Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması

Measurement of Turkish Word Semantic Similarity and Text Categorization Application

M.Fatih Amasyalı [1], Aytunç Beken [1]

1. Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
mfatih@ce.yildiz.edu.tr, aytuncbeken@yahoo.com

Özetçe

Literatürde metin sınıflandırma çalışmalarında ya her boyutun bir kelimeye ya da ngrama karşılık geldiği çok büyük boyutlu uzaylarda (Bag-of-Words) işlem yapılmakta ya da kelimeler kümelenerek metinler daha az boyutlu uzaylarda temsil edilmektedir. Bu çalışmada önerdiğimiz yaklaşımda ise öncelikle metinlerde geçen kelimleler anlamsal bir uzayda konumlandırılmıştır. Bunun için iki kelimenin birbirlerine anlamsal benzerliğinin, kelimelerin birlikte geçtiği doküman sayısıyla doğru orantılı olduğunu öne süren Harris'in hipotezi kullanılmıştır. Daha sonra anlamsal uzaydaki konumlardan faydalanarak bu kelimeleri içeren metinlerin anlamsal uzaydaki konumları elde edilmiş ve bu konumlar kullanılarak metinler sınıflandırılmıştır. Bu yaklaşım Türkçe haber metinleri üzerinde uygulanmış ve geleneksel metotlara göre daha başarılı sonuçlar elde edilmiştir.

Abstract

In literature, texts to be classified are generally represented in the large dimensional Bag of Words space in which every dimension equals to a word or ngram. In this study, firstly the words are placed in a semantic space. The word's coordinates in semantic spaces needs the similarity of the words according to their meanings. Harris states that two words' semantic similarity is related to the number of documents which the words are both in. We used his hypothesis for Turkish words. Firstly, we obtained word co-occurrence matrix from a web corpus. Then, the numerical coordinates of the words are calculated by using multi dimensional scaling. Texts coordinates are obtained from word coordinates which passes in the texts. In our experiments, Turkish news texts are classified into 5 classes. We get more successful results than the traditional Bag of Words space. Our approach is not for only Turkish words/texts, but also for all other languages.

1. Giriş

İnsan dilinin bilgisayarlar kullanılarak işlenmesini üretilmesini amaçlayan çalışmalarda kelime anlamları (semantik) en güç konulardan biridir. Bununla birlikte kelime anlamları metin sınıflandırma, soru cevaplama, kelime anlamını durulaştırma, arama motorlarından daha iyi sonuçlar elde etme ve otomatik metin özetleme gibi birçok uygulama için vazgeçilmez kaynaklardır. Bu nedenle kelime anlamlarının bilgisayar ortamında ifade edilmesi için birçok çalışma yapılmıştır. Bunun için başlıca iki yöntem benimsenmiştir[1]. İlki insan gücüyle büyük hiyerarşik kelime haritalarının oluşturulmasıdır ki buna örnek olarak Wordnet[2] ve Verbnet[3] verilebilir. Diğeri ise büyük metin kütüphanelerinde istatistikî metotlar kullanılarak otomatik kelime haritalarının üretilmesidir. Bu yaklaşımda en önemli sorun kelimelerin birbirlerine anlamca benzerliklerinin ölçülmesidir. Harris [4], iki kelimenin birlikte geçtiği doküman / cümle sayısının iki kelimenin benzerliğiyle doğru orantılı olduğunu öne sürmüştür. Daha önceki çalışmamızda bu yaklaşım kullanılarak 3 adet küçük veri kümesi üzerinde kelimelerin anlamsal konumları, arama motorları kullanılarak bulunmuş ve bu konumlarına göre kelimeler anlamsal kategorilere başarıyla ayrılmıştır [5].

Kelimelerin anlamsal benzerliklerinin bulunması için literatürde birçok çalışma yapılmıştır. Yuhua Li ve Jay J. Jiang çalışmalarında kelimelerin benzerliğini büyük metin kütüphanelerinde (külliyat/corpus) kelimelerin birlikte geçme sıklıklarını ve kelimelerin Wordnet hiyerarşisinde birbirlerine olan uzaklıklarını birlikte kullanarak ölçmüşlerdir [1,6]. Sonuçlarını insan deneklerin verdiği cevaplarla karşılaştırmışlardır. Guihong Cao ve arkadaşları [7] ise kelimelerin birlikte geçme sıklıklarını Reuters külliyatında hesaplamışlar ve kelimeleri Fuzzy K-means algoritmasıyla kümelemişlerdir.

Bu çalışmamızın ise başlıca iki amacı bulunmaktadır. İlki önceki çalışmamızda elde ettiğimiz başarının kelime sayısı arttığında da sürüp sürmeyeceğinin araştırılması, ikincisi ise bu anlamsal konumların metin sınıflandırma alanında kullanılmasıdır. Metin sınıflandırma çalışmaları çok fazla farklı kelime içeren ve bu sebeple metinlerin büyük boyutlu veri kümeleri olarak ifade edildiği çalışmalardır. Bu nedenle seçtiğimiz Türkçe metin sınıflandırma problemi bizi iki amacımıza da ulaştırabilecek niteliktedir.

Metin sınıflandırma konularında önceden yapılan çalışmalarda ya çok büyük boyutlu uzaylarda (Bag-of-Word) işlem yapılmakta ya da kelimeler kümelenerek metinler daha az boyutlu uzaylarda temsil edilmektedir. Bu küçük boyutlu modelleme çalışmalarında sınıflandırma başarısı artmış, işlem zamanı ve hafıza ihtiyacı azalmış ve büyük boyutlu uzaylarda çalıştırılamayan kompleks algoritmalar bu küçük boyutlu uzayda uygulanabilmiştir [8-11]. Bizim çalışmamızda ise önceki çalışmalardan farklı olarak, öncelikle metinlerin içinde geçen kelimelerin anlamsal benzerliklerine uygun sayısal koordinatları bulunmuş ve daha sonra metinlerin sayısal koordinatları, içinde geçen kelimelerin koordinatları kullanılarak bulunmuştur. Daha sonra metinlerin koordinatları kullanılarak sınıflandırma yapılmıştır. Klasik BOW uzayına göre daha başarılı sonuçlar elde edilmiştir.

Bildirinin ikinci bölümünde önerilen metodun ayrıntıları ve deneysel sonuçlar anlatılmıştır. Son bölümde ise çalışmadan elde edilen bulgular özetlenmiştir.

2. Metinlerin Sınıflandırılması

Bu bölümde öncelikle sınıflandırılan haber verileri tanıtılmış daha sonra metinlerin sayısallaştırılması işlemi anlatılmıştır. Daha sonra metinlerin çeşitli algoritmalarla sınıflandırılmasıyla elde edilen sonuçlar verilmiştir.

2.1. Kullanılan Veriler

5 farklı haber türüne (ekonomi, magazin, sağlık, siyasi, spor) ait 230'ar metin sınıflandırma için kullanılmıştır. Her gruptan 150'şer haber metni eğitim için, 80'er adedi test için ayrılmıştır. Metinler günlük haber sitelerinden rastgele toplanmıştır. Türkçe eklemeli bir dildir. Bir kelime köküne çeşitli ekler getirilerek çok sayıda farklı kelime türetilebilir. Ekler işin içine katıldığında metinlerdeki farklı kelime sayıları çok büyümekte ve işlem zamanını çok fazla arttırmaktadır. Bu sebeple PCKimmo [12] kullanılarak kelimelerin gövdeleri bulunmuştur. Diğer bir ifadeyle metinleri ifade etmek için içerdikleri kelimelerin kendileri değil gövdeleri kullanılmıştır.

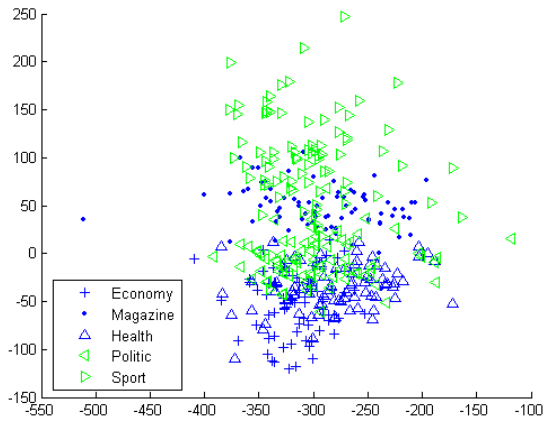
2.2. Metinlerin Koordinatlarının Bulunması

Metinlerde geçen toplam farklı gövde sayısı yaklaşık 4500'dür. Bu 4500 kelimenin sayısal karşılıklarını elde etmek 15.000 web sitesinden oluşan bir külliyat kullanılmıştır. Kelimelerin birbirlerine anlamsal yakınlık matrisi bu külliyatta birlikte geçtikleri doküman sayılarıyla oluşturulmuştur. Birbirlerine uzaklıkları bilinen ancak koordinatları bilinmeyen nesnelerin bir uzaklık matrisine uygun olan sayısal koordinatlarının bulunması için Çok Boyutlu Ölçekleme (Multi Dimensional Scaling) metodu kullanılır. Bu nedenle elimizdeki kelimelerin anlamsal yakınlık matrisi ters çevrilerek (uzaklık=1/yakınlık) uzaklık matrisi elde edilmiş ve çok boyutlu ölçekleme fonksiyonuna verilmiştir. Bunun sonucunda 4500 kelimenin 3602 boyutlu uzayda karşılık geldikleri sayısal vektörler / koordinatlar elde edilmiştir. 3602 boyutla işlem yapmak zor olduğundan ve çok boyutlu ölçeklemede boyutların anlamlılıkları başta sona doğru azaldığından baştan ilk 100 ve ilk 10 boyut alınarak denemeler yapılmıştır.

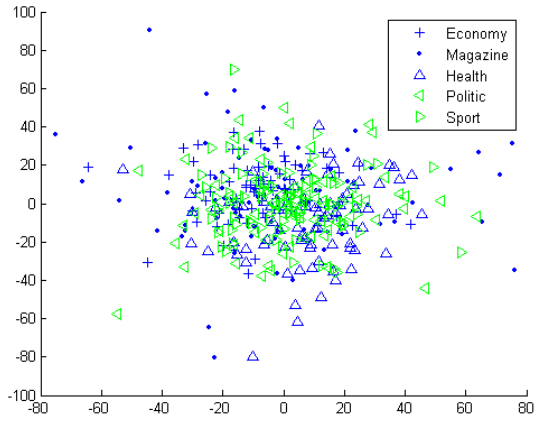
Metnin içinde geçen kelimelerin vektörlerinin ortalaması alınarak metin vektörleri elde edilmiştir. Metinler P boyutlu bir anlamsal uzayda ifade edilmek istenirse, hesaplanan kelime koordinatlarının ilk P adedi dokümanların koordinatlarının bulunmasında kullanılacaktır. K adet kelimeden (w) oluşan bir D dokümanın (P boyutlu uzayda) koordinatlarının hesaplanması Eşitlik 1'de verilmiştir. Dokümandaki her bir w kelimesi P boyutlu bir vektördür.

$$D_p = \frac{\sum_{i=1}^{K} w_{ip}}{K} (p = 1..P) \quad (1)$$

Bu şekilde kelimeler gibi metinler de anlamsal bir uzayda ifade edilmişlerdir. Şekil 1'de 400 test haber metninin çeşitli boyutlardaki dağılımları verilmiştir.



(a) 1-2. boyutlar



(b) 23-24. boyutlar

Şekil 1.: Metinlerin 2 boyutlu uzayda çeşitli boyutlara göre dağılımları.

Örneğin, Şekil 1-a'da yatay düzlem 1. anlamsal boyuta, dikey düzlem ise 2. anlamsal boyutu göstermektedir. Şekil 1'den görüldüğü gibi metinlerin birbirlerinden 2 boyutlu uzayda ilk boyutlarda tam olmasa da birbirlerinden bir ölçüde ayrılabilmektedirler. Ayrıca ilk boyutların sonrakilerden daha anlamlı oldukları da gözükmektedir.

2.3. Metinlerin Sınıflandırma Sonuçları

Metinleri ifade eden sayısal vektörler bulunduğundan veriler çeşitli makine öğrenmesi algoritmalarıyla birlikte kullanılmaya uygun hale gelmiştir. 2. bölümde adı geçen WEKA [14] yazılımında yer alan algoritmalar kullanılarak 750 metnin verileriyle sistem eğitilmiş, 400 metin ile de test edilmiştir. Test sonuçları Tablo 1 ve 2'de verilmiştir. Tablo 1'de metinlerin kelimelerin anlamsal vektörleriyle ifade edildiğindeki, Tablo 2'de ise metinlerin klasik metotlarla ifade edildiğindeki başarı oranları görülmektedir.

Tablo 1: Metinlerin kelimelerin anlamsal vektörleriyle ifade edildiğindeki başarı oranları

Sınıflandırma Algoritması	**100 Boyutlu Metinler**	**10 Boyutlu Metinler**
Lineer Regresyonla Sınıflandırma	**93.25**	89
Pace Regresyonla Sınıflandırma	92.75	88.5
En Yakın Komşu	70	79
Destek Vektör Makineleri	90	89.75
Rastgele Ormanlar (100 ağaçlı)	90.75	87.75

Metinler kelimelerin frekanslarıyla ifade edildiğinde her bir metin tüm metinlerdeki farklı kelime sayısı boyutunda bir vektörle gösterilmektedir. Literatürde bu metin vektörü farklı şekillerde elde edilmektedir [15]. Karşılaştırma yapmak için en sık kullanılan 2 tanesi ile yapılan denemeler Tablo 2'de verilmiştir. Metinler vektörlerle gösterildiğinde (metin sayısı * farklı kelime sayısı) boyutlu bir matris oluşturmaktadır. Bu matrisin elemanlarının 2 farklı şekilde elde edilişi Eşitlik 2 ve Eşitlik 3'te verilmiştir.

$$M_{ij} = tf_{ij} \tag{1}$$

$$M_{ij} = \log(tf_{ij} + 0.5) * \log(\frac{D}{df_i}) \tag{2}$$

Eşitlik 2 ve 3'te tf_{ij}, i kelimesinin j dokümanında geçme sayısını, D, toplam doküman sayısını, df_i ise i kelimesini içeren doküman sayısını göstermektedir.

Tablo 2'de tüm işlemler bu iki kelime frekansı temsili kullanılarak elde edilen veriler üzerinde gerçekleştirilmiştir. 4500 boyutlu veriler üzerinde bazı algoritmalar çalışma zamanı ve/veya hafıza problemleri yüzünden çalıştırılamamıştır. Bu nedenle 4500 boyuttan 100 tanesi WEKA'nın infogain metodu kullanılarak seçilmiş ve algoritmalar çalıştırılmıştır. Klasik metotlardaki iki kelime frekansında en başarılı sonuç Eşitlik 3 ile elde edilmişken Eşitlik 2'in ortalama başarısı Eşitlik 3'ten daha yüksektir ve daha güvenilir sonuçlar üretmektedir.

Tablo 2: Metinlerin klasik metotlarla ifade edildiğindeki sınıflandırma başarı oranları (X, uygulanamadı anlamındadır)

Sınıflandırma Algoritması	**Metinlerin Boyut Sayısı**	**Başarı yüzdesi (Eşitlik 2)**	**Başarı yüzdesi (Eşitlik 3)**
Klasik Naive Bayes	4500	87.25	
Diskrit Naive Bayes	4500	85.75	**89.25**
En Yakın Komşu	4500	34.25	43.5
Destek Vektör Makineleri	4500	87	86.5
Rastgele Ormanlar (100 ağaçlı)	4500	X	X
Lineer Regresyonla Sınıflandırma	4500	X	X
Pace Regresyonla Sınıflandırma	4500	X	X
C4.5	4500	74.75	23.5
Lineer Regresyonla Sınıflandırma	100	81.75	84.5
Pace Regresyonla Sınıflandırma	100	81.5	83.25
En Yakın Komşu	100	71.25	76.25
Destek Vektör Makineleri	100	80.25	87.75
Rastgele Ormanlar (100 ağaçlı)	100	85	84.5

Tablo 1 ve 2 incelendiğinde Tablo 1'deki başarının çok daha yüksek olduğu ve dolayısıyla metinleri kelimelerin anlamsal uzayında ifade etmenin klasik kelime frekansı uzayında ifade etmekten daha başarılı olduğu görülmektedir. Tablo 2'nin en iyi performansı 89.25% iken Tablo 1'in en iyi performansı Lineer Regresyonla Sınıflandırma algoritmasıyla elde edilen 93.25%'dir. Bu denemeye ait karışım (confusion) matrisi Tablo 3'te verilmiştir.

Tablo 3: 400 Test Metnine ait Karışım Matrisi (Lineer Regresyonla Sınıflandırma – Metinler 100 boyutlu)

Gerçekler ↓ / Bulunanlar →	Ekonomi	Magazin	Sağlık	Siyasi	Spor
Ekonomi	73	4	1	2	0
Magazin	2	73	3	1	1
Sağlık	2	1	75	1	0
Siyasi	4	1	2	73	0
Spor	0	1	0	1	78

Tablo 3 incelendiğinde diğerlerinden en iyi ayrılan sınıfın 97.5%'luk başarıyla spor haberleri olduğu görülmektedir.

3. Sonuç ve Gelecek Çalışmalar

Türkçe kelimelerin anlamsal sınıflara ayrılması ve bu özellikler kullanılarak metin sınıflandırılması konusunda yapılan ilk çalışma sunulmuştur. Kelimelerin sınıflandırılabilmesi için bir şekilde kelimelerin benzerliklerinin ölçülmesi gerekir. Türkçe iki kelime arasındaki benzerliğin büyük metin kütüphanelerinde belirli bir pencere içinde birlikte geçme sayılarıyla ölçülebileceği öne sürülmüş ve bu konuda çeşitli denemeler yapılmıştır. Daha sonra önerilen metodun metin sınıflandırma da kullanılabileceği düşünülmüş ve bir uygulama yapılmıştır. Elde edilen bulgular aşağıda listelenmiştir:

- Benzerlik matrislerinin elde edilmesinde arama motorlarının kullanımı daha büyük miktarda veri üzerinde çalışma olanağı vermesine rağmen arama zamanı çok uzun olacaktır. N adet kelimenin benzerlik matrisi için arama motoruna N*(N–1)/2 adet sorgu gönderilmesi gerekmektedir. Arama motorları bir kullanıcıdan bir günde yapılabilecek sorgu sayısını yaklaşık 1000'le sınırlamaktadırlar. Bu sebeple N'in örneğin 1000 değeri için benzerlik matrisinin oluşturulabilmek için yaklaşık 500 gün beklemek gerekecektir. Bunun yerine bu çalışmada yaklaşık 15.000 web sitesinden oluşan bir metin kütüphane kullanılarak bu sorun çözülmüştür ve elde edilen başarıya bakıldığında bu rakamın yeterli olduğu görülmektedir. Bununla birlikte daha büyük boyuttaki metin kütüphaneleri kullanıldığında başarının daha da artabileceği düşünülmektedir.

- Önerilen kelime ve buna bağlı olarak metin koordinatlarının bulunması metodu sadece Türkçe'ye özgü bir metot değildir. Her dil için kullanılabilir.

- Eşanlamlı kelimelerin tek bir kelime olarak değerlendiriliyor oluşu metodun dezavantajıdır.

- Benzerlik matrisi bulunurken, kelimelerin birlikte geçtikleri doküman sayısı yerine belirli bir boyuttaki bir pencere içerisinde birlikte geçme sayısı da denenmiş ancak sınıflandırma başarısını çok fazla etkilemediği görülmüştür.

- Kelime koordinatlarının bir uygulaması olarak gerçeklenen metin sınıflandırma işleminde klasik yollarla (Naive Bayes, Terim Frekansı) sınıflandırmaya göre daha başarılı sonuçlar (93.25%) elde edilmiştir. Üstelik metinler bu yeni metotla çok daha az boyutta ifade edilmiştir.

- Kelimelerin koordinatlarından metinlerin koordinatlarının bulunmasında kullanılan metnin içerdiği tüm kelimelerin ortalamasını almak basit olmasına rağmen başarılı bir metottur. Bununla birlikte gelecekte bu alanda başka yaklaşımlar da geliştirilmelidir.

Sonuç olarak metin sınıflandırma alanında elde edilen yüksek başarılar, kelimelerin sayısal koordinatlarının daha başka uygulama alanlarında (kelime anlamı durulaştırma, otomatik soru cevaplama vs.) da kullanmamız için ümit vericidir.

4. Teşekkür

Çalışmamıza yaptıkları değerli katkılardan dolayı Tunga Güngör, Nilgün Dursunoğlu ve Seyhan Amasyalı'ya teşekkürler ederiz.

5. Kaynakça

[1] Yuhua Li, Zuhair A. Bandar, David McLean, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", IEEE Transactions on Knowledge and Data Engineering, vol. 15, pp. 871-882, 2003.

[2] http://wordnet.princeton.edu

[3] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[4] Haris Zelig S., "Mathematical structures of language", Wiley, pp.12, 1968.

[5] Amasyalı M.F. "Arama Motorları Kullanarak Bulunan Anlamsal Benzerlik Ölçütüne Dayalı Kelime Sınıflandırma", Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 2006.

[6] Jay J. Jiang, David W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", International Conference Research on Computational Linguistics (ROCLING X), Taiwan, 1997.

[7] Guihong C., Dawei S., Peter B., "Fuzzy K-Means Clustering on a High Dimensional Semantic Space", Advanced Web Technologies and Applications (LNCS 3007) - The Sixth Asia Pacific Web Conference (APWeb'04), 2004.

[8] Bekkerman R., Ran El-Yaniv, Naftali T., Yoad W., "Distributional Word Clusters vs. Words for Text Categorization", Journal of Machine Learning Research, pp.1-48, 2002.

[9] Inderjit S. D., Subramanyam M., Rahul K., "Enhanced Word Clustering for Hierarchical Text Classification" The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), 2002.

[10] Baker L. D., McCallum A. K., "Distributional Clustering of Words for Text Classification", 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98), 1998.

[11] Courtney C., Rada M., "Measuring the Semantic Similarity of Texts", ACL Workshop on Emprical Modeling of Semantic Equivalance and Entailment, pp. 13-18, 2005.

[12] Oflazer, K. "Two-level description of Turkish morphology. Literary and Linguistic Computing", 9(2), pp. 137-148, 1994.

[13] Martinez W. L., Martinez A. R., "Exploratory Data Analysis with MATLAB", Chapman & Hall/CRC, p.61, 2004.

[14] www.cs.waikato.ac.nz/ml/weka

[15] Ciya L., Shamim A., Paul D., "Feature Preparation in Text Categorization", Oracle Text Selected Papers and Presentations ,2001.