Türkçe Sözcük Temsillerinde Dilbilimsel Özellikler
Linguistic Features in Turkish Word Representations

Onur Güngör
Huawei Türkiye Ar-Ge Merkezi & Boğaziçi Üniversitesi
İstanbul, Türkiye
Email: onur.gungor@huawei.com

Eray Yıldız
Huawei Türkiye Ar-Ge Merkezi & İstanbul Teknik Üniversitesi
İstanbul, Türkiye
Email: eray.yildiz@huawei.com

***Özetçe* —Eğiticisiz öğrenme yöntemleriyle elde edilen dağıtılmış sözcük temsillerinin birçok Doğal Dil İşleme (DDİ) uygulamasında başarıyı artırdığı, hatta birçok dil için bilinen en iyi sonuçların güncellenmesine yol açtıkları bilinmektedir. Bu temsillerin hem anlambilimsel hem sözdizimsel hem de biçimbilimsel bilgileri içerdiklerine dair çalışmalar bulunmaktadır. Bu alandaki ilk çalışmalardan sonra, biçimbilimsel bilginin daha iyi temsil edilmesine yardımcı olacak temsiller de tasarlanmıştır. Ancak içerilen biçimbilimsel bilginin özelliklerini araştırmak için yapılan çalışmalar zengin biçimbilimsel özelliklere sahip Türkçe gibi diller açısından sınırlı kalmıştır. Bu çalışmada amacımız atlagram yöntemi ile öğrenilen Türkçe sözcük temsillerinin sözdizimsel ve biçimbilimsel seviyelerde taşıdığı bilgileri araştırmaktır. Bunun için kök sözcük ile bunların çekimli ve türetilmiş hallerinden oluşan ikililer otomatik olarak üretilmiş, benzeşim yöntemiyle bu ikililer arasındaki ilişkiler incelenmiştir. Sonuç olarak, sözcüklerin sadece yüzeysel biçimi kullanılarak çalıştırılan atlagram algoritması ile üretilen temsillerin, Türkçe'deki farklı çekim ve yapım eki grupları açısından farklı temsil yeteneklerine sahip olduğu görülmüştür. Çalışmada oluşturulan deney kümeleri ve öğrenilmiş sözcük temsilleri ilerideki çalışmalarda kullanılması amacıyla paylaşılmıştır.**

Anahtar Kelimeler—dağıtılmış sözcük temsilleri, word2vec, atlagram, Türkçe, doğal dil işleme, eğiticisiz öğrenme dil modelleri, biçimbilim

***Abstract*—Distributed word representations which are learned using unsupervised methods are employed in many Natural Language Processing (NLP) tasks. They have led to state-of-the-art results in many NLP tasks for many languages. There have been studies reporting that word representations include morphological and semantical information. There are also work that aim to propose word representations which handle the morphological and syntactical information better. However, studies that evaluate the quality of the word representations for morphologically rich languages like Turkish are limited. In this study, we aim to explore the syntactic and morphological information captured by the Turkish word representations which are learned using skip-gram method on a large corpus. To assess the quality of information found in relations between Turkish word embeddings, analogical reasoning task is performed using couples consisting of root words and their inflected or derivative forms. We contribute with detailed experiments and show that word embeddings trained with skip-gram method have differing capabilities in capturing information for inflection and derivation groups in Turkish. We make the test sets and word embeddings publicly available to other researchers for further research.**

Keywords—distributed word representations, word2vec, skip-gram, Turkish, natural language processing, unsupervised learning

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

I. GİRİŞ

Metinlerin nasıl temsil edileceği birçok DDİ görevinde (örn. metin sınıflandırma, metin özetleme) temel sorunların başında gelmektedir [1]. Geleneksel yaklaşımda metinler, içerisinde geçen sözcüklerin ve bu sözcüklerden elde edilen çeşitli özelliklerin (örn. sözcük kökleri, n-gramlar ve sözcük türleri) bulunma sayıları ile gösterilmektedir. Bu temsil etme yöntemi ile bir metin, tüm metinlerdeki farklı sözcük veya sözcüğe dayalı özelliklerin sayısı boyutunda bir vektörle gösterilmektedir [2]. Bu metin vektörleri literatürde farklı yöntemlerle elde edilmektedir (örn. Bag-of-words, TF-IDF). Bu metin temsil yöntemlerinin en önemli dezavantajı sözcükler arasındaki anlamsal ve sözdizimsel yakınlıkları içermemesidir. Nispeten daha yeni olan çalışmalarda metinleri anlamsal bir uzayda temsil etmek için sözcüklerin birlikte geçtikleri doküman bilgilerinden faydalanan saklı anlam indeksleme [3] ve sınıf bilgilerinden ve kümeleme yöntemlerinden faydalanan sınıf bilgisiyle kümeleme [4] gibi çeşitli yöntemler önerilmiştir.

Son zamanlarda gerçekleşen makine öğrenmesi konusundaki önemli ilerlemeler sonucunda çok daha büyük veri kümelerini kullanarak daha karmaşık modellerin eğitilmesi mümkün hale gelmiştir. Bu gelişmeler önemli değişikliklere yol açmış ve ilk olarak Hinton ve arkadaşları [5] tarafından önerilen dağıtılmış sözcük temsilleri DDİ alanında giderek popülerleşmiştir. Sözcüklerin düşük boyutlu sürekli vektörlerden oluşan temsillerini eğiticisiz bir şekilde öğrenen yapay sinir ağı tabanlı yöntemler istatiksel dil modellemede oldukça başarılı sonuçlar vermiştir [6]. Daha sonra birçok DDİ görevinde kullanılan dağıtılmış sözcük temsilleri bu alanlarda en iyi sonuçların alınmasına yol açmıştır [7], [8]. Yapay sinir ağı tabanlı atla-gram algoritmasıyla üretilen İngilizce sözcük temsilleri ise [9] diğer yöntemlerle oluşturulan sözcük temsillerinden daha başarılı sonuçlar elde edilmesini mümkün kılmıştır. Bu sözcük temsil üretim yöntemi Türkçe'nin biçimbilimsel özelliklerini hesaba katmadığı halde Türkçe DDİ çalışmalarında da kullanılmış ve başarının artmasını sağladığı görülmüştür [10], [11].

Bunlarla birlikte, sözcüklerin biçimbilimsel özelliklerini daha iyi temsil edebilmek için tasarlanan sözcük temsilleri üzerine de çalışılmıştır. Örneğin, Luong ve arkadaşları [12] sözcük içinde tespit ettikleri ön-ek, gövde ve son-eklerin sıralamasına, cümle bağlamında ise bir dil mode-

line dayanan bir sözcük temsil modeli üzerine çalışmıştır. Güncel bir çalışma kapsamında, sözcüklerin biçimbilimsel çözümleyici ile işlendiği Luong'un çalışmasından farklı olarak, sözcük temsilleri eklerin eğiticisiz bir şekilde tespit edildiği halde öğrenilmiştir [13]. İngilizce DDİ alanında yaşanan bu gelişmelere karşın, biçimbilimsel olarak çok daha karmaşık olan Türkçe diline özgü dağıtılmış sözcük temsilleri konusunda yapılan çalışmalar oldukça sınırlıdır. Şen ve Erdoğan [14] tarafından yapılan çalışmada, atla-gram yöntemini kullanarak Türkçe için sözcük temsilleri öğrenilmiş, anlamsal ve sözdizimsel seviyelerde yapılan deneylerde İngilizce için önerilen deney kümelerinden [9] esinlenilmiştir. Anlamsal seviyedeki bilgiler "devlet - başkent", "il - ilçe" gibi ilişkiler taşıyan kelime çiftleri ile sorgulanırken, sözdizimsel seviyedeki bilgiler ise sadece çoğul eki, olumsuzluk eki, geniş zaman eki ve geçmiş zaman eki kullanılarak üretilen kök ve çekimli sözcük çiftleri ile sorgulanmıştır. Türkçe, İngilizce'ye kıyasla biçimbilimsel açıdan çok daha zengin bir dildir. Teorik olarak Türkçe sözcükler sınırsız sayıda ek alabilmekte ve aynı sözcük köküne yapım ve çekim ekleri eklenerek sınırsız sayıda yeni sözcük türetilebilmektedir [15]. Bu sebeple Türkçe sözcüklerin biçimbilimsel seviyede taşıdığı bilgiler İngilizce'ye göre çok daha fazladır. Sözcüklerin yüzeysel biçimiyle kullanılan atla-gram algoritmasıyla öğrenilen sözcük temsillerinin sözcüklerin taşıdığı bu tür bilgileri ne kadar yansıtabildiği Türkçe DDİ çalışmaları açısından önemli bir merak konusudur. İngilizce için önerilmiş olan deney kümeleri üzerindeki başarı bu merakı gidermekte yetersiz kalmaktadır. Şen ve Erdoğan'ın çalışmasında [14] anlamsal bilgiyi ölçmek için kullanılan deney kümeleri eksiz halde bulunan özel isimler iken, sözdizimsel bilgiyi ölçmek için kullanılan deney kümesi ise yalnızca dört farklı ek kullanılarak üretilen çekimli sözcüklerden oluşmuştur. Bizim çalışmamızdaki temel motivasyon, zengin biçimbilimsel yapıya sahip ve İngilizce diline kıyasla DDİ açısından daha fazla zorluk barındıran Türkçe dilinde atla-gram algoritması ile öğrenilen sözcük temsillerinin içerdiği bilgilerin biçimbilimsel ve sözdizimsel açıdan tutarlılığını daha geniş olarak incelemektir. Bunun için önce Türkçe için şu ana kadar farklı kaynaklardan toplanılmış en büyük derlem [11] kullanılarak atla-gram algoritması ile Türkçe sözcüklerin vektör temsilleri elde edilmiştir (Bölüm III-A). Bu temsillerin Türkçe'deki sözdizimsel ve biçimbilimsel özellikleri temsil etme yeteneği farklı köklere çeşitli çekim ve yapım ekleri eklenerek üretilen deney kümeleri kullanılarak benzeşim yöntemiyle [16] ölçülmüştür (Bölüm III-B). Elde edilen sonuçlar çekim eklerinin sözcüğe kattığı dilbilimsel özelliklerin oldukça iyi öğrenilebildiğini gösterirken, buna karşılık yapım eklerinin etkisinin göreli olarak daha kötü öğrenildiği gözlemlenmiştir (Bölüm IV).

II. Dağıtılmış Sözcük Temsilleri

Bu çalışmada bahsi geçen sözcük temsilleri atla-gram olarak adlandırılan yöntem kullanılarak üretilmiştir [16]. Atla-gram yöntemi bir metinde bir sözcüğün çevresindeki sözcükleri tahmin etmeye çalışarak sözcük temsillerini bulmaya çalışır. Formüllerle ifade edersek, derlemdeki tüm w_t sözcükleri için $\sum_{t=1}^{T}\sum_{-c\leq j<c, j\neq 0} \log P(w_{t+j}|w_t)$ maliyet formülü eniyileştirilmeye çalışılmaktadır. Görüldüğü üzere bu değerleri hesaplamak için her sözcüğün etrafındaki $2c$ sözcük değerlendirilir. Model, bu formüldeki $P(w_{t+j}|w_t)$ olasılığını $\exp(V_{w_t}^T U_{w_{t+j}})/\sum_{w\in W}\exp(V_{w_t}^T U_{w_{t+j}})$ olarak alır. Bu formülde W sözlükteki tüm sözcüklerin kümesini, V_w ve U_w ise $\mathbb{R}^d$ içindeki bu sözcüklere denk gelen vektörleri temsil ederler. Buradaki sorun bu formülün bölen kısmının hesaplanmasının bizim ilgilendiğimiz kadar büyük derlemler için çok zaman almasıdır. Bu yüzden diğer çalışmalarda da [16] yapıldığı gibi $P(w_{t+j}|w_t)$ yerine

$$\log\sigma(V_{w_t}^T U_{w_{t+j}}) + \sum_{1}^{g} \mathbb{E}_{w_i\sim P_{neg}(w)}[\log\sigma(-V_{w_t}^T U_{w_i})]$$

yaklaşıklığını kullandık. Bu formülde $\sigma(x) = (1+e^{-x})^{-1}$ formülünü ifade eder. Çalışmamızda negatif örnek sayısını belirten g değerini 5 aldık.

III. Deneyler

A. Türkçe Sözcük Temsillerinin Öğrenilmesi

Türkçe sözcük temsillerini öğrenmek için bildiğimiz kadarıyla şu ana kadar toplanılan en büyük Türkçe derlem olan Yıldız ve arkadaşlarının [11] derlemini kullandık. Herhangi bir önişleme tabi tutmadan çalıştırdığımız atla-gram algoritması ile Türkçe sözcük vektörlerini elde ettik. Haber siteleri, romanlar ve çeşitli web sayfalarından elde edilen ve toplam 941 milyon sözcük içeren bu derlemden yaklaşık 2 milyon adet farklı sözcük için temsil öğrenilmiştir. Atla-gram algoritması için `gensim` kütüphanesi[1] kullanılmıştır. Eğitim sırasındaki pencere boyutu 5, sözcük vektörlerinin boyutu ise 300 olarak ayarlanmıştır. Derlemde 10 defadan az görülen sözcükler elenmiş ve yanlış örnekler üretmek için negatif örnekleme [16] metodundan faydalanılmıştır. Öğrenilen sözcük temsilleri ve derlem bu konuda yapılacak gelecek çalışmalar düşünülerek kamu kullanımına açılmıştır[2].

B. Türkçe Deney Kümelerinin oluşturulması

İngilizce sözcüklerin temsillerinin sunulduğu çalışmada [16], sözcük temsillerinin kalitesini değerlendirmek için benzeşim yöntemi kullanılmıştır. İngilizce dili için oluşturulan deney kümesi aralarında belirli bir ilişki olan sözcük çiftleri (örn. "Berlin - Germany") ile aynı ilişkiyi içeren diğer çiftlerin eşleşmelerinden oluşmaktadır. Sözcük temsillerinin taşıdığı bilginin değerlendirilmesi için, bu eşleşleşmelerdeki bir çiftin bir elemanı dışarıda tutularak bu elemanı diğer üç elemanın sözcük temsilleriyle aritmerik işlemler yapılarak tahmin edilmeye çalışılır. Örneğin "Berlin - Germany <=> Paris - France" dörtlüsünde "France" dışarıda bırakılırsa, Germany'nin temsilinden Berlin'in temsili çıkartılarak "başkentlik" ilişkisi bulunur ve bu fark Paris'ın temsiline eklenerek ulaşılan vektöre en yakın sözcük temsilinin "France" olması beklenir. Vektörler arasındaki uzaklık kosinüs benzerliği ile ölçülmektedir. Bu deney kümesinde anlamsal seviyede taşınan bilginin ölçümü için devlet - başkent, şehir - eyalet, devlet - para birimi ilişkileri ve aile ilişkileri ile ilgili örnekler yeralmaktadır. Sözdizimsel seviye için ise "quick - quickly <=> slow - slowly" gibi örnekler bulunmaktadır. Toplam örnek sayısı 19,559'dur. Benzer yaklaşımla Türkçe için yapılan çalışmada [14] ise bu deney kümeleri Türkçeleştirilmiş ve atla-gram algoritmasının Türkçe dilinde de başarılı sonuçlar verdiği gözlemlenmiştir. Bu çalışmada ise Türkçe'nin zengin biçimbilimsel yapısı dikkate alınarak oluşturulan deney kümeleri üzerinde benzeşim yöntemi

[1] https://radimrehurek.com/gensim/l

[2] https://github.com/onurgu/linguistic-features-in-turkish-word-representations

TABLO I: Deney kümelerinden örnekler

Deney Kümesi	Örnekler
Çekim eki	güç - gücün <=> üye - üyenin yol - yolumuz <=> sahip - sahibimiz bilir - bilse <=> açıklar - açıklasa kullanır - kullanmalı <=> hazırlar - hazırlamalı
Yapım eki	yer - yersiz <=> komisyon - komisyonsuz eğitim - eğitimci <=> saat - saatçi devlet - devletimsi <=> insan - insanımsı
Hem yapım hem de çekim eki	yapmak - yapamayacaksam <=> gitmek - gidemeyeceksem ağaç - ağacıyım <=> araba - arabasıyım dolap - dolaplarındaki <=> araba - arabalarındaki

ile deneyler yapılmış ve sözcük temsillerinin biçimbilimsel ve sözdizimsel seviyelerdeki taşıdığı bilgiler değerlendirilmiştir.

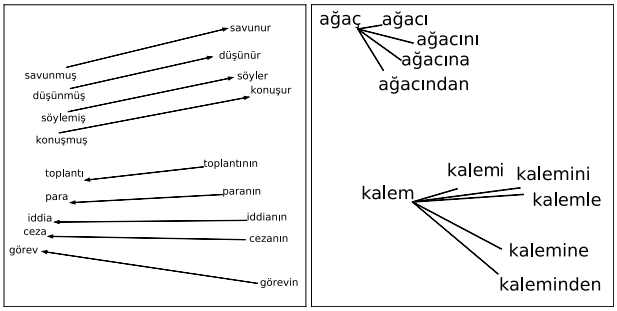
Sözcük temsillerinin taşıdığı sözdizimsel bilgi, bazı sözcüklere çeşitli ekler ekleyerek oluşturulan deney kümeleri kullanılarak değerlendirilmektedir [16] (örn. "slow - slowly <=> quick - quickly"). Türkçe'de ek sayısı oldukça yüksek olduğundan ve ekler sözcüğe eklenirken birçok ses olayı meydana geldiğinden Türkçe için sözdizimi ölçen deney kümesinin oluşturulması İngilizce'ye göre daha zordur. Ayrıca Türkçe'nin zengin biçimbilisel yapısı sebebiyle biçimbilimsel ve sözdizimsel özellikler iç içe geçmiştir. Bu sebepten biçimbilimsel ve sözdizimsel değerlendirme için aynı deney kümesi kullanılmıştır. Bu deney kümesini elde etmek için *i*) biçimbilimsel belirsizlik giderme problemi için hazırlanmış, bir milyon Türkçe sözcük ve biçimbilimsel çözümlemelerini içeren veri kümesindeki [15] isim ve fiil köklerinden türetilmiş sözcüklerin tüm biçimbilimsel çözümlemeleri tekilleştirilerek 60,745 biçimbilimsel etiket elde edilmiştir. *ii*) Elde edilen biçimbilimsel etiketlerden sadece bir çekim eki içeren, sadece bir yapım eki içeren ve birer adet hem yapım eki hem de çekim eki içeren olmak üzere üç adet küme oluşturulmuştur. *iii*) Türkçe'de yaygın kullanılan 25 adet fiil kök ve 25 adet isim kök rastgele seçilmiş ve biçimbilimsel etiketler arasından rastgele seçilen etiketler ile birleştirilerek yeni çözümlemeler türetilmiştir. *iv*) Sonlu durum dönüştürücüsü tabanlı bir biçimbilimsel çözümleyici [17] kullanılarak derlemdeki sözcüklerin çözümlemeleri bulunmuş ve yukardaki adımlar sonucunda üretilen biçimbilimsel çözümlemelerden yüzeysel biçimdeki sözcükler elde edilmesinde kullanılmıştır.

Tablo I'de üretilen deney kümelerinden çeşitli örnekler gösterilirken, Şekil 1'de t-SNE [18] algoritması ile iki boyutlu uzaya indirgenen sözcük vektörleri arasındaki ilişkiler resmedilmiştir. Deneylerde kullanılmak üzere oluşturulan bu veri kümeleri ticari olmayan araştırma projeleri ve akademik çalışmalar için kamuya açık olarak paylaşılmıştır.[3]

C. Deneysel Sonuçlar

Yapılan deneylerde, her bir deney kümesindeki dörtlü sözcük gruplarının ilk üçünün temsili kullanılarak Bölüm III-B'de tarif edilen aritmetik işlemler ile 4. sözcüğü bulma başarısı ölçülmüştür. Tablo II ve III'de deney sonuçları gösterilmiştir. Bu tablolarda deney kümeleri biçimbilimsel çözümleyiciden elde edilen veriler ışığında kategorilere ayrılmıştır. Bunlardan birkaçı ismin "bulunma hali", "yönelme hali" ve fiiller için "geçmiş zaman" ve "üçüncü tekil şahıs" ekleri

[3]https://github.com/onurgu/linguistic-features-in-turkish-word-representations



Şekil 1: Deney Kümelerinden Seçilen Sözcüklerin Temsillerinin t-SNE Algoritması ile İki Boyutlu Uzayda Gösterimi

TABLO II: Çekim ekleri ile yapılan deney sonuçları

Biçimbilimsel Kategoriler	Örnek Sayısı	MRR	Biçimbilimsel Kategoriler	Örnek Sayısı	MRR
Tamlayan eki	1176	0.488	Şimdiki z. eki (-yor)	861	0.847
3. tekil k. iyelik eki	1176	0.448	-mişli geçmiş z. eki	1176	0.774
İsmin -i hali	1176	0.439	Gelecek zaman eki	1176	0.707
2. tekil k. iyelik eki	1176	0.321	Olumsuzluk eki	1176	0.67
Vasıta eki (-le)	1128	0.253	Şimdiki z. eki (-mekte)	1176	0.646
1. çoğul k. iyelik eki	1176	0.25	Geçmiş z. eki	1176	0.535
İsmin -den hali	1128	0.238	İstek kipi	1176	0.534
İsmin -e hali	1176	0.236	Gereklilik kipi	1128	0.467
2. çoğul k. iyelik eki	1081	0.196	Emir kipi	1176	0.465
3. çoğul k. iyelik eki	1176	0.178	3. çoğul kişi çekimi	1128	0.42
1. tekil k. iyelik eki	1128	0.148	2. tekil kişi çekimi	946	0.365
İsmin -de hali	1128	0.136	1. tekil kişi çekimi	990	0.339
Eşitlik eki (-ce)	276	0.027	2. çoğul kişi çekimi	1035	0.327
İsim çekim ekleri toplam	14101	0.273	**Fiil çekim ekleri toplam**	14320	0.547
			Çekim ekleri toplam	28421	0.411

eklenerek oluşturulmuş dörtlü sözcük gruplarıdır. Başarıyı ölçmek için $^1/_{|Q|}\sum_{i=1}^{|Q|} {}^1/_{S_i}$ formulü ile hesaplanan MRR (Mean Reciprocal Rank - Ters Sıralamaların Ortalaması) metriği kullanılmıştır. Bu formülde bir kümedeki örnek sayısı Q ile gösterilirken, S_i doğru yanıtın en yakın sözcükler listesindeki sırasını göstermektedir. Bu metrik, olası yanıtların bir listesini üreten sistemlerin başarılarının istatistiksel olarak ölçülmesinde sıklıkla kullanılan bir metriktir [19].

Fiiller için öğrenilen temsiller isimler için öğrenilen temsillerle karşılaştırıldığında, fiillere eklenen çekim ekleri sonucunda oluşan sözcüklerle kök fiil arasındaki ilişkinin, isimlerdeki benzer ilişkilere göre daha iyi öğrenilmiş olduğu görülmektedir. Sadece fiil gruplarına bakıldığında, zaman eklerinin kök ve çekimli fiil arasında oluşturduğu ilişkinin diğer çekim eklerine göre daha iyi yakalandığı görülmüştür. Bunun altında yatan neden üzerine fikir yürütmek için derlemde bulunan sözcük istatistiklerine dayanan bir araştırma yapılmıştır.

TABLO III: Yapım ekleri ile yapılan deney sonuçları

Biçimbilimsel Kategoriler	Örnek Sayısı	MRR
-lık eki	406	0.023
Yokluk eki (Without) [-sız, -siz]	946	0.013
Birliktelik eki [-le, la]	1128	0.011
Küçültme eki (Diminutive) [-cik, -cık]	45	0.01
Meslek eki [-cı, -ci]	378	0.009
Yapım ekleri toplam	2906	0.012
Hem yapım eki hem de çekim eki	125	0.14

Bu araştırmada, fiil gruplarında geçen sözcüklerin derlemde bulunma sayılarının ortalamaları, isim gruplarındaki sözcüklerin ortalamalarına göre daha az çıkmıştır. Buna benzer olarak bulunma sayıları sıklık aralıklarına dağıtıldığında fiil grupları daha dengeli bir dağılım izlerken, isim gruplarındaki sözcüklerin genelde en yüksek sıklık aralığına toplandığı gözlenmiştir. Bu bulgular, seyrek geçen sözcüklerin daha kötü temsil edildiğine dair hipotezi güçsüzleştiriyor gibi gözükse de, en başta isim ve fiil gruplarındaki sözcüklerin seçilmesi sırasında sıklıkların eşit dağılması gözetilmediği için yanıltıcı olabilir.

Yapım ekleri ile üretilen deney kümelerinde ise MRR değerlerinde önemli ölçüde düşüş gözlenmiştir (Tablo III). Bunun sebebi, sözcük köklerine eklenen aynı yapım eklerinin sözcüğe çok farklı anlamlar katabilmesidir. Örneğin "-li" ekinin "İzmirli" ve "gizli" sözcüklerine kattıkları anlamlar birbirinden oldukça farklıdır. Bu durum "-li" eki almış olan sözcüklerin farklı bağlamlarda kullanılmalarına yol açmaktadır. Dolayısıyla "-li" ekinin sözcüğe taşıdığı anlam farklılaşarak belirsizleşmektedir. Bunun yanında, bazı durumlarda yapım ekleri kök ile türetilmiş sözcük arasında da büyük bir anlam farkı yaratmaktadırlar. Dolayısıyla, yapım eklerinden oluşan deney kümelerindeki düşük MRR değerlerine dayanarak türetilmiş sözcüklerin kötü temsil edildiğini söylemek yanlış olur. Çıkarılacak doğru sonuç, deney kümelerindeki kök ve türetilmiş sözcükler arasındaki ilişkilerin yeterince öğrenilememiş olduğudur. Hem yapım eki hem de çekim eki alan sözcüklerden üretilen deney kümesinde ise başarı sadece yapım eki alan sözcüklerdeki gibi düşüktür.

IV. Sonuçlar ve Tartışma

Türkçe sondan eklemeli ve biçimbilimsel olarak oldukça zengin bir dildir. Bu sebepten Türkçe sözcükler aldıkları çekim ve yapım eklerinin katkısıyla farklı DDİ seviyelerinde oldukça fazla bilgi içerebilmektedir. Bu durum DDİ görevleri açısından birçok zorluğu beraberinde getirmektedir. Diğer taraftan İngilizce başta olmak üzere birçok dil üzerinde sözcüklerin vektörel temsillerinin anlamsal ve sözdizimsel seviyede karmaşık bilgiler taşıdığı literatürde gösterilmiş ve bu sözcük temsillerinin kullanımıyla birçok DDİ görevinde en iyi sonuçlar alınmıştır. Bu çalışmada, Türkçe sözcüklerin zengin biçimbilimsel yapılarından dolayı edindikleri anlamsal veya sözdizimsel bilgilerin dağıtılmış sözcük temsilleri uzayında nasıl temsil edildikleri benzeşim yöntemi vasıtasıyla araştırılmıştır. Belirtmek gerekir ki, sözcük temsillerinin kalitesinin ölçümü konusunda literatürde çeşitli tartışmalar yürütülmüştür [20].

Sonuçlar göstermiştir ki, Türkçe gibi biçimbilimsel açıdan zengin bir dilde dahi sözcükler herhangi bir önişleme tabi tutulmadan sadece yüzeysel biçimleri kullanılarak elde edilen sözcük temsilleri, biçimbilimsel ve sözdizimsel seviyelerde önemli bilgiler içermektedir. Sadece çekim eki alan sözcüklerden oluşan deney kümesinde daha yüksek MRR değerleri elde edilirken yapım ekleriyle elde edilen deney kümelerinde MRR değerlerinin önemli ölçüde azaldığı gözlemlenmiştir. Bu sonuçlara göre Türkçe sözcüklerin aldığı ekler arttıkça bu eklerin sözcüğe kattığı bilgilerin temsil edilme kabiliyetinin azaldığı anlaşılmaktadır. Bu durum Türkçe'nin zengin biçimbilimsel yapısını da dikkate alan sözcük temsili üretme yöntemleri üzerine çalışma yapılması ihtiyacını da göstermektedir. İlk adım olarak, bu çalışmadaki deneyleri sözcüklerdeki biçimbilimsel bilgileri iyi temsil edebilmek üzere özellikle tasarlanmış sözcük temsilleri [12], [13] için de uygulayıp aralarında bir karşılaştırma yapmak mümkündür.

Kaynakça

[1] M. F. Amasyalı, M. Çetin, and C. Akbulut, "Metinlerin anlamsal uzaydaki temsil töntemlerinin sınıflandırma performansına etkileri," *Sigma*, vol. 5, pp. 8–14, 2013.

[2] L. Ciya, A. Shamim, and D. Paul, "Feature preparation in text categorization," *Oracle Text Selected Papers and Presentations*, 2001.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[4] M. F. Amasyalı, S. Balcı, E. Mete, and E. N. Varlı, "Türkçe metinlerin sınıflandırılmasında metin temsil yöntemlerinin performans karşılaştırılması/a comparison of text representation methods for turkish text classification," *EMO Bilimsel Dergi*, vol. 2, no. 4, 2012.

[5] G. E. Hinton, J. L. Mcclelland, and D. E. Rumelhart, "Distributed representations, parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations," 1986.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[8] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 129–136.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[10] H. Demir and A. Özgür, "Improving named entity recognition for morphologically rich languages using word embeddings," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 117–122.

[11] E. Yildiz, C. Tirkaz, B. Sahin, M. T. Eren, and O. Sonmez, "A morphology-aware network for morphological disambiguation," 2016.

[12] T. Luong, R. Socher, and C. D. Manning, "Better word representations with recursive neural networks for morphology," in *CoNLL*, 2013.

[13] K. Cao and M. Rei, "A joint model for word embedding and word morphology," *CoRR*, vol. abs/1606.02601, 2016.

[14] M. U. Şen and H. Erdoğan, "Learning word representations for turkish (türkçe için kelime temsillerinin öğrenimi)," 2014.

[15] D. Yuret and F. Türe, "Learning morphological disambiguation rules for turkish," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 328–334.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] K. Oflazer, "Two-level description of turkish morphology," *Literary and linguistic computing*, vol. 9, no. 2, pp. 137–148, 1994.

[18] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[19] E. M. Voorhees *et al.*, "The TREC-8 question answering track report." in *Trec*, vol. 99, 1999, pp. 77–82.

[20] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, "Problems with evaluation of word embeddings using word similarity tasks," *arXiv preprint arXiv:1605.02276*, 2016.