VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Otomatik Konuşma Tanıma Sistemlerinde Kullanılan Gerçek Metin Verisinde Biçimbilimsel-Sözdizimsel Hataların Tespiti ve Düzeltilmesi

Hüseyin POLAT[1], Hayri SEVER[2], Saadin OYUCU[3]*, Şükran TEKBAŞ[4]

[1-3]*Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, TÜRKİYE*
[2]*Çankaya Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, TÜRKİYE*
[4]*EMFA Yazılım Danışmanlık A.Ş., Ankara, TÜRKİYE*

Makale Bilgisi

Başvuru:
01/11/2019
Kabul:
13/12/2019

Özet

Türkçe Otomatik Konuşma Tanıma (ASR: Automatic Speech Recognition) sistemlerinde kullanılan akustik model gürbüz bir dil modeli ile desteklenmediği durumlarda kelime hata oranı yüksek olmaktadır. İyi dizayn edilmiş bir dil modeli ile akustik modelin birlikte ASR'de kullanılması kelime hata oranını düşürmektedir. ASR için gerekli dil modelinin eğitiminde düz metin verisi kullanılmaktadır. Kullanılan metin verisinin doğruluğu ASR modellerinin eğitimi için oldukça önemlidir. Bu çalışmada, doğal dil işlemeye dayalı bir yöntem kullanılarak Türkçe ASR sisteminin eğitilmesinde kullanılan metin verisi içerisindeki yazım hatalarının tespiti ve düzeltilmesi gerçekleştirilmiştir. Öncelikle metin verisi içerisinde dil bilgisi olarak yanlış yazılmış olan kelimeler bulunmuştur. Bir kelimedeki karakter eksikliği, karakter fazlalığı, karakterlerin yer değiştirmesi veya karakteri yanlış yazılmış olan kelimeler hatalı olarak kabul edilmiştir. Metin verisi içerisinde hatalı olarak kabul edilen kelimeler morfolojik analiz ile tespit edilmiştir. Yanlış kelimelerin yerine atanacak olan kelimeler belirlenmiştir. Yanlış yazılmış olan kelimeler doğru kelimeler ile değiştirilmiştir. Gerçekleştirilen çalışma hatalı kelimeleri tespit etme ve doğru kelimeler ile yer değiştirme işleminde %93 oranında başarı göstermiştir.

Anahtar Kelimeler: *Konuşma Tanıma, Doğal Dil İşleme, Düz Metin Hataları, Gramatik Kelime Hatası*

Detecting and Correcting Morpho-Syntactic Errors Used in Automatic Speech Recognition Systems Real Texts Data

Abstract

Word error rate is high when the acoustic model used in Turkish Automatic Speech Recognition (ASR) systems is not supported by a robust language model. The use of a well-designed language model in combination with an acoustic model in the ASR reduces the word error rate. In the training of the language model required for ASR, plain text data is used. The accuracy of the text data used is very important for the training of ASR models. In this study, the detection and correction of spelling errors in the text data used in the education of Turkish ASR system were realized by using a method based on natural language processing. Firstly, misspelled words were found in the text data. Lack of characters, surplus characters,

* İletişim e-posta: saadinoyucu@gazi.edu.tr
** Bu çalışmanın bir kısmı ICONDATA 2019 konferansında sözlü olarak sunulmuştur.

displacement of characters or misspelled words in a word are considered incorrect. The words which were accepted as wrong in the text data were determined by morphological analysis. The words to be assigned in place of the wrong words have been determined. The misspelled words have been replaced with the correct words. The study performed showed 93% success in detecting the wrong words and replacing them with the right words.

Keywords: *Speech Recognition, Natural Language Processing, Plain Text Error, Grammatical Word Error*

1 Giriş

İnsanların makine veya elektronik cihazlarla ana dillerini konuşarak iletişim kurması, makina veya elektronik cihazların komut ve kontrol mekanizmasının daha hızlı bir şekilde çalıştırabilmesine olanak vermektedir. Bu nedenle Automatic Speech Recognition (ASR) sistemleri geliştirilmekte ve kullanım alanı her geçen gün artmaktadır. ASR, insanlar tarafından konuşulan kelimeleri bilgisayar tarafından okunabilir metne dönüştüren bir teknolojidir [1]. Konuşma tanıma sisteminin görevi konuşma sinyalini birtakım modeller kullanarak konuşma bilgisini işleyip metne dönüştürmektir. Bu dönüşüm sırasında gerekli olan ASR modelleri ile temelde iki farklı veri ile eğitilmektedir. Bu veriler akustik modelleme için kullanılan konuşma verisi ve akustik bilgi ile eşleştirilecek metin verisidir. Bazı çalışmalarda ise akustik bilgiden elde edilen akustik model, dil modeli ile desteklenmektedir. ASR için gerekli olan dil modeli, bir sistem dizisinde hangi kelimelerin olası bir kelime dizisi oluşturduğunu, hangi kelimelerin birlikte ortaya çıkacağını vermektedir [2]. Dil modelinin oluşturulması ve eğitilmesi işlemi için ise metin verisi kullanılmaktadır.

Dil modeli oluşturulurken metin verisi içerisinde yer alan kelime sıralamaları Markov varsayımları ile modellenmektedir. Böylelikle bir kelimeden sonra hangi kelimenin gelmesi gerektiği olasılıksal olarak modellenmektedir. Bu modellerin başarısı ASR sistemlerinin çıkışında elde edilen metin başarımını doğrudan etkilemektedir. Modellerin eğitimi için kullanılan veri setindeki metinlerde yer alabilecek hatalı kelimeler ASR sistemlerinin başarımını doğrudan etkilemektedir. Bu durum ASR çıkışında yanlış kelimelerin sunulmasına neden olmaktadır. Dolayısıyla ASR için gerekli olan akustik ve dil modelinin geliştirilmesinde kullanılan metin verilerindeki hataların düzeltilmesi gerekmektedir.

Bu çalışmada, Doğal Dil İşleme (NLP: Natural Language Processing) tekniği kullanılarak ASR için gerekli olan akustik ve dil modelini oluşturan metinler içerisindeki yazım hatalarının tespit edilerek düzeltilmesi gerçekleştirilmiştir. Metin içerisinde karakter eksikliği, karakter fazlalığı, karakterlerin yer değiştirmesi ve yanlış yazılmış karakterlere sahip olan kelimeler hatalı olarak kabul edilmiştir. Yazım yanlışı olan kelimeler metin içerisinde morfolojik analiz ile tespit edilmiş ve yanlış yazılmış kelimelerin düzeltilmesi gerçekleştirilmiştir.

2 İlgili çalışmalar

Düz metin içerisinde yazım denetimi yapan erken dönem çalışmalarda sadece yanlış yazılmış kelimelerin tespiti yapılmıştır. Yakın zamanda gerçekleştirilen çalışmalarda ise yazım hatası olan kelimeler yerine kelime önerisi yapabilen sistemler üzerine yoğunlaşılmıştır [1]. Ancak bu çalışmaların çok azı Türkçe üzerine gerçekleştirilmiştir. Bu durumun nedeni ise Türkçe'nin diğer dillere göre çok farklı yapısal kurallara ve sorunlara sahip olmasından kaynaklanmaktadır. Türkçe'de kelimelerin alabileceği eklerin yanlış sıralama ile köke eklenmesi ya da sesli uyumuna uygun olmayan eklerin kök kelimeye eklenmesi önemli sorunlar arasındadır [3].

Yazım denetimi üzerine Starlander ve Belis'in geliştirdiği ve Fransızca yazım denetimi gerçekleştiren FipsOrtho, Fransızca öğrenenlere yönelik olarak hazırlanmış bir çalışmadır. Bu çalışmada yanlış yazılmış kelimeleri bulmak ve düzeltmek için kullanılan yöntemler; alpha-code yöntemi, ses bilimsel inceleme yöntemi ve ad-hoc kuralları yöntemidir. Belirtilen yöntemler ile bulunan hatalı kelimeler arasında Levenshtein-Damerau düzeltme mesafesi hesaplaması yapılmış ve uzaklığı en küçük olan kelime ya da kelimeler doğru kelime olarak önerilmiştir [4].

Almanca için Gabriele Kodydek tarafından yapılmış olan kelime çözümleme çalışmasında kelimeler en küçük bileşenlerine parçalanmıştır. Atomlar Almancanın en küçük anlamlı birimini temsil etmektedir. Bu çalışma da kelime analiz işlemi iki ana bölümden oluşmaktadır. Bu bölümlerden ilki

atom tablosu ve özyineleyici çözümleme işlemidir. Özyineleyici çözümleme algoritması Almancanın kurallarına uygun olarak kelimeleri atomlarına ayırmaktadır. Girilen kelimenin atom tablosundaki tüm alt kelimeleri bulunmaya çalışılır ve Alman dilinin dilbilgisi kurallarına göre bulunan tüm atomları birleştirilir [5]. Böylelikle hatalı olabilecek kelimeler tespit edilerek atom tablosundaki doğru kelimeler ile değiştirilir.

S. Dembitz tarafından yapılan çalışmada Hırvat dili için yazım denetimi aracı geliştirilmiştir. Önerilen yöntem öğrenme algoritmasına dayalı bir yapıya sahiptir. Yanlış yazılmış olan kelimelerin değerlendirilmesinde bulanık mantık ve n-gram tabanlı olasılıksal yaklaşımları kullanmıştır [6-7]. Gerçekleştirilen benzer bir çalışmada da ise Hintçe Tamil dili yazım denetleyicisi geliştirilmiştir. Bu dil Hindistan'ın güneyinde kullanılan bir dildir ve biçim açısından zengin bir yapıya sahiptir. Denetleyiciye girilen metindeki kelimeler sözlükte bulunmaz ise hata düzeltme yöntemi uygulanmaktadır [8].

Çakmak ve Diri'nin Türkçe üzerine yapmış olduğu çalışmada maksimum entropi yaklaşımı ele alınarak makine öğrenmesine dayalı bir yöntem kullanılmıştır. Bu yöntemde olasılıkların p(a|b) dağılımı metin verilerinden hesaplanmaktadır. Her kelimemin olasılığı kayıt dağılım olasılığı kayıt edilmektedir. Karar listesine dayanan yöntemde her bir kelimenin olasılığı diğer bir kelimenin olasılığı ile hesaplanmaktadır. Bu hesaplama işlemlerinin ardından en düşük olasılığa sahip olan kelime hatalı olarak kabul edilmektedir. Hatalı olan kelimeyi düzeltmek için kelimenin yerine geçecek, en yüksek olasılığa sahip kelime önerilmektedir [9].

Türkçe metinlerdeki yazım hatalarının denetlenmesi ve yabancı kelimelerin bulunması için yapılan diğer bir çalışmada kelimelerin Türkçe ses bilgisi kurallarına uygun olup olmadığının tespiti yapılmıştır. Bu tespitin yapılabilmesi için ilk olarak heceleme algoritması yardımıyla kelimelerin hecelenebilir olup olmadığının denetimi yapılmıştır. Bu denetimi geçemeyen kelimeler Türkçe hece yapısına uygun olmadıkları için doğrudan elenmiştir [3].

Türkçe yazım hatalarının denetlenmesi üzerine Çakıroğlu ve Özyurt'un yaptığı çalışmada istatistiksel yöntemler kullanarak hata tespiti yapılmıştır. Türkçe metinler için otomatik yazım denetimi ve düzeltme yöntemi önerilmiştir. Kelime frekansları ve yazım hataları istatistiksel bir çalışma ile elde edilmiştir. Sistemin bütünü ve alt modülleri bu istatistiksel veriler aracılığıyla tasarlanmıştır. Sistemin performansı ve başarımı "Ms Word" kelime işlemci programı ile analiz edilmiş ve çalışma sonunda sistem hata düzeltmede % 71, doğru kelime önermede %98 başarı sağladığı görülmüştür [10].

Yazım hatalarının denetlenmesi ve düzeltilmesi konusunda yapılmış çalışmalara bakıldığında çoğunun Hint-Avrupa dil ailesi üzerine özellikle de İngilizce üzerinde gerçekleştirildiği görülmektedir. Ancak Ural-Altay dil ailesi üyesi olan Türkçe üzerinde yapılan çalışmaların da sayısı giderek artmaktadır. Türkçe sondan eklemeli bir yapıya sahip olduğu için yazım hatalarının tespiti oldukça zorlu bir görevdir. Bu çalışmanın kapsamı, ASR sistemi için gerekli olan iki farklı model için gerekli olan metin verisindeki yazım hatalarının tespiti, hataların yüksek başarımla düzeltilmesi ve ASR modellerinin eğitimi için verilerin hazır hale getirilmesidir.

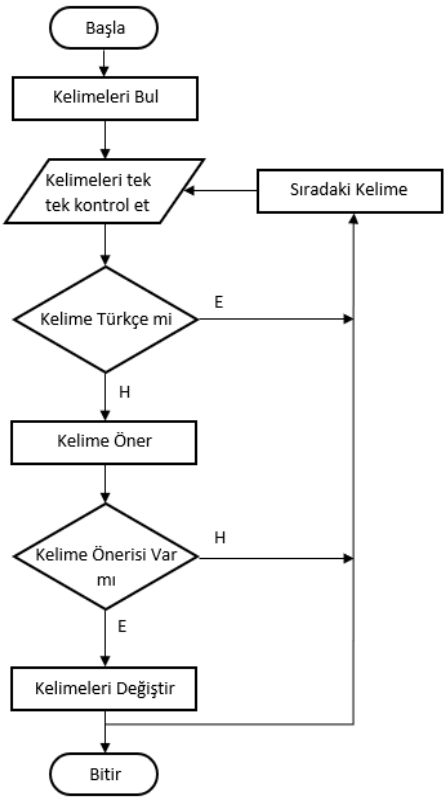
3 Hatalı yazılmış kelime tespiti ve kelime önerme

Klasik yöntemlerde kullanılan hatalı yazılmış kelime tespiti ve düzeltme yöntemleri için bir sözlük kullanılmaktadır. Genel olarak yazım denetimi yapılacak olan dile ait bütün kelimeler bu sözlüğe yerleştirilmektedir. Yazım denetimi yapılacak olan kelime sözlük içerisinde yer alıyorsa kelime doğru olarak kabul edilmektedir. Eğer kelime sözlükte yer almıyorsa yanlış olarak kabul edilmektedir.

Türkçe'nin sondan eklemeli bir yapıya sahip olması nedeniyle sadece sözlüğe dayalı bir yaklaşım tek başına yeterli değildir. Ancak Türkçe metinlerdeki yabancı kelimelerin tespit edilmesi için kelimelerin Türkçe ses bilgisi veya hece yapısı kurallarının denetlenmesi yeterlidir. Türkçede bulunan kelimeler ortalama olarak 3-5 ek almaktadır. Bu nedenle yazım hatalarının bulunabilmesi için morfolojik analiz gerekmektedir. Kelimedeki yazım hatasının kök kelimeden mi yoksa kök kelimeye eklenen eklerden mi kaynaklandığını anlamak için biçim bilimsel çözümleme yapılması gerekmektedir. Bu nedenle bu çalışma kapsamında kelimelerin morfolojik analizi gerçekleştirilmiş kök ve ekler birbirinden ayrılmıştır. Türkçe yazım denetimi için açık kaynak kodlu bir NLP Kütüphanesi olan Zemberek aracı kullanılmıştır. Morfolojik analizi yapılan kelimede hata bulunması

durumunda hatalı olan kök veya ek bilginin en doğru şekilde düzeltilmesi amaçlanmıştır.

Yazım hatalarının otomatik olarak tespiti için Java programlama dili kullanılarak bir uygulama geliştirilmiştir. Uygulama .txt uzantılı dosyaların içindeki kelimeleri bulup Türkçe olup olmadıklarını kontrol etmektedir. Kelimeler morfolojik analiz yöntemi ile hecelerine ayrılmıştır. Kelime analizinde kelimenin Türkçe olduğuna karar verildikten sonra sıradaki diğer kelime ele alınmıştır. Eğer kelime Türkçe değil ise kelime önerme işlemine gidilmekte ve hatalı olan kelimeye en yakın kelime önerisi sunulmaktadır. Öneri olarak sunulan kelime ile hatalı olan kelime yer değiştirme işlemine gönderilmektedir. Şekil 1'de geliştirilen uygulamanın çalışmasına ait akış şeması verilmiştir.

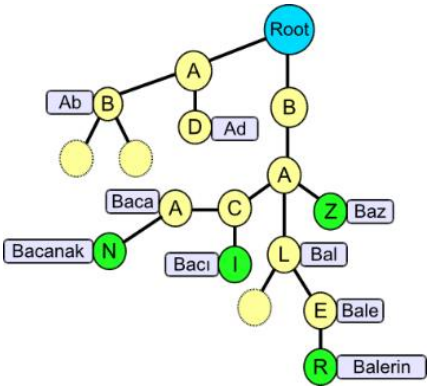


Şekil 1. Akış şeması

Bir kelimenin morfolojik analiz ile çözümlenebiliyor olması bu kelimenin Türkçe olması ile ilgili bilgi vermektedir. En çok kullanılan klasik yazım denetimi yöntemlerinde dile ait bütün kelimeler bir sözlüğe eklenir ve yeni gelen bir kelime sözlük içerisindeki kelimeler ile kıyaslanır. Eğer kelime bu sözlükte var ise kelime doğru kabul edilir. Ama bu yöntem ilk bakışta doğru gibi görünse de sözlükte bulunmayan ama o dile ait bir kelimeyi hatalı olarak kabul etmektedir. Bu tür yöntemlerin yüksek doğruluk oranı ile çalışması için kelime dosyasında milyarlarca kelimenin bulunması gerekmektedir.

Doğal dil işlemede kullanan uygulamalar verilen kelimeyi morfolojik analiz ile incelemektedir. İlk olarak kelimenin kökü olabilecek olan aday kelimeler belirlenmektedir. Belirlenen bu kelimelere uygun olabilecek ekler sırası ile kelime köklerine eklenmektedir. Kök ekleme işlemi sonucunda kelimenin aynısı elde edilmiş ise bu kelime Türkçedir, eğer kök adaylarının hiçbirinde aynı sonuç elde edilememişse o kelime Türkçe değildir sonucuna varılmaktadır.

Morfolojik analizde kök arama işlemi yapılabilmesi için Türkçede bulunan bütün kök kelimelerin bulunması gerekmektedir. Verilen bir kelimenin kök adaylarının bulunması için kökler bir ağaca yerleştirilir. Bu ağaçta kökler içeriklerine göre yerleşmektedir. Örneğin Şekil 2'de gösterilen örnekte "baz" kökü sırasıyla B- A -Z ile etiketlenmiş ve düğümlerin en sonuncusuna bağlanmıştır. Dikkati çeken bir nokta da uzun köklerin gereksiz ve fazladan düğüm oluşturmayacak şekilde ağaca bağlanmasıdır. Yani "Balerin" kökü B-A-L-E düğümlerinden sonra gelen R düğümüne bağlanmıştır.

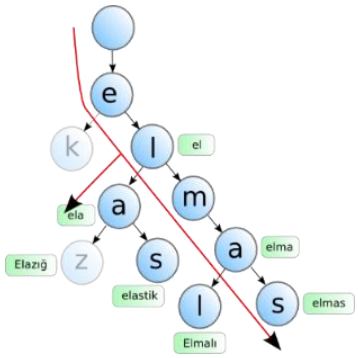


Şekil 2. Kelime kök ağacı [11]

Kök ağacı kök düğümleri ve bu düğümlerin bağlantılarından oluşmaktadır. Kitap (kitabı) gibi bir ek geldiği zaman yumuşama olasılığı olan köklerin değişmiş halleri de kök ağacına eklenmelidir. Ancak bu değişmiş biçimler orijinal kökü işaret edecek şekilde işlem gerçekleştirmelidir. Kelime kök sözlüğünde ilk aramada bulunamadı ise kelimede sondan itibaren ek araması yapılmaktadır. Örneğin "elmalar" kelimesi kök arama işleminde kök olarak bulunmamaktadır. Kök olarak bulunamayan kelime üzerinde ikinci aşama olan ek arama işlemi yapılmaktadır. Ek arama işleminden sonra "lar" ekine ulaşılır. Ek olarak belirlenen kısım kelimeden

çıkarılır ve kelime tekrardan kök arama işlemine tabi tutulur. Kök arama işleminde "elma" kelimesine ulaşılmış ve kelime bulunmuştur.

Kelime önerme işleminde ise nispeten karmaşık bir kök seçici ve özel bir çözümleyici kullanılmaktadır. Bu özel seçici ağaç üzerinde ilerlerken belli bir hata toleransını da hesaba katmaktadır. Bu hata toleransını kelimenin içerisinde bulunan kalın ünlüler ve ince ünlüler gibi benzerlikler oluşturmaktadır. Örneğin "kişi" kelimesi ile "kışı" kelimesi birbirine benzemektedir.



Şekil 3. Kelime önermede kullanılan kök ağaç şeması temsili [12]

Şekil 3'te gösterilen ağaç üzerinde "elxalarının" hatalı kelimesi incelenirken sadece "ela, elma ve elmas" değil "ela" kökü de seçilmektedir. Zemberek bu kelime için "elmaslarının ve elalarının" önerilerini üretmiştir. Ağaç üzerinde ilerleme hata toleransı ile sınırlıdır. Algoritma hata toleransının belli bir değerini aşıncaya kadar ağaçta ilerle ve rastladığın kökleri seç ifadesi ile özetlenebilir. Zemberek kelimelerin birbirlerine ne kadar benzediğini bulmak için Damerau-Levenshtein düzeltme mesafesi algoritmasını kullanmaktadır. Kelime önerme aşamasından sonra eğer bir kelime önerisi var ise hatalı olan kelime yerine önerilen ilk kelime seçilmiştir. Seçilen kelime ile hatalı olan kelime yer değiştirmiştir.

3.1 Çalışmada kullanılan veri seti

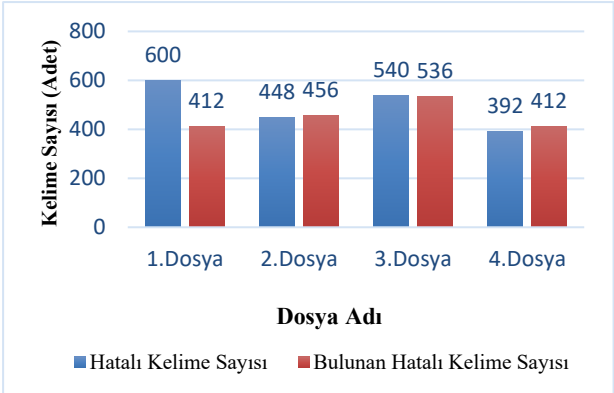
Bu çalışmada kullanılan veri seti dört adet farklı metin dosyasından oluşmaktadır. Bu dosyalardan üç tanesi internet ortamında bulunan kısa hikâyeleri, bir tanesi de Oğuz Atay'ın "Tutunamayanlar" adlı kitabının bir bölümünden oluşmaktadır. Dosyalarda bulunan kelimelerde yazım hatası oluşturulmuştur. Dosyalarda bulunan yazım hataları; kelimede eksik karakter, fazla karakter, yanlış yazılmış karakter ve yer değiştirmiş olan karakterleri içerecek şekilde organize edilmiştir. Dosyalarda oluşturulmuş olan yazım hataları Tablo 1'de gösterilmiştir.

Tablo 1. Dosyalarda yapılan yazım yanlışları.

	1.Dosya	2.Dosya	3.Dosya	4.Dosya
Eksik Karakter	163	120	200	85
Yanlış Karakter	139	120	108	113
Yer Değiştirme	151	124	112	85
Fazla Karakter	147	84	120	109

4 Deneysel sonuçlar

Deneylerde kullanılan metin dosyalarındaki yazım hataları tespit edilmiş ve düzeltilme işlemine tabi tutulmuştur. Bu işlemlerden sonra orijinal metin ile uygulama sonucunda hataları düzeltilmiş olan metin karşılaştırılmıştır. Karşılaştırma işlemi metin içerisindeki kelimeler ile aynı sırada olacak şekilde gerçekleştirilmiştir. Karşılaştırma işleminin sonucunda uygulamanın başarısı ortaya çıkmıştır. Tablo 1'de özellikleri ve yapılan hata türleri verilen dosyalar, geliştirilen uygulama ile test edilmiş ve sonuçları Şekil 4.'te gösterilmiştir.



Şekil 4. Dosyalardaki mevcut hatalı kelime sayısı ve uygulama tarafından bulunan hatalı kelime sayısı

Şekil 4'te görüldüğü gibi geliştirilen uygulama metin içerisindeki yazım hatalarının %90'ını tespit etmiştir. Bazı dosyalarda belirtilen yazım hatalarından çok daha fazlasının tespit edildiği görülmüştür. Bu durumun nedeni metinler içerisinde çok sayıda özel ismin yer almasıdır.

Kullanılan yöntemde sadece Türkçe kelimeler kontrol edildiği için morfolojik analizi yapılamayan kelimeler ve bazı özel isimler Türkçe olarak algılanmamıştır. Özel isimlerin bazılarında herhangi bir kök bulunamadığı için kelime üretilememiştir. Yapılan hataların düzeltilmesi ve kelime önerme işleminin karşılaştırılması için "MS Word" kelime işlemci programı kullanılmıştır. Çalışmada önerilen yöntemin eksik karakterden kaynaklı yazım hatalarının tespiti MS Word'e göre yüksek, fazla karakter içeren yazım hatalarının ise düşük olduğu görülmüştür.

5 Sonuç ve öneriler

Bu çalışmada ASR sistemleri için gerekli olan modellerin eğitiminde kullanılacak Türkçe metinlerdeki yazım hatalarının tespiti ve düzeltilmesi çalışılmıştır. Bu işlemler geliştirilen bir uygulama sayesinde otomatik olarak ele alınmıştır. Böylelikle büyük metin verileri üzerinde kısa sürede hata tespit ve düzeltme işlemleri gerçekleştirilmiştir.

Gerçekleştirilen uygulamanın yazım hatası olarak nitelendirdiği hatalar kelimede karakter eksikliği, kelimede karakter fazlalığı, kelimede yer değiştirmiş karakter ve kelimede fazladan yazılmış olan karakter olarak belirlenmiştir. Bu hataları tespit edebilecek yöntemler araştırılmış ve uygun yöntemlerin ele alınabilmesi için bir uygulama geliştirilmiştir. Uygulama yazım hatası yapılmış olan kelimeyi bulmakta ve hatalı kelimenin yerine uygun kelimeler önermektedir. Önerilen kelimeler içerisinden bir kelime seçilerek hatalı kelime ile yer değiştirilmiştir. Bir metin verisinde bulunan hatalı kelimelerin tespiti ve düzeltilmesi işlemi üç sayfalık bir metin verisi için yaklaşık 10-15 dakika sürmektedir. Geliştirilen uygulama ile bu süre kısaltılmış ve büyük metin dosyaları üzerinde otomatik işlem yapılması sağlanmıştır.

Çalışma sonucunda metin verisi içerisinde hatalı kelime bulma oranı yaklaşık olarak % 95 iken bu kelimelere doğru öneri verme oranının %97 olduğu görülmüştür. Bu değerler metinden metine ve hata türüne göre değişkenlik göstermektedir. Yapılan çalışma da eksik karakter hatasının tespiti ve bu tespite kelime öneri hatası oranı %5 iken fazla karakter içeren kelimeye hatalı öneri yapma oranı %1 civarındadır.

Uygulamanın testi için kullanılan veriler aynı zamanda MS Word ile test edilmiş ve bir karşılaştırma yapılmıştır. Belirtilen hata oranları MS Word'de eksik karakter hatası %7 iken fazla karakter içerme hatası % 4 olarak gözlemlenmiştir.

Yapılan çalışmada metin içerisindeki hatalı kelimelerin tespitinin başarımı her ne kadar yüksek olsa da bu kelimelere yapılan önerilerin hata oranının da yüksek olduğu gözlemlenmiştir. Bu nedenle gelecek çalışmalarda hatalı kelimelere yapılan önermelerin doğruluk oranını artırmak için n-gram tabanlı yöntemler kullanılarak farklı istatistiksel yaklaşımlar geliştirilebilir. Ayrıca uygulamaya ek olarak özel isimlerin bulunduğu geniş bir kelime sözlüğü verilmesi özel isimlerin tespitindeki başarımı arttıracaktır.

Kaynaklar

[1] Prakoso H, Ferdiana R, Hartanto R. "Indonesian Automatic Speech Recognition system using CMUSphinx Toolkit and Limited Dataset". *International Symposium Electronic Smart Devices*, Bandung, Indonesia, 29-30 November 2016.

[2] Stolcke A. 'Entropy-based Pruning of Backoff Language Models'. *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, arXiv:cs/0006025, 1988.

[3] Delibas A. "Doğal Dil İşleme İle Türkçe Yazım Hatalarının Denetlenmesi". *Yüksek Lisans Tezi*, İstanbul Teknik Üniversitesi, Bilgisayar Mühendisliği, İstanbul, Türkiye, 2008.

[4] Starlander M, Popescu-Belis A. "Corpus-based Evaluation of A French Spelling and Grammer Checker". *Third International Conference On Language Resources And Evaluation*, Las Palmas, Canary Islands, Spain, 29 -31 May 2002.

[5] Gabriele K. "A Word Analysis System for German Hyphenation Full Text Search and Spell Checking with Regard to the Latest Reform of German Orthography" *Institute of Computer Graphics*, Algorithms and Data Structures Group, Vienna University of Technology, Vienna, Austria, 2011.

[6] Dembitz S, Knezevich P, Sokele M. "Developing A Spell Checker As An Expert System" *Journal of Computing and Information Technology*, 1(4), 285-291, 2004.

[7] Dembitz S, Knezevich P, Sokele M. "Hascheck – the Croatian Academic Spelling Checker" *18th Annual International Conference of The British Computer Society Specialist Group on Expert System*, Cambridge, England, December 1998.

[8] Dhanabalan T, Parthasarathi R, Geetha TV. "Tamil Spell Checker" *Tamil Internet*, Chennai, India 2003.

[9] Murata M, Utiyama M, Uchimoto K, Ma Q, Isahara H. "Correction of Errors in a Modality Corpus Used for Machine Translation Using Machine-learning". Japan *Communications Research Laboratory*, Kyoto, Japan, 2001.
[10] Cakiroglu U, Ozyurt, O. "Türkçe Metinlerdeki Yazım Yanlışlarına Yönelik Otomatik Düzeltme Modeli", *Eleco*, Bursa, Türkiye, 7-9 Haziran 2016.
[11] Zemberek. "Ağaç Performansı ve Kök Seçiciler". http://zembereknlp.blogspot.com/2007/04/zemberek-nasl-alr-2aa-performans-ve-kk.html (15.04.20189).
[12] Akgul O. "Türkçe Kelimelerin Morfolojik Analizi". https://akgulomer.wordpress.com/2011/01/23/turkce-kelimelerin-morfolojik-analizi/ (13.04.2019).