansiklopedi tarzında olan verilerin daha çok bilgi içereceğine inanarak, Türk Dil Kurumunun güncel Türkçe sözlüğünü (TS) kullanmaya karar verdik.

Türkçe sözlüğün ilk baskısı 1945 yılında yayınlanmıştır. O zamandan beri, birçok kez gözden geçirilmiş ve güncellenmiştir. Şu anda kullanmakta olduğumuz sürümü ise 2011 yılına aittir. Bu sürümde 65000 özgün kelime tanımları ile birlikte mevcuttur.

Türkçe sözlükteki kelime tanımlarına bakılınca hypernym-hyponym ilişkileri bazı ipuçları ile belirlenebilir. Örneğin, Türkçe sözlük'te *orkide* kelimesi tanımında "Salepgillerden" kelimesi geçmektedir. Bu gözlemlere dayanarak, hypernym-hyponym ilişkisini çıkarmak için mevcut kalıplardan kurallar çıkartılabilir. Bu kalıpların en sık görülenleri ise, "-gillerden" kalıbı, "çeşidi", "türü", "giller" gibi kalıplarıdır. Tablo I bu kalıpları ve bu kalıpları kullanarak bulduğunuz hypernym-hyponym ilişkisi olan kelime sayılarını göstermektedir.

TABLO I: Türkçe sözlük kelime tanımları üzerinde kullanılan kalıplar

Kalıp	İlişki Sayısı	Açıklama
-lerden	949	Zambakgillerden bir bitki
-giller	916	Tarla kuşugillerden bir kuş
-türü	231	Uzun taneli bir üzüm türü
-gillerden	54	Patlıcangillerden bir biber türü
-çeşidi	43	Çok tatlı bir kayısı çeşidi

B. Vikipedi

Bir sonraki adım olarak, Türkçe'nin başka bir kaynağı olan Vikipedi online ansiklopedisini kullandık. Öncelikle, Türkçe sözlük'te 65000 özgün kelime için Vikipedi'deki kelimelere baktık. Bu kelimelerden Vikipedi'de sadece 9624 tanım bulunmaktadır.

Vikipedi verisinin tamamını almak yerine, sadece genel tanımı içeren kısmı almak için ön eleme kuralını kullandık. Vikipedi ön eleme kuralı, Vikipedi'de kelimenin ilk cümle başında gelip ve cümle sonunun "dir", "dır", "dür", "dur" ile bitmesine bakmaktadır. Bu kurala uyan girdiler ilk olarak bizim büyük derlemimizi oluşturmaktadır. Böylelikle elimizde elenmiş olarak 6100 kelime tanımı kalmaktadır. Bu tanım cümleleri ve hypernym-hyponym ilişkilerini gözlemleyerek, çeşitli kalıplar çıkartılıp kurallar oluşturulmuştur. Bu kalıplardan bazıları aşağıda listelenmiştir. Burada kalıptaki SUP hypernym'e, SUB ise hyponym'e işaret etmektedir:

1) SUB ... SUP verilen genel (addır, isimdir).
   Rupi, bazı Asya ülkelerinde kullanılan paralara verilen genel addır.
   SUP: *para*, SUB: *rupi*
2) SUB ... bir SUPDIr.
   Bezelye, baklagiller familyasından tırmanıcı bir bitkidir.
   SUP: *bitki*, SUB: *bezelye*
3) SUB ... bir SUP çeşididir.
   Kadayıf, isim benzerliğine karşın tel kadayıftan çok farklı birer tatlı çeşididir.
   SUP: *tatlı*, SUB: *kadayıf*
4) SUB ... SUP türüdür.
   Kızılçam, kurak koşullara son derece dayanıklı, çok farklı toprak koşullarında başarıyla yetişen ve yetiştirilen, Türkiye'nin de en hızlı gelişen ağaç türüdür.
   SUP: *ağaç*, SUB: *kızılçam*
5) SUB ... SUPlerden birisidir.
   Frenk maydanozu, genelde mutfakta hafif yemeklere aroma vermek için kullanılır ve Fransız mutfağında aromatik bitki karışımının içindeki bitkilerden birisidir.
   SUP: *bitki*, SUB: *frenk maydanozu*
6) SUB ... SUP (bütünüdür, tümüdür).
   Tedavi, sağlığı bozulmuş olan bireyi sağlıklı duruma kavuşturma amacıyla yapılan tıbbi işlemler bütünüdür.
   SUP: *işlem*, SUB: *tedavi*
7) SUB ... SUPlArın (bütünüdür, tümüdür).
   Dekor, bir oyun sırasında sahnede kullanılan ve oyunu tamamlayan aksesuarların tümüdür.
   SUP: *aksesuar*, SUB: *dekor*
8) SUB ... SUPdır.
   Astronomi, kökenleri, evrimleri, fiziksel ve kimyasal özellikleri ile gök cisimlerini açıklamaya çalışmak üzere gözleyen bilim dalıdır.
   SUP: *bilim dalı*, SUB: *astronomi*

Türkçe sözlük'te 6500 özgün kelime içinde çoklu kelimeler mevcuttur. Bu çalışma ise çoklu kelimeleri de baz alıyor, dolayısıyla, hypernym-hyponym ilişkisinde olan kelimeler, bir çok kelimeden oluşabiliyor. Tablo II çoklu kelime örneklerinden bir kaçını göstermektedir.

TABLO II: Çoklu Kelime Örnekleri

Hyponym	Hypernym
teokratik monarşi	*yönetim biçimi*
teokrasi	*devlet biçimi*
astronomi	*bilim dalı*
gemi	*ulaşım aracı*
bulaşık makinesi	*makine*
yılan balığı	*yılan*

Ön eleme kuralı sonucunda 6100 kelime tanımı çıkarmıştık. Bu kelime tanımları üzerinde yukarıda gösterdiğimiz çeşitli kurallar uygulayarak, Vikipedi'den son olarak 1600 adet hypernym-hyponym ilişkisi çıkarılmıştır.

C. Vikisözlük

Bir sonraki adımda, Vikisözlük'ü kullandık. Vikisözlük, Türkçe için online bir ansiklopedi ve genel bir sözlüktür. Vikipedi'den farklı olarak, daha farklı bir veri yapısı kullanmaktadır. Daha fazla etiket kullanarak bir sözlüğe daha yakın bir yapı içermektedir. Bu etiketlerden, "eşanlam", "kategori" örnek verilebilir.

Bu aşamada da Türkçe sözlük'teki 65000 özgün kelime Vikisözlük'te aranmış ve oradaki tüm kelimelerin tanımlarının bulunduğu ilk paragrafları alınmıştır. Bu kelimelerden sadece 11410 kelimesi Vikisözlük'te mevcuttur. Vikisözlük'te de ön eleme kuralları kullanıldı. Vikisözlük'ün yapısı Vikipedi'den farklı olduğundan dolayı Vikipedi'de kullanılan ön eleme kuralı burada kullanılmamıştır. Ön eleme kuralı, kelimenin tanımının geçtiği kısmı alır. Vikisözlük'te kullanılan ön eleme

kuralı şu şekilde özetlenebilir. "kategori" olarak işaretlenen bölümde genellikle kelimenin hypernym'i geçmektedir. Fakat bazen bu "kategori" kısmında kelimenin ait olduğu "alan" da geçebilmektedir. Kelimenin "alan" hypernym'den farklı olarak, bir kelimenin gerçek hayatta kullanıldığı ve atfettiği alanı belirtir. Örnek olarak, *futbol* kelimesinin alanı "spor" olarak belirlenmiştir, fakat hypernym olarak "oyun" geçmesi gerekir. Bu şarta uyan bazı genel alanları elenmiştir. Örnek olarak bu alanlar, "ad", "isim", "cümle", "deyim", "terim", "kişi" ve buna benzer genel isimlerdir. Bu ön eleme kuralı ile toplam 3075 girdi elenmiştir. Böylelikle, toplam geriye kalan kelime tanım sayısı, 8335 olmuştur.

Aşağıda bu kurallar kullanılarak üretilen bir kaç örnek verilmiştir. Burada SUP örnekteki hypernym'e, SUB ise örnekteki hyponym'e işaret etmektedir.

1) SUB: *mango* SUP: *bitki*, *meyve*
2) SUB: *üzüm* SUP: *asmagiller*, *meyve*
3) SUB: *altıntop* SUP: *sedef otugiller*, *bitki*

Vikisözlük verisinden bir kaç örnek Tablo III de verilmiştir.

TABLO III: Vikisözlük'ten çıkan hypernym-hyponym örnekleri

Hyponym	Hypernym
abanoz	*bitki*
alüminyum	*element*
altıparmak	*bitki*
anakonda	*yılan*
antilop	*boynuzlugiller*
arapça	*dil*
gül	*çiçek*
mango	*bitki*
üzüm	*meyve*
altıntop	*bitki*
balina	*memeli*

Vikisözlük'ten sonuç olarak 718 tane hypernym-hyponym ilişkisi elde edilmiştir.

III. Alan Bilgisi

Bölüm II-C'de, hypernym ile alan karışıklığından bahsetmiştik. Aslında, Vikisözlük ve Vikipedi'de bu ikisi arasında her zaman bir ayrım yapılmamaktadır. Bu çalışmanın bir aşaması olarak, Vikisözlük, Vikipedi'den elde ettiğimiz hypernym-hyponym adayları arasından, aslında kelime-alan olan ilişkileri temizledik.

Bunun için, Türkçe sözlük'te "alan" olarak işaretlenmiş kelimeleri çıkarıp, onları eleme kurallarımıza ekledik. Bu eleme sonucunda kelime-alan ilişkisini de elde etmiş olduk.

Tablo IV bazı önemli alanları ve kurallar sonucu çıkardığımız sayılarını gösteriyor.

Listelenen alanlar dışında Vikipedi, Vikisözlük kaynaklarından toplam 124 adet alan elde edilmiştir. Vikipedi de kullanılmadan attığımız kelimeler de kategori sınıfını oluşturduğu için hem Vikipedi hem de Vikisözlük'ten atılmıştır.

IV. Katmanlar

Türkçe sözlük'te işaretlenmiş olarak bulunan kelime-alan ilişkilerini de eledikten sonra Vikipedi ve Vikisözlük'ten elde ettiğimiz 2193 ilişki üzerinde daha derin ilişkiler bulmak

TABLO IV: Alanlardaki Kelime Sayısı

Alan	Sayı
Tıp	291
Din	248
Müzik	237
Matematik	242
Coğrafya	202
Edebiyat	191
Hukuk	188
Toplum bilimi	179
Biyoloji	176
Anatomi	158
Akrabalık	53
Eğitim	52
Asker	40

adına denemeler yapmaya başladık. Elde ettiğimiz sonuçlarda 2,3,4,5'li katmanlar halinde ilişkiler tespit ettik. Her bir katman için aşağıda örnekler listelenmiştir. Burada SUP örnekteki hypernym'e, SUB ise örnekteki hyponym'e işaret etmektedir.

1) 2 Katmanlı: *eflatun->renk*
   a) SUB: *eflatun* SUP: *renk*
2) 3 Katmanlı: *ebegümeci->bitki->canlı*
   a) SUB: *ebegümeci* SUP: *bitki*
   b) SUB: *ebegümeci*, *bitki* SUP: *canlı*
3) 4 Katmanlı: *lahana->turpgiller->bitki->canlı*
   a) SUB: *lahana* SUP: *turpgiller*
   b) SUB: *lahana*, *turpgiller* SUP: *bitki*
   c) SUB: *lahana*, *turpgiller*, *bitki* SUP: *canlı*
4) 5 Katmanlı: *amber balığı->balina->memeli->hayvan->canlı*
   a) SUB: *amber balığı* SUP: *balina*
   b) SUB: *amber balığı*, *balina* SUP: *memeli*
   c) SUB: *amber balığı*, *balina*, *memeli* SUP: *hayvan*
   d) SUB: *amber balığı*, *balina*, *memeli*, *hayvan* SUP: *canlı*

Her katman için yukarıda hyponym/hypernym ilişkilerini gösterdik. Elde ettiğimiz bu kelimeler ile daha düzgün ilişkilerin olduğunu kanıtlamış olup daha fazlası için yeni veriler bulmaya çalıştık. Tablo V'te bazı istatistikler verilmiştir.

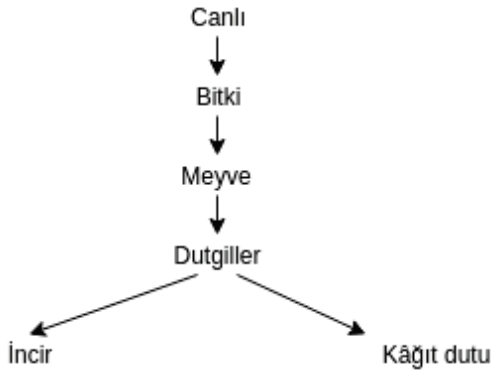
TABLO V: Hypernym Katmanları

	2 Katman	3 Katman	4 Katman	5 Katman	Toplam
Hypernym	652	55	40	22	1739

Hypernym katmanları için verdiğimiz tabloda sadece hangi katmandan kaç adet olduğu belirtilmektedir. Toplamda verilen katmanlar dışında kalan 579 adet ilişki 2 katmanlılardan oluşmaktadır. Tablo V'te verilen ve dışarda kalan diğer ilişkiler birer ilişki göstermemektedir. Örneğin *kükürt element*'in hyponymidir. Ve *element* için 58 adet hyponym bulunmaktadır. Bu ve diğerleri bu şekilde ilişkilendirilmiştir. Sonuç olarak bir iki hyponym'e sahip olan ilişkilerde elde ettiğimiz için her hypernym değerini 1 tane olarak kaydettik (Şekil 2).

V. Sonuç

Sonuç olarak, Türkçe için mevcut üç büyük online erişebilir ansiklopedi ve sözlük, Türkçe sözlük, Vikipedi, Vikisözlük'ü kullanarak, kelimeler arasında anlamsal ilişkiler bulmaya çalıştık. Bu anlamsal ilişki, hypernym-hyponym ilişkisidir. Bu çalışma çeşitli kurallar kullanarak bu ilişkileri



Şekil 2: Üretilmiş katman örneği

metinlerden çıkarıp ve yan bir çıktı olarak alan-kelime ilişkilerini çıkarmaktadır. Tablo VI'da bazı istatistikler verilmiştir.

TABLO VI: Hypernym İstatistikleri

	Vikipedi	Vikisözlük	TDK	Toplam
Hypernym	1600	718	2193	4511

Bu çalışmanın devamı olarak, başka anlamsal ilişkiler olan, meronym-holonym gibi ilişkileri bulunabilir.

KAYNAKÇA

[1] Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. Building a wordnet for turkish. *Romanian Journal of Information Science and Technology*, 7(1-2):163–172, 2004.

[2] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[3] Haşim Sak, Tunga Güngör, and Murat Saraçlar. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in natural language processing*, pages 417–427. Springer, 2008.

[4] Savas Yıldırım and Tugba Yıldız. Automatic extraction of turkish hypernym-hyponym pairs from large corpus. In *24th International Conference on Computational Linguistics*, page 493, 2012.