Sözlük Tabanlı Kavram Madenciliği: Türkçe için bir Uygulama

Cem Rıfkı Aydın
Boğaziçi Üniversitesi, Bebek 34342, Beşiktaş, İstanbul
cemrifkiaydin@gmail.com

Ali Erkan
Boğaziçi Üniversitesi, Bebek 34342, Beşiktaş, İstanbul
alierkan@gmail.com

Tunga Güngör
Boğaziçi Üniversitesi, Bebek 34342, Beşiktaş, İstanbul
gungort@boun.edu.tr

Hidayet Takçı
Cumhuriyet Üniversitesi, Kampüs 58140, Sivas
htakci@cumhuriyet.edu.tr

ÖZET
Bu makalede, dokümanlardan anlamlı kavramlar çıkarmak için sözlük-tabanlı bir metot geliştirilmiştir. Şu ana kadar İngilizce'de kavram çıkarma üzerine birçok çalışma yapılmıştır; ama sondan eklemeli bir dil olan Türkçe için bu alanda henüz geniş kapsamlı bir çalışma olmamıştır. Bu çalışmada biz resmi Türkçe sözlüğünden faydalandık. Normal sözlükler kavram madenciliğinde az kullanılmaktadır; ama kelimelerin sözlükteki anlam cümlelerindeki diğer kelimelerle başta hipernim, hiponim, eşanlam olmak üzere bazı ilişkilere sahip olduğu göz önünde bulundurularak yürütülen çalışma sonunda başarı oranları yüksek çıkmıştır.

Anahtar Kelimeler
Kavram madenciliği, biçimbirimsel analiz, biçimbirimsel muğlaklık giderme

SUMMARY
In this study, a dictionary-based method is used to extract expressive concepts from documents. So far, there have been many studies concerning concept mining in English, but this area of study for Turkish, an agglutinative language, is still immature. We used dictionary for concept extraction. The dictionaries are rarely used in the domain of concept mining, but taking into account that dictionary entries have synonyms, hypernyms, hyponyms and other relationships in their meaning texts, the success rate has been high for determining concepts.

Keywords
Concept mining, morphological analysis, morphological disambiguation

GİRİŞ

Kavram, olayların ve nesnelerin ortak özelliklerini barındıran ve ortak bir ad altında toplayan genel tasarımdır. Kavramlar, soyut veya somut olabilir. İnsanlar her ne kadar gördüğü, okuduğu veya duyduğu şeylerden genel bir anlam çıkarmada başarılı olabilse de, bilgisayarlar için aynı durum geçerli değildir, bu yüzden birçok istatistiksel veya makine öğrenme algoritmalarından faydalanılmıştır.

Kavram madenciliği, yazınsal, görsel ve işitsel materyallerden kavramlar çıkarmak için kullanılan bir terimdir. Kavram madenciliği metotlarından çoğu yazınsal materyallerden kavram çıkarmak için geliştirilmiştir. Bu makalede de sadece yazınsal dosyalardan kavram çıkarmakla ilgilenilmiştir. Kavram madenciliği zor; ama kullanımı oldukça geniş ve yararlı bir alandır. Örneğin hukuksal alanlarda davaları sınıflandırmada kullanılabilir [1], ayrıyeten tıp alanında da teşhisleri koymada ve hasta sınıflandırmalarında yardımcı bir rol üstlenebilir [2] [3]. Açık uçlu anketlerin değerlendirilmesinde, dokümanların sınıflandırılmasında ve müşteri profillerinin değerlendirilmesi gibi alanlarda da kavram madenciliğinden yararlanılabilir.

Şu ana kadar kavram madenciliği alanındaki çalışmaların çoğunda, sözlüksel bir veritabanı olan WordNet'ten yararlanılmıştır. WordNet'te kelimelerin birbiriyle olan ilişkileri tanımlayan *synset* adında ilişki kümeleri vardır. Bu ilişki kümelerinde hiponim, eşanlamlılık, zıt anlamlılık gibi bir sürü ilişki türleri vardır. Kavram madenciliğinde ise en çok başvurulan ilişki türü hipernimi ilişkisidir; çünkü hipernimi bir kelimenin genel anlamını barındıran bir ilişkidir ve kavramlar da şeylerin genel tasarımını ifade eder. İngilizce gibi çok yaygın dillerde WordNet oldukça titizlikle çok kapsamlı olarak hazırlanmıştır; ama Türkçe dahil birçok dilde WordNet hala birçok özelliği eksik, tamamlanmamış bir yapıya sahiptir.

Bu çalışmada, kavram madenciliği için Türkçede daha önce denenmemiş bir metot olarak Türk Dil Kurumu (TDK) sözlüğü kullanılmıştır. Sözlük kelimelerinin anlam cümlelerinde, kelimenin kendisiyle hipernimi, hiponimi, eş anlamlılık ve nedensellik gibi anlam ilişkisi olan birçok sözcük bulunmaktadır ve bu ilişkilerden bir kelimenin genel anlamını yani kavramını çıkarmada faydalanılmıştır. Ayrıca frekans gibi birçok faktör kavram çıkarmada göz önünde bulundurulmuştur. Yaptığımız çalışmada çıkan başarı oranı, Türkçe üzerine yapılan diğer çalışmalara oranla daha yüksek çıkmıştır.

Literatür araştırması bölümünde, kavram madenciliği üzerine yapılan çalışmalardan kısaca bahsedilmektedir. Metodoloji bölümünde, bu çalışma için geliştirdiğimiz algoritma anlatılmaktadır. Değerlendirme sonuçları bölümünde, geliştirdiğimiz metodolojinin başarı oranları verilmektedir. Son olarak ise sonuç bölümü gelmektedir.

LİTERATÜR ARAŞTIRMASI

Kavram madenciliği üzerinde birçok dilde çalışma yapılmıştır, bu çalışmaların çoğunda makine öğrenme algoritmaları ile WordNet'ten faydalanılmıştır. Türkçede de kavram madenciliği konusunda geliştirilen metotlarda makine öğrenme algoritmalarının kullanıldığı görülmüştür.

Bir çalışmada WordNet'in *synset* ilişkileri kullanılarak kümeleme (clustering) algoritması izlenmiştir [4]. Ama bu çalışmada muğlaklık giderme yapılmamıştır ve kümeleme algoritmasında *synset* kelimelerinin tümünün kullanılması başarı oranını azaltmıştır.

Yürütülen başka bir çalışmada *ConceptNet* denen, kelimelerin birbiriyle olan fiziksel, sosyal, zamansal ve diğer birçok ilişkilerini kapsayan bir bilgi veritabanı kullanılmıştır. Bu veritabanı WordNet ile kıyaslandığında daha zengin bir yapıya sahiptir. Burada veritabanındaki ilişkiler grafiksel bir biçimde ifade edilmiştir ve frekanslar da göz önünde bulundurularak grafikte en çok girdi alan düğümler (node) ana kavramlar olarak belirlenmiştir [5].

Türkçede ise kavram madenciliği üzerine çok az çalışma olmuştur. Çalışmaların birisinde kümeleme algoritması izlenmiştir. Başta manüel olarak kelimeler belli kümelere atanmıştır, sonra ise test dokümanlara bu kümeler atanarak kavram çıkarılmıştır [6]. Geliştirilen metodolojinin başarı oranı % 51'dir.

Çalışmalardan birinde ise metin sınıflandırılmasında kavram madenciliğinden faydalanılmıştır. Kavramlar belli faktörler (frekans, dokümanlarda bulunma sıklığı) göz önünde bulundurularak çıkarılmıştır [7]. Fakat çıkarılan kavramlar sadece dokümanda bulunan kelimeler arasından seçilmektedir, dolayısıyla çok başarılı sonuçlar çıkmamıştır.

Ayrıyeten kavram çıkarmada yapay zeka algoritmalarından olan Latent Dirichlet Allocation da kullanılmıştır [8]. Bu algoritma aslında dokümanlardan konu çıkarmakta kullanılmaktadır. Bizim izlediğimiz metodolojide ise herhangi bir makine öğrenme algoritması kullanılmamıştır, bunun yerine istatistiksel bir metot izlenilmiştir.

METODOLOJİ

Kavram madenciliği alanında en çok başvurulan yöntemlerden biri WordNet kullanımıdır; çünkü WordNet'teki hipernimi özelliği bir kelimenin genel anlamını ifade eder ve kavramlar da olayların ve nesnelerin genel anlamı olduğu için bu ilişkinin kullanımı oldukça yararlı olmuştur. Fakat hipernimi dışındaki eşanlamlılık gibi ilişkiler de bir dokümandan kavram çıkarma da kullanılabilir. Örneğin eğer bir dokümanda *doktor* kelimesi çok sık geçiyorsa, bu

```
<entry>
   <name> jaguar </name>
   <affix>undefined</affix>
   <lex_class>isim, zooloji </lex_class>
   <stress>undefined</stress>
   <pronunciation> Fransızca jaguar </pronunciation>
   <origin> Fransızca</origin>
 - <meaning>
      <meaning_class>undefined</meaning_class>
      <meaning_text>  Kedigillerden, Orta ve Güney
         Amerika'da yaşayan, postu iri benekli memeli
         türü (Felis onca).</meaning_text>
    - <quotation>
         <author>undefined</author>
         <quotation_text>undefined</quotation_text>
      </quotation>
   </meaning>
   <atasozu_deyim_bilesik>undefined</atasozu_deyim_bilesik>
   <birlesik_sozler>undefined</birlesik_sozler>
</entry>
```

Şekil 1. Jaguar kelimesinin XML formatındaki sözlükteki yapısı.

dokümandan çıkarılan kavramlardan birisi *hekim* olabilir. WordNet ne kadar sık kullanılsa da Türkçe için pek gelişmiş bir veritabanı değildir, bunu da hesaba katarak biz bu çalışmada TDK sözlüğünden faydalandık.

Sözlük Yapısı

Biz bu çalışmada XML formatındaki TDK sözlüğünden faydalandık. Şekil 1, bu formattaki sözlüğün bir kelimesini ve bunun özelliklerini göstermektedir.

Burada biz sadece <*name*> (ad), <*lex_class*> (sözcük kategorisi) ve <*meaning*> (anlam) etiketlerinden faydalandık ve diğer özellikleri eledik. Burada biz sadece sözcük kategorisi isim olan kelimeleri hesaba kattık; çünkü kavramlar genelde isim olan kelimelerden çıkartılabilir. Burada birçok farklı anlam olabilir o yüzden doğru anlamı seçmek için bağlam analizi gerçekleştirdik. Sözlük anlam cümlelerinde geçen ilişki türlerinden bazıları aşağıdaki gibidir:

- *Eşanlam:* Eş anlamları olan kelimeler. Simetrik bir ilişkidir.
- *Bileşen:* Bir kelimenin başka bir kelimenin bileşeni olması durumudur. Simetrik bir ilişki değildir.
- *Yer:* Kelimenin bulunduğu yeri tanımlayan ilişki.
- *Kullanabilirlik:* Kullanım amacını içeren ilişki. (Örneğin diş fırçası diş fırçalama için kullanılır.)
- *Etki:* Bir eylemin sebep olduğu sonucu içeren ilişkidir.
- *Hipernim:* Kelime ile daha genel/üst kavram arasındaki ilişkidir.
- *Hiponim:* Kelime ile daha özel/alt kavram arasındaki ilişkidir.
- *Altolay:* Bir eylemin içerebileceği alt eylemle olan ilişkidir. (Sabah uyanırken esnemek gibi.)
- *Önkoşul:* Bir eylemin başka bir eylemin ön şartı olması durumu.

Bu sözlük anlam ilişkileri bir çok yolda kullanılabilir, örneğin iki kelime arasındaki anlam yakınlığını bulmak ve benzerlik kurmak için bu ilişkileri kullanabiliriz. Kümeleme yoluyla anlam benzerliği bulunan kelimeler aynı kümeye atanabilir. (Örneğin *el* ve *parmak* sözcükleri bileşen ilişkisine sahiptir, böylece bunlar aynı kümede bulunabilir.) Fakat biz bu çalışmada kümeleme yerine daha basit bir istatistiksel yol izledik.

Sözlük Ön-İşleme

WordNet gibi sözlüksel veritabanları yapısal (structured) olmasına rağmen, TDK sözlüğü belli bir yapıya sahip değildir, bu yüzden Boğaziçi Üniversitesi'nin geliştirdiği muğlaklık giderici araçlar olan BoMorP ve BoDis kullanılmıştır [9] [10]. Bu araçlar kelimenin kökünü, eklerini ve sözcük kategorisini (isim, sıfat vs.) bulmaktadır. Kavramlar isimler olduğu için biz bu araçların döndürdüğü kelimeler arasından çekim ekleri elenmiş isimleri seçtik. Böylece kelimelerin anlam cümlelerindeki isimleri de sonraki işlemlere tabi tuttuk.

Sözlüğü Kullanarak Bağlamsal Muğlaklık Giderici

Bir sözcük sözlükte bir sürü farklı anlama sahip olabilir, biz bu yüzden dokümanlardaki kelimelere bakarak, hangi anlamlarda kullanıldıklarını o kelimelerin bağlamlarına bakarak belirledik. Bunun için dokümandaki çok anlamı bulunan kelimelerin sağındaki ve solundaki 15 kelimeye baktık ve sözlük anlam cümlelerinden hangisi bu 30 kelimelik bağlam sözcüklerinden en fazlasını kapsıyorsa, o anlam cümlesini kelimenin kullanılmakta olan anlamı olarak belirledik.

Kavram Atama

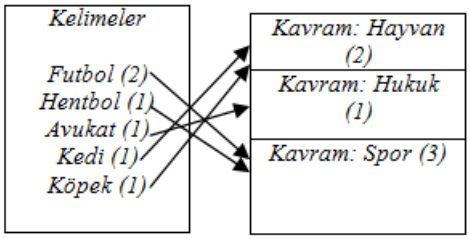
Basit Sözlük Algoritması

Bir dokümanın kavramı genelde o dokümanda en sık geçen kelimelerle alakalıdır. Örneğin bir dokümanda *futbol* kelimesi çok sık geçiyorsa, bu dokümanın kavramı *spor* diyebiliriz. Kavram madenciliği konusunda hesaba katılan faktör genelde sadece frekans iken, biz iki unsuru daha göz önünde bulundurduk:

$$\text{Kavram}(k) = \text{Frek}(k) \times \text{İlkYer}(k) \times \text{Kapsam}(k) \quad (1)$$

Yukarıdaki formüle göre bir kelimenin diğer kelimelere oranla daha baskın bir kelime olması için dokümandaki frekansı yanında, bulunduğu ilk konum ve kelimenin kapsamı da göz önünde bulundurulmalıdır. Bir kelime ne kadar dokümanın ilk yerlerinde geçtiyse (başlık, ilk paragraf vs.) o kadar önemli diyebiliriz. Aynı zamanda bir kelime dokümanın başındaki ve sonundaki cümlelerde geçiyorsa onun kapsamı daha fazla diyebiliriz ve bu kelime sadece dokümanın ortasındaki paragrafta geçen diğer bir kelimeye göre daha önemlidir diyebiliriz.

Bu algoritmaya göre bir matris oluşturulur ve matrisin sıra kelimeleri dokümanda geçen kelimeleri, kolon kelimeleri de bu doküman kelimelerinin sözlükteki



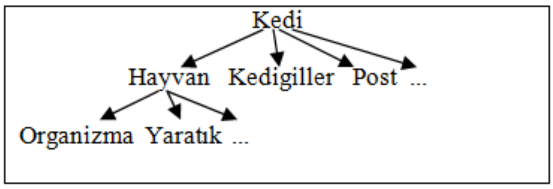
Şekil 2. Bir dokümandan kavram çıkarma örneği.

anlam cümlelerindeki sözcükleri temsil etmektedir. Kelimeler matristeki sıra ve kolonlarda en fazla bir defa gösterilmektedir. Sonra matris sıfır veya bir değerleriyle doldurulur.

Eğer dokümandaki kelime *futbol* ise ve bunun anlam cümlesindeki kelimeler *takım*, *spor* ve *oyun* ise matriste *futbol* satırındaki *takım*, *spor*, *oyun*, *futbol* kolonlarının değerleri bir olur, bu sıradaki diğer matris değerleri sıfır olur. (Ayrıyeten bütün matristeki sıraları temsil eden kelimeler kolon kelimelerine bir kere eklenir; çünkü bir dokümandaki kavram o dokümanda geçen kelimeler arasından da seçilebilir.) Matristeki bütün değerler böyle doldurulunca sonunda bütün değerler, o değerlerin bulunduğu sırayı temsil eden kelimelerin frekans, kapsam ve ilk bulunduğu yer faktörleriyle çarpılır. Sonra matristeki kolon değerleri toplanır ve hangi kolon kelimeleri en yüksek değerleri veriyorsa, onlar dokümanın asıl kavramları diyebiliriz. Şekil 2'ye göre dokümanın asıl kavramı spordur diyebiliriz. (Soldaki kolonda doküman kelimeleri gösterilmektedir, sağda ise bunların anlam cümlelerindeki kelimeler gösterilmektedir, parantez içerisindeki değerler, frekans ve diğer faktörlerin çarpım değeridir.)

Yukarda anlatıldığı gibi sözlükten faydalanarak kavram çıkarma modeli geliştirilmiştir, dokümandaki sözcükler ve onların anlam cümlelerindeki kelimeler hesaba katılıyorken iki seviyeli bir yapı kurduğumuz göz önünde bulunduruluyor; fakat aynı zamanda bu çalışmada üç ve dört seviyeli hiyerarşik yapılar da kullanılarak kavram çıkarılmaya çalışılmıştır. Bu yapıya göre, doküman kelimesi hiyerarşik yapının en üstünde, bu kelimenin anlam cümlesindeki kelimeler bir alt seviyede, bu anlam cümlesi kelimelerinin de sözlükteki anlam cümlesi kelimeleri bir alt seviyededir. Kedi kelimesi için örnek Şekil 3'te verilmiştir.

Üç veya daha fazla basamaklı yapılar kurarken, her basamağa, yukarıdan aşağıya giderken azalan bir değer atanmıştır ve matristeki değerler bu sayılarla çarpılarak kavramlar çıkarılmıştır; çünkü yukarıdan aşağıya giderken en üstteki kelimenin alt basamaklardaki kelimelerle anlam ilişkisi zayıflamaktadır. Yaptığımız çalışmada en başarılı sonuçlar iki seviye için alınmıştır.



Şekil 3. Kedi kelimesi için kurulan 3 basamaklı hiyerarşik yapı.

Ayrıyeten bu algoritma için sözlükteki kelimelerin anlam cümlelerinde en fazla bulunan %1 kelime elenmiştir. Bu algoritmaya göre dokümanda geçmeyen kelimeler de kavram olarak önerilebilmektedir.

Bağlam Algoritması

Basit sözlük algoritması her ne kadar büyük derlemlerin ve uzun dokümanların kavramını çıkarmada başarılı olsa da, küçük metinlerden bu algoritmaya göre anlamlı kavramlar çıkarılamamaktadır. Bunun nedeni ise bir kelimenin kavramının, o kelimenin anlam cümlesindeki bütün kelimelerle alakasının olmamasıdır. Örneğin *futbol* kelimesinin bir anlam cümlesi kelimesi *kale*dir ve algoritmada matris oluşturulurken bu da hesaba katılmaktadır; ama *futbol*un genel kavramının *kale* kelimesi ile alakası yoktur. O yüzden bu algoritmanın bir alternatifi üretilmiştir.

Bu yeni algoritmaya göre dokümandaki kelimelerin bütün derlemde sağındaki ve solundaki 15 kelimeye bakılmıştır ve eğer bu kelimeler, asıl kelimenin anlam cümlesinde de geçiyorsa bunlar hesaba katılmıştır, diğer geçmeyen anlam cümlesi kelimeleri elenmiştir. Burada da iki yol izlenmiştir:

1. Bir kelime için sadece hem anlam cümlesinde hem de o kelimenin bütün derlemdeki dokümanlardaki 30 kelimelik bağlamlarında geçen kelimeler hesaba katılmıştır. Bu seçilen kelimelere de derlemdeki bağlam frekanslarına göre değer atanarak, matriste bu değerler de hesaba katılmıştır.
2. Hangi anlam cümlesi kelimesi bağlamlarda en sık geçiyorsa, sadece o hesaba katılmıştır, diğer anlam cümlesi kelimeleri elenmiştir.

Matris her iki yönteme göre birinci algoritmada olduğu gibi doldurulur ve hangi kolon değeri en yüksek değeri veriyorsa o dokümanın asıl kavramıdır diyebiliriz. İkinci metoda göre her doküman kelimesi için o kelimenin kendisi ve anlam cümlesindeki kelimelerden derlem bağlamlarında en fazla geçen kelime olmak üzere en fazla iki kelime hesaba katılırken, birinci metoda göre bu sayı ikiden fazla olabilir. Bu iki metot önceki algoritmaya göre daha iyi sonuçlar vermektedir; çünkü her anlam cümlesi kelimesi, o kelimenin kavramıyla alakalı değildir.

DEĞERLENDİRME SONUÇLARI

Geliştirdiğimiz metodolojiyi dört derlemdeki dokümanlardan kavram çıkarmak için kullandık. Bu derlemlerden ikisi hukuk alanındaki makale ve haberlerden, birisi spor alanındaki haberlerden, bir diğeri de birçok çeşitli mühendislik ve mimarlık raporlarından oluşmaktadır. Hukuk alanındaki dokümanları içeren derlemlerde fazla bir konu farklılığı yokken, diğerlerinde konu farklılığı gözlenmektedir. Toplamda 368 dokümandan kavramlar çıkarılmıştır.

Değerlendirme metriği olarak *doğruluk* kullanılmıştır. Bu metrik aşağıda verilmiştir.

$$Doğruluk = \frac{Doğru\ bulunan+}{\frac{Doğru\ olarak\ bulunmayan}{Doğru\ bulunan+\ Bulunamayan+\ Yanlış\ bulunan+\ Doğru\ olarak\ bulunmayan}} \quad (2)$$

Değerlendirme işlemleri için geliştirdiğimiz metodolojinin çıkardığı kavramlarla, manüel olarak çıkardığımız kavramları karşılaştırdık. Bu karşılaştırmayı yaparken ilk 3, 5, 7, 8, 9, 10 ve 15 kavramları karşılaştırdık. Örneğin bir dokümanda algoritmanın önerdiği üç baş kavram ile manüel olarak çıkarttığımız üç baş kavram arasındaki kelime sayısı ne kadar ortaksa doğruluk oranı o kadar artmıştır. Bu 3 kelimeden ikisi ortaksa başarı oranı 2 / 3 = 0.67 olarak belirlenir.

Değerlendirme sonuçlarında izlenen başka bir yol da algoritmanın bulduğu baş kavramları o dokümandan manüel olarak çıkarttığımız bütün kavramlarla kıyaslamaktır. Doğal olarak buradaki başarı oranı, sınırlı kıyaslama (3'e 3, 10'a 10 gibi) yollarına göre çok daha yüksek çıkmaktadır.

Tablo 1 değişik derlemler için elde edilen doğruluk oranlarını göstermektedir. Birinci algoritma için (Basit sözlük algoritması) en başarılı sonuçlara iki seviyeli hiyerarşik yapıyı ve frekans gibi faktörleri hesaba katan alt-metot ile ulaşılmıştır. 2-seviyeli olup frekans gibi faktörleri hesaba katmayarak veya 3-seviyeli yapıyla oluşturulan matrislerin kavram çıkarma başarısının düşük olduğu gözlenmiştir. İkinci algoritma ortalama olarak daha iyi sonuçlar vermiştir. Karşılaştırma pencereleri açısından baş 3 kelimeleri kıyaslarken daha yüksek yüzdelik başarı oranına ulaşılmıştır. Değişik derlemler için farklı sonuçlarla karşılaşılmasının nedeni de derlemlerin bazılarının tek birkaç konuda yoğunlaşırken bazılarının bir sürü farklı konuda hazırlanmış dokümanlardan oluşmasıdır. İlk algoritmanın 3, 5, vs. gibi algoritmanın bulduğu baş kavramlarla manüel olarak bulunanlarla sınırsız karşılaştırma kullanıldığında % 65,86 başarı oranına sahip olurken, ikinci algoritmanın ikinci alternatifinin % 78,88 gibi bir başarı oranına sahip olduğu görülmüştür.

Tablo 2 ise örnek amacıyla birinci algoritmanın bütün alt-metodlarıyla ikinci algoritmanın ikinci alternatifini bir derlem için kıyaslanmasını göstermektedir. Birinci

Derlemler	Kıyaslama Pencere Büyüklüğü	Basit Sözlük Algoritması (2-seviye, faktör)	Bağlam Algoritması (Alt-metot 1)	Bağlam Algoritması (Alt-metot 2)
Gazi	k = 3	62.30	73.90	74.00
	k = 5	59.00	69.60	70.10
	k = 7	58.30	67.30	67.70
	k = 9	57.60	65.40	65.60
Spor Haberleri	k = 3	60.80	70.10	71.00
	k = 5	59.60	66.30	67.80
	k = 7	58.00	64.90	66.80
	k = 10	57.10	62.20	65.40
Yargıtay Haberleri	k = 3	64.60	73.43	71.97
	k = 5	60.13	70.09	70.57
	k = 7	58.46	68.33	68.62
	k = 8	58.07	67.93	67.47
Yargıtay Kararları	k = 3	73.70	99.20	95.30
	k = 5	73.10	88.10	84.00
	k = 7	69.00	82.30	79.40
	k = 10	64.20	76.00	75.60

Tablo 1. Farklı derlemler için doğruluk oranları.

algoritma (2-seviye, 1-0) matrisi sadece kolon kelimelerinin sırayı temsil eden kelimelerin anlam cümlelerinde geçip geçmemesine bakarak 0 ve 1 değerleriyle doldurmaktadır, frekans ve diğer faktörler hesaba katılmamaktadır. Birinci algoritma (2-seviye, faktör) ise önceki metoda ilaveten frekans, ilk konum ve kapsam faktörlerini hesaba katmaktadır. Birinci algoritma (3-seviye, katsayı) ise 3-basamaklı katsayılı kelime yapısını göz önünde bulundurmaktadır. İkinci algoritma (2) ise ikinci algoritmanın ikinci alt metodudur.

	Karşılaştırma Pencere Büyüklüğü			
Algoritma	**k = 3**	**k = 5**	**k = 7**	**k = 10**
Birinci Algoritma (2-seviye, 1-0)	76.81	67.45	63.21	59.44
Birinci Algoritma (2-seviye, faktör)	73.70	73.10	69.00	64.20
Birinci Algoritma (3-seviye, katsayılı)	68.80	63.00	59.90	58.30
İkinci Algoritma (2)	95.30	84.00	79.40	75.60

Table 2. Yargıtay Kararları derlemi için değişik karşılaştırma pencere büyüklükleri ve dört değişik algoritma kullanımı ile elde edilen performans sonuçları.

SONUÇ

Bu makalede, kavram madenciliği için Türkçede daha önce denenmemiş bir metot izlenmiş, Türkçe sözlüğü kullanılmıştır. Şu ana kadar olan çalışmaların çoğunda WordNet'in hipernimi özelliği kullanılmıştır; fakat şu da hesaba katılmalıdır ki tek bir ilişki özelliği dokümandan kavram çıkarmada yetersiz olabilmektedir.

Ayrıyeten sözlük kullanımı sayesinde metinde geçmeyen kelimeler kavram olarak belirlenebilmektedir.

Kavram çıkarmak için sözlüğü kullanırken, aynı zamanda derlemlerde bağlam bazlı bir yöntem izlenmiş ve bazı anlam cümleleri kelimeleri elenmiştir. Bu mantıklı bir yaklaşımdır; çünkü bir sözlüğün genel anlamı, sözcük anlam cümlesindeki bütün kelimelerinden çıkarılamaz, bazıları elenmek zorundadır.

İleride geliştirdiğimiz bu metodoloji başka Türkçe derlemleri için de çalıştırılabilir ve dokümanlardan kavramlar çıkartılabilir. Algoritmanın biraz değişmesi şartıyla İngilizce derlem dokümanlarından da kavramlar çıkartılabilir. Bu makalenin konusuyla bağlantılı olarak sözlük yardımı ile kümeleme (clustering) yapılabilir ve anlamlı kavramlar elde edilebilir.

KAYNAKÇA

[1] Moens, M., Angheluta, R., 2003. 'Concept Extraction from Legal Cases: The Use of a Statistic of Coincidence', *International Conference on Artificial Intelligence and Law*, ICAIL, ACM.

[2] Faber, V., Hochberg, J.G., Kelly, P.M., Thomas, T.R., White, J.M., 1994. 'Concept

Extraction – a datamining technique', Los Alamos Science.

[3] Bennett, N.A., He, Q., Chang, C.T.K., Schatz, B.R., 1999. 'Concept Extraction in the Interspace Prototype', Technical Report, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL.

[4] Pennock D., Dave K. and Lawrence S., 2003. 'Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews', *Twelfth International World Wide Web Conference (WWW'2003),* ACM

[5] Liu, H., Singh, P., 2004. 'ConceptNet - a practical commonsense reasoning toolkit', BT Technology Journal.

[6] Uzun, M., 2011. 'Developing a concept extraction system for Turkish', *International Conference on Artificial Intelligence*, ICAI.

[7] Chengzhi, Z., Dan, W., 2008. 'Concept Extraction and Clustering for Topic Digital Library Construction', *International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE/WIC/ACM.

[8] AlSumait, L., Barbar´a, D., Domeniconi, C., 2008. 'OnLine LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking', *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining,* ICDM.

[9] Sak, H., Güngör, T., and Saraçlar, M., 2008. 'Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus', *GoTAL 2008*, vol. LNCS 5221, pp. 417-427, Springer.

[10] Sak, H., Güngör, T. and Saraçlar, M., 2007. 'Morphological disambiguation of Turkish text with perceptron algorithm', *CICLing 2007*, vol. LNCS 4394, pp. 107-118.

ÖZGEÇMİŞLER

Cem Rıfkı Aydın

Yazar Lisans derecesini (B.Sc.) 2011 yılında Bahçeşehir Üniversitesi, Bilgisayar Mühendisliği bölümünden almıştır. Şu anda Boğaziçi Üniversitesi Bilgisayar Mühendisliğinde Yüksek Lisans (M.Sc.) öğrencisidir ve Master öğrenimi süresince TÜBİTAK tarafından BIDEB bursuyla desteklenmiştir. Araştırma alanları doğal dil işleme, yapay zeka uygulamaları, şekil tanıma, oyun programcılığı, bilgi alma ve genetik algoritmalarıdır.

Ali Erkan

Yazar Lisans (B.Sc.) ve Yüksek Lisans (M.Sc.) derecelerini Bilkent Üniversitesi, Endüstri Mühendisliğinden, ayrıyeten Yüksek Lisans (M.Sc.) derecesini 2010 yılında Boğaziçi Üniversitesi Yazılım Mühendisliğinden almıştır. Şu anda Boğaziçi Üniversitesi Bilgisayar Mühendisliğinde doktora (Ph.D.) öğrencisidir.Araştırma alanları doğal dil işleme, makine öğrenmesi, şekil tanıma, bioinformatik ve istastiktir.

Tunga Güngör

Yazar doktora (Ph.D.) derecesini 1995 yılında Boğaziçi Üniversitesi, Bilgisayar Mühendisliğinden almıştır. Şu anda Boğaziçi Üniversitesi, Bilgisayar Mühendisliği bölümünde doçenttir. Araştırma alanları doğal dil işleme, makine öğrenmesi ve şekil tanımadır. Yazarın 60 civarında yayınlanmış bilimsel makalesi vardır ve birçok araştırma projelerine ve konferans düzenlemelerine katılmıştır.

Hidayet Takçı

Yazar Lisans derecesini (B.Sc.) Trakya Üniversitesi, Bilgisayar Mühendisliği bölümünden 1997 yılında almıştır. Yüksek Lisans (M.Sc.) eğitimini Bilgisayar Mühendisliği bölümünde Gebze Yüksek Teknoloji Enstitüsü'nde 1999 yılında tamamlamış, Doktora derecesini ise Bilgisayar Mühendisliği bölümünde 2005 yılında Gebze Yüksek Teknoloji Enstitüsü'nden almıştır. Post Doktora çalışmalarına Boğaziçi Üniversitesi'nde devam etmektedir. Araştırma alanları veri ve metin madenciliği, dil tanıma, yapay sinir ağları, risk analizi bilişim suçları, yazar tanıma, kriminal veri madenciliği, bilgisayar ağları ve güvenliktir.