Türkçe ile Avrupa Dilleri Arasındaki Anlamsal Benzerliğin Kelime Temsilleri ile Gösterimi

Semantic Similarity between Turkish and European Languages Using Word Embeddings

Lütfi Kerem Şenel^{1,2,3}, Veysel Yücesoy¹, Aykut Koç¹, Tolga Çukur^{2,3,4}

¹ASELSAN Research Center, Ankara, Turkey

²Elektrik ve Elektronik Mühendisliği Bölümü, Bilkent Üniversitesi, Ankara, Türkiye

³Sabancı Beyin Araştırmaları Merkezi, UMRAM, Bilkent Üniversitesi, Ankara, Türkiye

⁴Sinirbilimi Programı, Bilkent Üniversitesi, Ankara, Türkiye

Email: {lksenel,vyucesoy,aykutkoc}@aselsan.com.tr, cukur@ee.bilkent.edu.tr

Özetçe —Kelime temsilleri, bir dilin kelime hazinesindeki kelimelerin yüksek boyutlu bir uzayda gerçel vektörler ile temsil edilmesidir. Elde edilen vektörlerin kelimeler arasındaki anlamsal ilişkileri başarılı bir şekilde modellediği ve birçok doğal dil işleme uygulamasında performansı arttırdığı gözlemlenmiştir. Kelime temsilleri, temel olarak İngilizce için geliştirilmiş olmakla beraber birçok başka dil için de başarılı sonuçlar vermiştir. Bu çalışmada, oluşturulan bir kelime listesi üzerinden kelime temsilleri kullanılarak Türkçe (iki farklı derlem) ile beş temel Avrupa dili (İngilizce, Almanca, Fransızca, İspanyolca, İtalyanca) arasındaki anlamsal benzerlik ölçülmüş ve yapılan istatistiksel testler ile sonuçların geçerliliği doğrulanmıştır. Bunun yanı sıra derlem seçiminin ve ek ön işleme adımlarının Türkçe için hazırlanan benzetim ve farklılık bulma testlerindeki performansı nasıl etkilediği incelenmiştir.

Anahtar Kelimeler—kelime temsilleri, doğal dil işleme, diller arası anlamsal benzerlik.

Abstract—Representation of words coming from vocabulary of a language as real vectors in a high dimensional space is called as word embeddings. Word embeddings are proven to be successful in modelling semantic relations between words and numerous natural language processing applications. Although developed mainly for English, word embeddings perform well for many other languages. In this study, semantic similarity between Turkish (two different corpora) and five basic European languages (English, German, French, Spanish, Italian) is calculated using word embeddings over a fixed vocabulary, obtained results are verified using statistical testing. Also, the effect of using different corpora, and additional preprocess steps on the performance of word embeddings on similarity and analogy test sets prepared for Turkish is studied.

Keywords—word embeddings, natural language processing, semantic similarity between languages.

This work was supported in part by a Marie Curie Actions Career Integration Grant (PCIG13-GA-2013-618101), by a European Molecular Biology Organization Installation Grant (IG 3028), and by a TUBA GEBİP fellowship awarded to T. Çukur.

I. GİRİŞ

Diller arasındaki benzerlik farklı alanlardaki araştırmacılar tarafından istatistiksel, dilbilimsel ya da sinirbilimsel açılarından araştırılmıştır. Örneğin Mikolov [1], zengin veritabanına sahip olmayan dillere kelime temelli çeviri yapabilmek için diller arası anlamsal benzerliği göz önünde bulunduran doğrusal bir model geliştirmeye çalışmıştır. Başka bir çalışmada [2] Almanca ve Yunanca gibi morfolojisi zengin diller arasındaki anlamsal benzerlik, bağlamsal ve eşdizilimsel metrikler kullanılarak ölçülmeye çalışılmıştır. Correia [3] yaptığı çalışmada çift anadili olan bireylerin beyninde, iki dilde aynı anlama gelen kelimelerin fonetik olarak farklı olsalar da benzer şekilde temsil edildiğini göstermiştir.

Bir dilin anlamsal özelliklerini modellemek amacıyla kelime temsilleri günümüzde sıkça kullanılan bir yöntemdir. Kelime temsilleri bir dilin sözcük hazinesindeki kelimeleri, kelimeler arasındaki anlamsal ilişkileri koruyacak şekilde belirli boyuttaki bir vektör uzayına yansıtılmasıyla elde edilir. Kelimeleri gerçel vektörler kullanarak temsil etme fikri LSA ve rastgele yerleştirme (randomized embedding) yöntemleri ile başlamıştır [4]. O zamandan beri kelime vektörlerini sözdizimsel, morfolojik ve anlamsal yönlerden geliştirmek ve optimize etmek için çok sayıda çalışma yapılmıştır [5], [6].

Kelime temsillerinin bir dili başarılı bir şekilde modellemesini sağlayan ve onlara günümüzdeki popülerliğini kazandıran, son yıllarda ortaya atılmış olan word2vec [7] ve GloVe [8] algoritmalarıdır. Farklı yöntemler kullanmalarına karşın iki algoritma da benzetim testlerinde (i.e. *anne - teyze + baba = ? hala*) oldukça başarılı sonuçlar veren kelime vektörleri oluşturabilmektedir.

Bahsedilen algoritmaların bu başarısından dolayı, kelime temsilleri günümüzde doğal dil işleme (DDİ) uygulamaları ve DDİ'yi kullanan sinirbilim, bilgisayarla görü, bilişimsel dilbilim gibi alanlar için bir temel oluşturmaktadır. Karpathy [9] ve Huth [10] tarafından yapılan çalışmalar, kelime temsillerinin bilgisayarla görü ve sinirbilime ilişkin uygulama alanlarına örnek olarak verilebilir.

Bu makale; Türkçe ile beş temel Avrupa dili (İngilizce,

Almanca, Fransızca, İspanyolca, İtalyanca) arasındaki kelime tabanlı anlamsal yakınlığı kelime temsilleri kullanarak seçilen bir sözcük listesi üzerinden incelemeyi amaçlamaktadır. Türkçe kelime vektörlerini öğrenmek için kullanılan derlem seçiminin etkisini gözlemlemek için analizler farklı bir Türkçe derlemi kullanılarak tekrarlanmıştır. Diller arası benzerlik analizinin yanı sıra farklı derlem seçiminin, derlemin gereksiz kelimelerden temizlenmesinin ve kelimelere gelen eklerin atılmasının elde edilen kelime vektörleri üzerindeki etkisi, Türkçe için oluşturulan bir benzerim ve farklılık test kümeleri üzerinde test edilmiştir.

Bölüm II, çalışmanın ana konusu olan diller arası anlamsal benzerlik çalışmalarını kapsamaktadır. Bölüm III aynı dilin farklı derlemlerinden oluşturulan temsillerin farkının incelendiği bölümdür. Bölüm IV ile makalenin sonuçları belirtilmiş ve çalışma sonlandırılmıştır.

II. DİLLER ARASI BENZERLİK

A. Derlemler

Boyut olarak yeterince büyük olmaları, dilbilgisi olarak yeterli seviyede düzgün olmaları ve çok çeşitli konuları kapsamaları sebebiyle seçilen Avrupa dillerindeki bütün Wikipedia¹ sayfa ve makaleleri o dillerin kelime vektörlerinin öğrenilmesi için derlem olarak seçilmiştir. Bu Avrupa dilleri için kullanılan derlemlerin uyumlu olması sayesinde, derlemlerin yapısından ya da aralarındaki farklardan kaynaklanabilecek diller arası anlamsal farklılıkların da önüne geçilmiştir. Türkçe Wikipedia içerik olarak uyumlu olmasına karşın boyut olarak Avrupa dillerinden önemli derecede küçüktür (8-40 kat). Bu nedenle Türkçe için ikinci bir derlem olarak Boğaziçi Üniversitesi'nin hazırladığı internet haberlerinden oluşan derlem (Boun Web) de [11] kelime vektörlerinin öğrenilmesi için kullanılmıştır.

B. Ön İşleme

İndirilen Wikipedia veritabanı yığınlarını temizlemek ve metin olarak depolamak için bir Python kodu² kullanılmıştır. Bu işlemten sonra, elde edilen dosyalar bir araya getirilmiş, başlıklar ve alfabetik olmayan karakterler (rakamlar ve noktalama işaretleri gibi) kaldırılmıştır. Bütün metinler küçük harfe çevrilmiş ve son olarak, GloVe algoritmasının kelime vektörlerini öğrenmede kullanabilmesi için tek satır halinde yazılmıştır. Kesme işareti ile ayrılmış olan ekler ayrı birer kelime olarak alınmıştır (Türkiye'deki → türkiye deki). Bu işlemlerin ardından elde edilen İngilizce, Almanca, Fransızca, İspanyolca, İtalyanca, Türkçe Wikipedia ve Boun Web derlemleri sırasıyla yaklaşık 1170, 410, 295, 251, 220, 52 ve 412 milyon kelimeden oluşmaktadır.

C. Sözcük Listesinin Hazırlanması

İncelenen altı dil için kullanılan yedi farklı derlem, ön işlemeden geçtikten sonra kelime temsillerini öğrenmek üzere GloVe algoritmasına girdi olarak verilmiştir. GloVe algoritması çalıştırılırken kullanılan VOCAB_MIN_COUNT parametresinin derlemlere göre değerleri Tablo I'te verilmiştir. Bu parametre kelime temsillerini öğrenme sırasında kullanılacak olan sözcük hazinelerinin birbirine yakın boyutlarda olmasını

TABLE I: VOCAB_MIN_COUNT parametresinin derlemlere göre değeri.

Parametre Adı	İng	Fra	Alm	İta	İsp	Tür	Boun
VOCAB_MIN_COUNT	50	17	52	12	15	5	30

sağlayacak şekilde seçilmiştir. VECTOR_SIZE, MAX_ITER, WINDOW_SIZE, X_MAX parametreleri bütün derlemler için sırasıyla 300, 50, 15 ve 100 olarak ayarlanmıştır.

Diller arasındaki anlamsal benzerliklerin ölçülebilmesi için dilleri temsil etmek üzere 2434 İngilizce kelimeden oluşan bir sözcük listesi oluşturulmuştur. Kelimeler seçilirken, dillerin anlamsal özelliklerinin iyi bir şekilde yansıtılabilmesi adına İngilizce'de sık kullanılan ve çok çeşitli konulara ait kelimeler seçilmiştir, bu İngilizce kelimeler profesyonel tercümanlar aracılığıyla diğer dillere tercüme edilmiştir. Çeviri sonucunda en az bir dille, birden fazla kelime ile ifade edilmiş olanlar listeden çıkarılmıştır. Natural language toolkit³ (NLTK) tarafından her dil için ayrıca tanımlanmış olan, anlamsal ilişki içermeyen sözcükler de (stop words) listeden çıkarılmıştır. Sonuç olarak 1869 kelimelik bir sözcük listesi kullanılmıştır.

D. Deneyler

Temsili benzerlik analizi (representational similarity analysis), sayısal modeller ile ölçülen sinirsel aktiviteleri ilişkilendirmek için literatürde sıklıkla kullanılan bir yöntemdir [12], [13]. Bu yöntemde sayısal modeller, temsili benzeşmezlik matrisi (representational dissimilarity matrix) adı verilen matrisler kullanılarak tanımlanır. Bu çalışmada hazırlanan sözcük listeleri üzerinden oluşturulan benzeşmezlik matrisleri kelime temsilleri yoluyla modellenmiş olan dillerin arasındaki anlamsal benzerliği hesaplamak için kullanılmıştır. Bu çalışmada diller arası kelime tabanlı anlamsal benzerliği bulmak için izlene yol aşağıda verilen iki varsayım üzerinden açıklanabilir:

Varsayım 1. L_1 ve L_2 iki farklı dil olmak üzere, L_1 'den M adet kelime seçerek $W_1 = \{w_{1,1}, w_{1,2}, \dots, w_{1,M}\}$, L_2 'den M adet kelime seçerek $W_2 = \{w_{2,1}, w_{2,2}, \dots, w_{2,M}\}$ tanımlanabilir. Burada bütün $i \in \{1, 2, \dots, M\}$ için $w_{2,i}$ $w_{1,i}$ 'nin L_2 diline çevirisidir.

Varsayım 2. E ve S vektör kümeleri N boyutlu ve sırasıyla W_1 ve W_2 kümelerindeki kelimelere karşılık gelen vektörlerden oluşmaktadır. (örneğin $i \in \{1, 2, \dots, M\}$ için $e_i \in \mathbb{R}^{N \times 1}$ $w_{1,i}$ 'e karşılık gelen ve $s_i \in \mathbb{R}^{N \times 1}$ $w_{2,i}$ 'e karşılık gelen kelime temsilleri olmak üzere $E = \{e_1, e_2, \dots, e_M\}$, $S = \{s_1, s_2, \dots, s_M\}$).

Bir dile ait benzeşmezlik matrisini hesaplamak için korelasyon tabanlı bir uzaklık metriği kullanılmıştır. Korelasyon tabanlı benzeşmezlik matrisi LE şu şekilde tanımlanmıştır:

$$LE_{i,j} = 1 - \frac{\hat{e}_i^T \hat{e}_j}{\|\hat{e}_i\|_2 \|\hat{e}_j\|_2} \quad (1)$$

ve $\hat{e}_i$ şu şekilde hesaplanmıştır:

$$\hat{e}_i = e_i - \mu_{e_i} \quad \text{ve} \quad \mu_{e_i} = \frac{1}{N} \sum_{l=1}^N e_i[l] \quad (2)$$

¹<https://www.wikipedia.org>

²<http://github.com/attardi/wikiextractor>

³www.nltk.org

Burada $e_i[l]$, e_i vektörünün l 'inci elemanıdır. S kümesinin benzeşmezlik matrisi LS aynı şekilde hesaplanmıştır. Sonrasında L_1 ve L_2 dillerinin arasındaki kelime tabanlı anlamsal benzerlik Denklem 3 kullanılarak hesaplanmıştır.

$$\rho = \text{corr}(\text{LE}, \text{LS}) \quad (3)$$

Burada kelimelerin kendileriyle olan ilişkilerini gözardı etmek için korelasyon benzeşmezlik matrislerinin ana köşegenleri haricindeki elemanları üzerinden hesaplanmıştır.

Denklem 3, bu çalışmada diller arasındaki anlamsal benzerliği hesaplamak için kullanılan temel yöntemi ifade etmektedir. Elde edilen korelasyonun 1'e yakın olması dillerin anlamsal yönden birbirine benzediklerini ifade ederken, 0'a yakın olması dillerin arasında zayıf bir ilişki olduğunu ifade eder. Bu deney sonucunda Türkçe derlem olarak Wikipedia kullanılarak elde edilen korelasyon değerleri Tablo II'te, Boğaziçi derlemi kullanılarak elde edilen değerler Tablo III'de verilmiştir.

Denklem 3 ile elde edilen korelasyon değerlerinin gerçekten diller arasındaki anlamsal benzerliği yansıttığını doğrulamak, bu değerlerin şans eseri veya GloVe algoritmasındaki bir yönelimin sonucu olmadığını göstermek ve güven aralığı tanımlayabilmek için çeşitli istatistiksel testler uygulanmıştır.

Sonuçların istatistiksel analizi için tekrarsız yeniden örnekleme yöntemi (Bootstrap without replacement) kullanılmıştır. Varsayım 1 ve 2'den yola çıkarak, tekrarsız b tane indis, testin her yinelenmesinde rastgele ve bütün $i \in \{1, 2, \dots, b\}$ için $t_i < M$ koşulunu sağlayacak şekilde seçilmiş ve $T = \{t_1, t_2, \dots, t_b\}$ elde edilmiştir. Daha sonra oluşturulan $X = \{e_{t_1}, e_{t_2}, \dots, e_{t_b}\} = \{x_1, x_2, \dots, x_b\}$ ve $Y = \{s_{t_1}, s_{t_2}, \dots, s_{t_b}\} = \{y_1, y_2, \dots, y_b\}$ kelime kümelerine karşılık gelen vektörler üzerinden yeni benzeşmezlik matrisleri denklem 4 kullanılarak elde edilmiştir.

$$\text{LE}_{i,j}^{(k)} = 1 - \frac{\hat{x}_i^T \hat{x}_j}{\|\hat{x}_i\|_2 \|\hat{x}_j\|_2} \quad (4)$$

Burada $k \in \{1, 2, \dots, r\}$ için $i, j \in \{1, 2, \dots, b\}$. $\text{LS}^{(k)}$ da aynı şekilde hesaplanmıştır.

$\text{LE}^{(k)}$ ve $\text{LS}^{(k)}$ benzeşmezlik matrisleri $b \times b$ boyutundadır ve r testin tekrarlanma sayısını gösterir. Diller için benzeşmezlik matrisleri oluşturulduktan sonra korelasyon (3) kullanılarak hesaplanmıştır. Test r kere tekrarlandıktan sonra diller arasındaki korelasyonların ortalamaları (μ) ve standart sapmaları (σ) hesaplanmıştır. $r = 1000$ için elde edilen ortalama ve standart sapma değerleri, %95 güven aralıkları ile birlikte Tablo II ve Tablo III'te verilmiştir. %95 güven aralıkları, 1000 denemenin sonucundaki korelasyonlar sıralandığında 25. ve 975. değerlere karşılık gelmektedir.

Elde edilen korelasyon değerlerini doğrulamak için iki farklı sıfır hipotezi testi yapılmıştır. Bunlardan birincisinde, birbirlerinin tercümelemleri olan W_1 ve W_2 kelime kümeleri yerine, elemanlarının yerlerinin rastgele karıştırılmasıyla elde edilen yeni kümeler ve bunlara karşılık gelen vektörler kullanılmıştır. Bu sayede, elde edilen korelasyonların iki dildeki kelimeler arasındaki anlamsal ilişkiden kaynaklanıp kaynaklanmadığı test edilmiştir.

İkinci sıfır hipotezi testinde W_1 ve W_2 kümelerindeki kelimelerin yerleri sabit tutulup E ve S vektör kümelerindeki vektörlerin elemanları rastgele bir şekilde karıştırılmıştır. Bu

TABLO II: Türkçe Wikipedia derlemi üzerinden Türkçe ile Avrupa dilleri arasındaki anlamsal benzerlikler, Bootstrap 1000 kere tekrarlanarak elde edilen güven aralıkları, ortalama ve standart sapma değerleri.

Dil	Anlamsal Benzerlik	Güven Aralığı (%95)	Bootstrap (μ)	Bootstrap (σ)
Türkçe - İngilizce	0.4986	0.4883 - 0.5086	0.4985	0.0053
Türkçe - Almanca	0.4253	0.4142 - 0.4363	0.4252	0.0056
Türkçe - Fransızca	0.4381	0.4271 - 0.4505	0.4383	0.0060
Türkçe - İspanyolca	0.4576	0.4452 - 0.4705	0.4576	0.0064
Türkçe - İtalyanca	0.4548	0.4436 - 0.4659	0.4545	0.0058

TABLO III: Türkçe Wikipedia derlemi üzerinden Türkçe ile Avrupa dilleri arasındaki anlamsal benzerlikler, Bootstrap 1000 kere tekrarlanarak elde edilen güven aralıkları, ortalama ve standart sapma değerleri.

Dil	Anlamsal Benzerlik	Güven Aralığı (%95)	Bootstrap (μ)	Bootstrap (σ)
Boun Web - İngilizce	0.4909	0.4814 - 0.5015	0.4910	0.0052
Boun Web - Almanca	0.4203	0.4094 - 0.4308	0.4201	0.0055
Boun Web - Fransızca	0.4286	0.4167 - 0.4408	0.4283	0.0061
Boun Web - İspanyolca	0.4509	0.4391 - 0.4614	0.4507	0.0057
Boun Web - İtalyanca	0.4524	0.4405 - 0.4636	0.4522	0.0058

sayede korelasyonların GloVe algoritmasının yapısından kaynaklanıp kaynaklanmadığı test edilmiştir.

İki sıfır hipotezi testi de 1000'er kere tekrarlanmış ve elde edilen korelasyonlardan p değerleri, sıfır testinden gelen korelasyonun asıl deney ile elde edilen korelasyondan büyük olma olasılığı şeklinde hesaplanmıştır. İki test için de $p < 0.001$ elde edildiğinden sıfır hipotezi reddedilmiş ve korelasyonların gerçekten diller arasındaki anlamsal benzerliklerden kaynaklandığı doğrulanmıştır.

Tablo II ve Tablo III'a bakıldığında Türkçe'nin anlamsal yönden İngilizce'ye diğer Avrupa dillerinden daha fazla benzediği görülmektedir.

Bunun yanı sıra iki Türkçe derlemin sonuçları karşılaştırıldığında, korelasyonlar genel olarak birbirine yakın olsa da, Wikipedia kullanılarak eğitilen vektörler ile hesaplanan korelasyon Boğaziçi derlemi ile eğitilen vektörlerle hesaplandan bir miktar daha yüksek çıkmıştır. Bu durum farklı dillerdeki derlemler arasındaki uyumun önemini göstermektedir.

III. ANLAMSAL İLİŞKİ TESTLERİ

Türkçe Wikipedia derleminin Boğaziçi derlemine kıyasla yüksek benzerlik sonuçları vermesinin, temsiller arasındaki kalite farkından kaynaklanabileceği düşünülmüştür. Bu düşüncüyü test etmek için çeşitli test kümeleri oluşturulmuştur. Testlerin çoğu İngilizce kelime temsillerinin testlerinde sıklıkla kullanılan, iki kelime arasındaki anlamsal ilişkinin verilen üçüncü bir kelime ile başka hangi kelime arasında olduğunu bulmaya yönelik testlerdir ve doğru cevaba kelime vektörleri üzerinde yapılan basit aritmetik operasyonlarla ulaşılmaya çalışılır (İspanya - Madrid + Ankara=? Türkiye). Bu formatta hazırlanan 7 farklı testin yanı sıra, verilen dördü anlamca ilişkili bir tanesi alakasız beş kelimeden alakasız olanı bulmaya yönelik bir test de hazırlanmıştır.

Derlem seçiminin dışında, çeşitli ön işlemler ile vektörlerin performanslarının artırılıp artırılamayacağı da test

TABLO IV: Farklı kaynak seçimi ve ön işlemler ile elde edilen derlemlerin benzetim ve farklı kelimeyi bulma testlerindeki performansları

Derlem	Bilindik Ülke-Başkent	Dünya Ülke-Başkent	Millet-Dil	Ülke-Millet	Ülke-Dil	Ülke-Para Birimi	Eş Anlamlılar	Farklı Bul
Wikipedia	82.39	64.07	66.19	65.12	94.29	1.54	16.04	81.22
eksiz Wikipedia	82.95	62.99	62.86	65.41	86.19	3.54	16.16	80.37
temiz Wikipedia	85.05	64.76	64.29	64.83	90.00	2.57	13.54	80.93
temiz eksiz Wikipedia	85.23	64.29	64.29	65.60	86.19	3.62	14.04	80.26
Boun Web	89.96	61.08	30.53	53.64	35.26	1.40	16.51	76.41
eksiz Boun Web	86.36	60.26	31.05	60.00	39.47	2.14	16.29	76.09
temiz Boun Web	89.20	61.48	26.32	57.68	33.16	1.33	15.38	75.59
temiz eksiz Boun Web	87.31	61.08	34.74	62.22	43.68	2.25	14.92	76.18

edilmiştir. Bu amaçla, derlemlerde NTLK tarafından Türkçe için belirlenmiş olan ve DDİ için gereksiz görülen kelimeler (stop words) atılarak temiz derlemler elde edilmiştir. Türkçenin sondan eklemeli yapısından dolayı kelimeler cümle içerisinde çok çeşitli ek ve ek gruplarıyla birleşerek bulunabilirler. Anlamlarının değişmemesine karşın farklı ekler alan aynı kelimeler, kelime temsilleri oluşturulurken farklı kelime olarak kullanılır. Bu durumun önüne geçebilmek için derlemlerdeki kelimeler Zemberek⁴ DDİ kütüphanesi kullanılarak eklerinden temizlenmiştir. Bu işlemler sonucu elde edilen sekiz farklı derlemden GloVe kullanılarak kelime vektörleri öğrenilmiş ve bu vektörler oluşturulan sekiz test kümesi ile test edilmiştir.

Tablo IV'te sunulan sonuçlara bakıldığında, ekleri veya gereksiz kelimeleri derlemlerden atmanın bazı testlerde performansı yükseltirken bazılarında düşürdüğü görülmektedir ve bu işlemlerin iki derlem üzerindeki etkilerinin de çok tutarlı olmadığı anlaşılmaktadır. Boğaziçi derlemi ile Wikipedia arasındaki belirgin performans farklılıkları Millet-Dil, Ülke-Millet ve Ülke-Dil testlerinde görülmektedir. Bu testlerde Wikipedia'nın daha iyi sonuç vermesinin sebebinin, içerdiği sayfalarda millet, dil ve ülke isimlerinin sık sık bir arada bulunuyor olması olabileceği değerlendirilmiştir.

IV. SONUÇLAR

Bu çalışmada Türkçe ile İngilizce, Almanca, Fransızca, İspanyolca ve Almanca arasındaki anlamsal benzerlik, kelime temsilleri yoluyla incelenmiştir. Dillerin anlamsal benzerliklerinin karşılaştırılabilmesi için bir sözcük listesi belirlenmiş ve incelenen diğer dillere tercüme edilmiştir. GloVe algoritması ve Wikipedialar kullanılarak kelime vektörleri öğrenilmiş, bu vektörlerden her dil için benzeşmezlik matrisleri oluşturulmuş ve bu matrisler arasındaki korelasyon hesaplanmıştır. Türkçe'nin anlamsal olarak bu beş Avrupa dilinden en çok İngilizceye (0.50), en az Almancaya (0.43) benzediği görülmüştür. Türkçe kelime vektörlerinin öğrenilmesi için Wikipedia'nın yanı sıra boyut olarak daha büyük olan Boğaziçi derlemi de kullanılmıştır. Türkçe Wikipedia'nın boyutça küçük olmasına karşın diğer diller için kullanılan veritabaları ile uyumlu olmasından dolayı nispeten daha yüksek korelasyonlar verdiği gözlemlenmiştir.

Aynı dilin farklı derlemlerinden öğrenilen vektörlerin performanslarını ölçmek için kelimeler arası anlamsal ilişkiye yönelik Türkçe test kümeleri hazırlanmıştır. Gereksiz kelimelerin temizlenmesi ve eklerin atılması işlemleriyle farklı derlemler

oluşturulmuş ve öğrenilen vektörler test edilmiştir. Elde edilen sonuçlara göre bu işlemlerin vektörlerin performansında tutarlı bir artış sağlamadığı anlaşılmıştır. Türkçe Wikipedia'nın özellikle belirli testlerde Boğaziçi derleminde daha iyi sonuç vermesi, test performansı için derlem içeriğinden ziyade içerik uygunluğunun önemli olduğunu göstermektedir.

KAYNAKLAR

- [1] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.
- [2] K. Zervanov, E. Iosif, and A. Potamianos, "Word semantic similarity for morphologically rich languages," in *LREC*, 2014, pp. 1642–1648.
- [3] J. Correia, E. Formisano, G. Valente, L. Hausfeld, B. Jansma, and M. Bonte, "Brain-based translation: fmri decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe," *The Journal of Neuroscience*, vol. 34, no. 1, pp. 332–338, 2014.
- [4] D. Ravichandran, P. Pantel, and E. Hovy, "Randomized algorithms and nlp: Using locality sensitive hash function for high speed noun clustering," in *Proc. of the 43rd Annual Meeting on Assoc. for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 622–629.
- [5] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proc. of the 24th International Conf. on Machine Learning*. ACM, 2007, pp. 641–648.
- [6] T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology," in *CoNLL*, 2013, pp. 104–113.
- [7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
- [9] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [10] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.
- [11] H. Sak, T. Güngör, and M. Saraçlar, "Resources for turkish morphological processing," *Language resources and evaluation*, vol. 45, no. 2, pp. 249–261, 2011.
- [12] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte, "A toolbox for representational similarity analysis," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003553, 2014.
- [13] N. Kriegeskorte and R. A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 401–412, 2013.

⁴<https://github.com/ahmetaa/zemberek-nlp>