İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KISA METİNLERDE VARLIK İSMİ TANIMA

YÜKSEK LİSANS TEZİ

Beyza EKEN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

OCAK 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KISA METİNLERDE VARLIK İSMİ TANIMA

YÜKSEK LİSANS TEZİ

Beyza EKEN
(504111504)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ

OCAK 2015

İTÜ, Fen Bilimleri Enstitüsü'nün 504111504 numaralı Yüksek Lisans Öğrencisi **Beyza EKEN**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"KISA METİNLERDE VARLIK İSMİ TANIMA"** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Yrd. Doç. Dr. Cüneyd TANTUĞ** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Yrd. Doç. Dr. Gülşen ERYİĞİT** İstanbul Teknik Üniversitesi	.....
	Yrd. Doç. Dr. Arzucan ÖZGÜR Boğaziçi Üniversitesi	.....

Teslim Tarihi : **15 Aralık 2014**
Savunma Tarihi : **22 Ocak 2015**

Aileme,

ÖNSÖZ

Tez çalışmamda bana yardımcı olan danışman hocam Sayın Yrd. Doç. Dr. Cüneyd TANTUĞ'a teşekkürü borç bilirim. Eğitim hayatım boyunca yardımlarını ve desteklerini esirgemeyen annem ve babam başta olmak üzere tez çalışmam süresince motivasyonuma katkıda bulunan tüm aileme ve arkadaşlarıma teşekkür ederim.

Aralık 2014 Beyza EKEN

İÇİNDEKİLER

Sayfa

KISALTMALAR

DDİ	**:** Doğal Dil İşleme
NLP	**:** Natural Language Processing
NER	**:** Named Entity Recognition
VİT	**:** Varlık İsmi Tanıma
MUC	**:** Message Understanding Conferences
CONLL	**:** Conference on Computational Natural Language Learning
CRF	**:** Conditional Random Fields
KRA	**:** Koşullu Rastgele Alanlar
HMM	**:** Hidden Markov Models
SMM	**:** Saklı Markov Modelleri
MEMM	**:** Maksimum Entropi Markov Modelleri

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

KISA METİNLERDE VARLIK İSMİ TANIMA

ÖZET

Ağ ortamının gelişmesi ve Web 2.0 olarak adlandırılan sürece girmesi ile birlikte içeriklerini kullanıcıların oluşturduğu sosyal medya siteleri, mikroblog (internet günlüğü) sitelerinin sayısı ve popülerliği artmıştır. Bu sitelerde paylaşılan metinlerin içinden istenilen bilgiye makineler yardımı ile kolayca ulaşma ve bilgileri yorumlama noktasında doğal dil işleme devreye girmektedir.

Doğal dil işleme insanların konuştuğu dili bilgisayarların yorumlayabilmesini amaçlayan çalışmaları kapsar. Yapay zeka ve dilbilimin bir alt dalıdır.

Varlık ismi tanıma ise doğal dil işlemenin çalışma alanlarından bilgi çıkarımının bir alt dalı olup metinlerden geçen varlık isimlerinin tanınması ile ilgilenir. Varlık ismi tanıma kapsamında varlık isimleri tespit edilir ve önceden tanımlanmış kategorilere göre sınıflandırılırlar. Bu kategorilerden en çok kullanılanları temel olarak kişi, organizasyon, yer, tarih, saat, para, yüzde varlıklarıdır.

Varlık ismi tanıma makine çevirisi, bilgi elde etme, duygu analizi, multimedia indeksleme, soru-cevap sistemleri gibi diğer doğal dil işleme uygulamalarınında kullanılabilmektedirler.

Kurallı yazılmış metinlerde varlık ismi tanıma problemi %95 seviylerine yaklaşmış iken, imla kurallarına uymadan kuralsız ve rahat bir dille yazılan metinlerde performans %50 seviyelerine düşmektedir. Kuralsız metinlerde başarımı yükseltmek adına litaratürde genel olarak iki eğilim görülmektedir; kuralsız metinler normalize edilerek kurallı formata dönüştürülür ve mevcut araçlara uygun hale getirilir ya da kuralsız metinlerin doğasına uygun yeni yöntemler ve araçlar geliştirilir.

Normazliasyon ciddi ve kompleks bir doğal dil işleme alanıdır. Bu çalışmada amaç kuralsız metinlerde kompleks bir normalizasyon uygulamadan varlık ismi tanıma yöntemleri geliştirmektir. İstatistiksel öğrenme tekniklerinden koşullu rastgele alanlar yöntemi kullanılarak tweet verileri üzerinde varlık ismi tanıma çalışmaları yapılmıştır.

Twitter adlı mikroblog sitesinde kullanıcıların paylaştığı 140 karakterle sınırlı kısa metinlere tweet denilmektedir. Bu metinler yazım ve dilbilgisi kurallarına uymadıkları gibi yabancı ve argo kelimeler, gülensuratlar (smiley) içerebilmektedirler. Tweetlerin rahat ve kuralsız doğası doğal dil işleme çalışmalarını zorlu bir hale getirmektedir.

Çalışmada iki yöntem kullanılmıştır. İlki daha önceki yöntemlere dayanan kurallı yazılmış haber metinlerde başarımı ispatlanmış, biçimbilimsel özelliklere dayanan bir yöntemdir. İkincisi ise biçimbilimsel özellikler yerine kabaca kelimenin ilk N ve son N harfleri kullanılarak geliştirilen bir yöntemdir.

Türkçenin zengin biçimsel yapısından dolayı biçimbilimsel özelliklerin doğal dil işleme çalışmalarında kullanılması kaçınılmaz bir durumdur. Fakat tweet metinlerinin biçimbilimsel çözümleme aşamasında başarılı sonuçlar elde edilememektedir, dolayısı ile bu çalışmada biçimbilimsel özellikler yerine kelimenin ilk N ve son N harfleri ile biçbilimsel bilgiler yakalanmıştır. Şöyle ki kelimenin ilk 3-4 harfi kelimenin kök bilgisini içerebilmekteyken, son 3-4 harfi ise kelimenin aldığı eklerin bilgilerini içerebilmektedir.

Her iki yöntemde kurallı ve kuralsız metinlerde denenmiştir, kurallı metinler türünde haber metinleri, kuralsız metinler türünde tweetler kullanılmıştır. Test sonuçlarına

göre heber metinlerinde her iki yöntemle neredeyse aynı başarım elde edilirken, tweet metinlerinde başarım yaklaşık %6 oranında yükselmiştir. Bu demektir ki ilk yöntemde kullanılan biçimbilimsel çözümleme ve belirsizlik giderme araçlarına ihtiyaç duymadan kelimenin ilk ve son N harfleri ile kabaca biçimbilimsel özellikler elde edilerek varlık ismi tanıma probleminde kullanılabilir.

Temel yöntem ve geliştirilen yöntemlerde biçimbilimsel aşamaların farklılıklarına ek olarak sözcüklerin küçük/büyük harf bilgileri, cümle başlangıcında bulunup bulunmamaları bilgisi ve sözlük bilgilerinden yararlanılmıştır. Geliştirilen yöntemde bunlara ek olarak tüm eğitim, sınama verileri ve sözlükler asciification işleminden geçirilmiştir. Asciification metindeki Türkçe karakterlerin (ö, ç, ş, ı, ğ, ü) ASCII tablosu karşılıklarına (o, c, s, i, g, u) dönüştürülmesini kapsayan bir terimdir. Tweet metinleri genellikle Türkçe karakterlerin kullanımına dikkat edilmeden yazıldıkları için Asciification işlemi uygulanmıştır, yanlış yazılan kelimeleri normalize etmek yerine tüm verilerden Türkçe karakterleri kaldırma yöntemi tercih edilmiştir.

Varlık isimlerini içeren geniş sözlüklerin kullanımına varlık ismi tanıma sistemlerinden sıklıkla başvurulur. Bu çalışmada kişi, yer ve organizasyon olmak üzere üç adet isim sözlüğü kullanılmıştır, sisteme verilen girdi metindeki sözcükler sözlükler ile eşleştirilerek varlık ismi olup olmadığına karar verilebilir. Geliştirilen yöntemde temel alınan yönteme ek olarak sözlük eşleştirmesinde kısmı eşleşme yöntemi kullanılmıştır. Kısmi eşleşmenin amacı tweetlerde kelimelerin birkaç harfini atlayarak yanlış yazılan varlık isimlerinin tespitini sağlamaktır. Örneğin “Mehmet” yerine “Mhmet” yazılabilmektedir. Böyle yanlış yazılmış varlık isimlerini tespit etmek amacıyla girdi metindek sözcükler ile sözlükteki kelimelerin Levenshtein mesafe algoritması ile uzaklıkları hesaplanmış ve küçük uzaklıktaki kelimeler aday varlık ismi olarak gösterilmiştir.

NAMED ENTITY RECOGNITION ON TURKISH SHORT TEXTS

SUMMARY

With the expansion of the web and material sharing through the internet, an excessive amount of information pollution has emerged. Microblog web sites has been rising as a new trend in alternative information sharing. Posts on these microblog sites sometimes contain important information about a product, a person or even crucial news about distant corners of the world. But maybe the most significant point about these microblog posts is that they hold great amounts of statistics. Therefore, gathering and processing this information has become essential. Thus, at this point, natural language processing is holds meaningful importance.

Natural language processing (NLP) is a sub-branch of artificial intelligence and it is concerned with interpretation of natural language by machines. NLP's aim is to make the interaction easier between human and computers.

Named entity recognition is a NLP term that refers to the recognition of named entities in natural language. It is a way of extracting information by detecting and classifying named entities in texts. Named entities mainly refer to persons, locations, organizations. Originally, named entity term just refers these three entity but later it expand to involve numeric expressions like date, time, money, percentage. Recently it has branched out to bioinformatics for protein recognition, gene recognition, medicine recognition etc.

There have been lot of studies in named entity recognition field and state of the art perfomance has reached to nearly human annotation performance in formal texts. But off the shelf natural language processing tools' performances decreases when they are applied to informal texts. Because informal texts may not fit to grammar rules, spelling rules unlike formal texts. As a consequence, the need arises to develope new tools that would be proper for informal structure of texts or normalizing informal texts to fit to off the shelf NLP tools.

The aim of this study is to increase performance of named entity reconition on Turkish tweets without appliying complex normalization. Tweets are micro texts that shared on Twitter website which are limited to 140 characters. Most of the time they can contain grammar and spelling errors, slang words and smileys, so unfortunately this nature of tweets make harder to process such data.

Turkish is a highly agglutinative language, so natural languge processing is much more challenging than non-agglutinative languages. Because of agglutinative and morphologically rich nature, morpholocigal features are indispensable in natural language processing applications.

Before works on named entity recognition in Turkish formal texts also have shown that it is useful to use morphological features of words. However for informal text domain it is not easy get morphological features with off the shelf morhphological analysis tools efficiently. It could be an option to normalize informal texts to fit existing tools, but in this work instead of normalizing, morphological information captured roughly without morphological processing tools. For example first N character and last N characters of words can hold morphological informations in Turkish, first N characters of the word nearly equals the stem information of the word and last N characters of the word can contain suffixes.

Based upon tweets' informal nature two different model are developed. First model is base model, which based on another named entity recognition work on Turkish news domain. Second method is developed method in this work as an alternative to base method. This developed method increased the performance tweet domain nearly %6, and nearly get same performance in news domain. Base model includes morphological analyzing and disambiguation process differently from developed model.

Despite the aim is to develop named entity recognition system on tweet domain, news domain also used as formal text domain to compare results of models. News data fit to spelling and writing rules as opposed to tweets data. Tweets have unstructural nature, can contain slang words, repeating characters for exclamation, on purpose spelling errors, abbreviations. In addition, Turkish tweets can be written with only using English alphabeth characters instead of using some Turkish characters (ö, ç, ş, ı, ğ, ü). Moreover, instead of using any alphabet characters some symbols can be used as letters in words ("$eker" instead of "şeker", "s€n" instead of "sen").

Morphological structure gives important clues about which is part-of-speech information, inflectional and derivational affixes. These morphological informations of words can give a clue about being a named entity or not. So in both models morphological informations are used.

Base model gets morphological information using external morphological tools for analyzing and disambiguating. Unfortunately, when these tools applied to tweets, its performance decreases and could not get morphological information of tweets accurately. As mentioned before, first and last N characters of Turkish words have significant information about morphological structure of the word. Instead of external tools, getting first and last N characters are easier.

Beside morphological features case feature, start of sentence feature, gazetteer features are used in both models.

Letter case of a word has an important clue about being a named entity. Named entities are mostly proper names and proper names should start with uppercase letter, so case information of word has significant increase in performance. Also sentences start with uppercase letter, it can be cause confusion, so being first word of sentence information and case information of word are used as features.

Most of the NER studies take advantage of gazetteers, which are large lists of entity names. In this project three gazetteers are used that are person names, location names and organization names. Organization gazetteer mostly consists of organization name generator words like "ministry", "university", "association" etc.

Differently in developed model some features are added to handle obstacles of twitter domain. They are simple normalization, asciification and apostrophe information. Moreover, partial looking up tecniques applied on both base and developed models.

Firstly simple normalization techniques are applied to data which are converting symbols in words to equivalent letter, removing repeated characters more than two. Two repetition of characters is allowed in normalization procedure because some words originally has two same adjacent letters. Than all Turkish characters in tweets are converted to equivalents in ASCII table which named asciification procedure.

Proper names' affixes should be separated with apostrophe so if a word has apostrophe it is probably a named entity. Apostrope information is also used as feature to create developed model.

Due to tweet domain peculiarities exact matching of words with gazetteer's elements yields to miss out some named entities which have spelling errors or which are contracted forms of entities. Some words can be written in contracted form or one or more characters in a word can be skipped on purpose or not. These contracted forms or incorrect written words can be a named entity which correct form of them exists in gazetteers. Partial matching techiques used not to miss out these incorrect entites in text. Words in input text compared to gazetteers and distances are calculated with Levenstein distance algorithm. When comparing words with gazetters' elements if any zero distance has occur that means exactly same word exists in relevant gazetteer and the input word is a named entity. If one distance has occure it means input word could be a named entity with one contraction and so on. To decrease time cost of calculating distances of input word, it is compared to words in gazetteers that only starts with same letter as input word instead of comparing each word in gazetteers.

As a result, base model reaches %90.38 f1-measure, and developed model reaches %90.23 f1-measure in news domain according to CoNLL metrics. In tweet domain, base model reaches %55.29 f1-measure and developed model reaches %61.05 f1-measure according to CoNLL metrics. Base model's performance could be achieved without using morphological processes with external tools. Moreover, developed model gives better results for informal domain.

So it means no need to use complex morphological and normalization procedure to get reasonable results on tweet domain.

Previous named entity recognition works have shown that conditional random fields (CRF) method reached good performance at recognizing entities. Consequently, in this work CRF have been chosen as the method to build named entity recognition model. CRF are a statistical modelling method used in sequential data classifying. CRF++ tool used to perform CRF algortihm in this study.

Herewith this study is an experiment of named entity recognition on Turkish tweet domain. Tweets have short and informal nature and this make this domain much harder than formal text domain. All of mentioned features has improved the performance of the system for tweets. To improve performance gazetteers can be update, because everday new entites can derive. In addition, partial matching methods can be enhance.

Developed model only recognizes seven entity types person, location, organization, date, time, money and percentage, but it can be extended to recognize other entity types for example product entity recognition can be important to analyze to get some statistics for companies about their products.

1. GİRİŞ

Ağ ortamındaki bilgi paylaşımının kolaylaşması ile paylaşılan bilgi de hızla artmaktadır. 2014 yılında dünya çapındaki toplam web site sayısı 1 milyarın üzerine çıkmıştır (Url-1, 2014). İnternet ortamı kullanıcıların sadece içerik aldıkları bir ortam değil içeriğine katkıda bulundukları bir ortama dönüşmüştür. Bu süreci tanımlamak için Web 2.0 kavramı ortaya atılmıştır (Url-2, 2014). Bu kavram kullanıcıların ortaklaşa içerik ürettiği ve paylaştığı ikinci nesil web site teknolojilerini kapsamaktadır. Blog (internet günlüğü) siteleri, mikroblog siteleri, sosyal paylaşım siteleri, vikiler Web 2.0 içeriğini oluşturan parçalardır. Facebook, Twitter, Vikipedi, Youtube, LinkedIn, Instagram en popülerlerinden birkaç tanesidir.

Giderek artan siteler ve kullanıcı sayısı sebebiyle üretilen ve paylaşılan bilgi de giderek artmaktadır. Bu paylaşımlardaki metinler ve yorumlar bir ürün, kişi, mekan, firma vs. hakkında bilgiler ve/veya kullanıcı görüşlerini içerebilmektedir. Örneğin bir firma yeni çıkan ürünü ve firma hakkındaki müşteri yorumlarını incelemek ve buna göre strateji belirlemek isteyebilir. Bu durumda firma isminin geçtiği metinlere ulaşmak isteyebilir. Bunun yanında teknolojinin hayatın her alanına girmesiyle belge yönetim ve arşivleme işlemleri elektronik ortamlarda yapılmaktadır. Örneğin belirli bir kişiye ait belgelere kolayca ulaşmak istenebililir, bu durumda kişi ismine göre metin bölümleri veya belgeler indekslenebilir. Ağ ve elektronik ortamlardaki bilgilere erişmek ve bu bilgilerden yararlanma konusunda her geçen gün yeni teknolojiler üretilmekte ve daha pratiklerine ihtiyaç duyulmaktadır.

Oluşan bilgi yığınları içindeki metinleri veya ses verilerini yorumlayarak ihtiyaç duyulan bilgiye makineler aracılığı ile daha kolay ulaşabilmek için yapay zekanın bir alt dalı olan doğal dil işleme (DDİ) tekniklerinden yararlanılmaktadır.

DDİ insan ve bilgisayar arasındaki iletişimi kolaylaştırmayı hedefleyen, dilbilim ve yapay zekanın kesiştiği bir alandır. İnsanların konuştuğu doğal dili makinelerin anlamasına ve yorumlamasına yönelik çalışmaların yanında üretilen makine çıktılarını insanın anlayabileceği doğal dil formatına uydurmaya yönelik çalışmaları

kapsamaktadır. DDİ uygulama alanlarına örnek olarak konuşma tanıma, konuşma sentezleme, makine çevrisi, bilgi çıkarımı verilebilir.

Varlık ismi tanıma (VİT) bilgi çıkarımının bir alt kategorisi olup metinlerde geçen varlık isimlerinin tespiti ve sınıflandırılması ile ilgilenir (Nadeau ve Sekine, 2007). VİT çalışmaları 1990' ların başında başlamış olup, varlık ismi kavramı ilk kez 1996 yılında 6. MUC (Message Understanding Conference) konferansında kullanılmıştır (Nadeau ve Sekine, 2007). Bu kavram ilk başlarda sadece kişi, yer, organizasyon isimlerini kastederken daha sonra tarih, zaman, yüzde gibi sayısal birimleri de kapsayacak şekilde genişlemiştir. Sonraki yıllarda daha da çeşitlenerek çalışılan metnin türüne göre tıbbi terimleri, protein isimlerini, gen isimleri vs. kastedecek formata bürünmüştür.

Metinlerdeki varlık isimlerinin tespiti bilgi çıkarımında önemli rollere sahiptir. Makine çevirisi, duygu analizi, soru cevap sistemleri gibi çeşitli doğal dil işleme uygulamalarının önemli bir parçası olabilmektedir. Örneğin metindeki söz konusu varlık isminin yanında varlık hakkında bir bilgi veya görüş verilmiş olabilir ya da varlık isminin kendisi bir sorunun öğrenilmek istenen cevabı olabilir.

Daha önce yapılan çalışmalar gösteriyor ki VİT neredeyse çözülmüş bir problemdir, çalışmalar çoğu dil için düzgün formatta yazılmış metinlerde genellikle %90 üzerinde başarım elde etmektedir. Fakat metinler her zaman düzgün formatta yazılmış olmayabilir. Özellikle web ortamındaki sosyal paylaşım sitelerinde paylaşılan günlük konuşma formatındaki metinlerde imla kurallarına, dil bilgisi kurallarına dikkat edilmemektedir. Metinlerin içerikleri genelde az olmakta ve argo kelimeler, kısaltmalar, kasıtlı yazım hataları içerebilmektedirler. Böylesi metinlerde doğal dil işleme çalışmaları yapmak zorlaşmaktadır ve mevcuttaki doğal dil işleme araçlarının performansı düşmektedir.

Formatsız metinlerdeki performansı iyileştirmek için bu metinlerin doğasına uygun olarak yeni doğal dil işleme teknikleri geliştirilmeli ya da normalizasyon yöntemleri ile bu metinler mevcuttaki doğal dil işleme araçlarında kullanılabilecek formata getirilmelidirler.

1.1 Tezin Amacı

Bu tez çalışmasının amacı web ortamında sayıca artmakta olan sosyal paylaşım ve mikroblog sitelerinde paylaşılan farklı sitelerde farklı isimlerle (statü, tweet) adlandırılabilen kısa metinler üzerinde varlık ismi tespiti yapacak bir model geliştirmektir.

Günümüzde sayıları artmış olan mikroblog sitelerinde kullanıcılar anlık olarak yaptıklarını; bir ürün, mekân, kişi veya olay hakkında yorumlarını; duydukları haberleri vs. paylaşmaktadırlar. Uzunluğu sınırlı olan bu paylaşımların içerdiği bilgilere veya yorumlara ulaşmada VİT önemli bir adım olacaktır.

Bu çalışmada günümüzde oldukça popüler olan Twitter adlı sosyal paylaşım sitesinden alınan Türkçe tweetler üzerinde çalışılmıştır. Tweetler üzerinde VİT çalışması düzgün formattaki metinlere oranla daha zordur çünkü tweetler içeriği 140 karakterle sınırlı, imla ve yazım kurallarına uymayan, günlük konuşma dilinde yazılmış metinlerdir. Bu durum sadece VİT değil diğer DDİ çalışmalarını da zorlaştırmaktadır.

Türkçenin biçimbilimsel olarak zengin bir dil olması sebebiyle Türkçe VİT çalışmalarında biçimbilimsel özelliklerin kullanılması başarımı ciddi oranda artırdığı gözlemlenmiştir (Yeniterzi, 2011; Şeker ve Eryiğit, 2012). Fakat tweet verilerindeki yazım hatalarından ötürü biçimbilimsel çözümleme zorlaşmaktadır ve biçimbilimsel çözümleme araçlarının performansı düşmektedir. Bu durum tweet metinlerinde varlık ismi tanıma için yeni bir teknik arayışına yönlendirmiştir.

Bu çalışamada öncelikle düzgün formattaki Türkçe haber metinleri kullanılarak Şeker ve Eryiğit' in (2012) çalışmasında kullandığı başarımı ispatlanmış teknikler kullanılarak bir VİT modeli geliştirilmiştir. Daha sonra aynı model ile tweet metinleri test edilmiş ve Çelikkaya ve diğ. (2013) çalışmasında da olduğu gibi başarımın ciddi oranda düştüğü gözlemlenmiştir. Tweet metinlerinin başarımını yükseltmek için yeni teknikler ve tweet verileri ile eğitilen yeni bir model oluşturulmuştur, bu yeni model ile başarımın yükseldiği gözlemlenmiştir.

1.2 Benzer Çalışmalar

Varlık ismi tanıma alanında İngilizce başta olmak üzere birçok dilde çalışmalar yapılmıştır. İngilizce için son elde edilen başarımlar %95 seviyelerindedir.

İlk oluşturulan VİT sistemleri kural tabalı iken zamanla makine öğrenmesi tekniklerinin kullanımı yaygınlaşmıştır (Nadeau ve Sekine, 2007).

İlk çalışmalar 1990' lı yıllarda başlamıştır, bunlardan biri Rau' nun (1991) İngilizce için yapılmış olup elle tanımlanmış kurallar yardımıyla metinden şirket isimlerini tanımaya yönelik çalışmasıdır.

Varlık ismi kavramı ilk kez MUC-6 (6. Message Undersanding Conference) konferansında ortaya atılmıştır (Nadeau ve Sekine, 2007). MUC-6 konferansı ile birlikte VİT çalışmaları hızlanmıştır (Grishman ve Sundheim, 1996). Sonraki senelerde varlık ismi tanıma alanında düzenlenen konferanslar artmıştır, CoNLL, IREX, HAREM, ACE bu konferanslardan bazılarıdır.

İngilizce için en yüksek başarım elde etmiş çalışmalardan biri Ratinov ve Roth' un (2009) çalışmasıdır, CoNLL03 verisinde %90.8 başarım elde etmişlerdir.

Yarowsky ve Cucerzan' ın (1999) dilden bağımsız çalışması Türkçe için yapılmış ilk çalışma olarak kabul edilir, bu çalışma Romence, Yunanca, İngilizce, Hintçe ve Türkçe üzerinde test edilerek Türkçe için %53'lük bir başarım elde edilmiştir.

Sadece Türkçeye özgü olan ilk çalışmada ise Tür ve diğ. (2003) tarafından yapılmış, haber metinleri üzerinde saklı markov modelleri (SMM) kullanarak %91.56' lık bir başarım elde edilmiştir. İstatistiksel modelin kurulmasında kelimelerin sözlüksel, bağlamsal ve biçimbilimsel özelliklerinden yararlanılmıştır. Kişi, yer ve organizasyon isimlerinin tespiti ile ilgilenilmiştir.

Türkçe finansal metinlerden kişi ismlerinin tespitine yönelik yapılmış olan Bayraktar ve Temizel' in (2008) çalışmasında ise örüntü modellerinden ve kelimelerin sıklık analizlerinden yararlanılmıştır.

Türkçe için Küçük ve Yazıcı (2009) tarafından yapılan çalışma ise kural tabanlı geliştirilmiş bir sistemdir, haber metinleri, tarihsel metinler, çocuk hikâyeleri üzerinde kişi, yer, organizasyon ve nümerik ifadelerin tespiti ile ilgilenilmiştir. Bu kural tabanlı sistem daha sonra istatistiksel öğrenme yöntemiyle geliştirilerek haber metinleri üzerindeki başarımı %87.96' dan %90.13' e çıkarılmıştır (Küçük ve Yazıcı, 2012).

Koşullu rastgele alanlar yöntemi ile Türkçe e-posta metinlerinde kişi, yer, organizasyon tespitine yönelik yapılmış Özkaya ve Diri' nin (2011) çalışmasınnda %92.89 başarım elde edilmiştir. Bu çalışmada sözlüklerden; kelimelerin uzunluk,

kısaltma olup olmadığı, noktalama işareti içerip içermediği, cümlenin ilk kelimesi olup olmadığı vs. bilgilerinden yararlanılmıştır.

Otomatik kural öğrenme tabanlı oluşturulmuş bir sistemde Tatar ve Çiçekli (2011) tarafından haber metinleri üzerinde %91.08' lik bir başarım elde edilmiştir. Kelimelerin büyük/küçük harf bilgisi, uzunluğu, bağlamsal özellikler, biçimbilimsel özellikler ve sözlüklerden yararlanılmışır. Oluşturulan sistemin Türkçe gibi sondan eklemeli olan başka diller için de kullanılması hedeflenmiştir.

Yeniterzi (2011) tarafından kelimelerin biçimbilimsel özelliklerinden ve büyük/küçük harf bilgisinden yararlanarak koşullu rastgele alanlar (CRF) yöntemi ile Türkçe için yapılan bir çalışmada ise haber metinlerinde CoNLL metriğine göre %88,94' lük bir başarım elde edilmiştir.

Kelimelerin biçimbilimsel özelliklerininin yanısıra büyük/küçük harf bilgisi, cümle başlangıcı bilgisi ve sözlük bilgileri kullanılarak koşullu rastgele alanlar (CRF) yöntemi ile Türkçe için Şeker ve Eryiğit (2012) tarafından yapılan bir başka çalışmada ise haber metinlerinde CoNLL metriğine göre %92' lik bir başarım elde edilmiştir.

Şimdiye kadar bahsedilen çalışmalar düzgün formatlı, imla ve dilbilgisi kurallarına uygun yazılmış metinler üzerinde yapılmış çalışmalardır. Fakat özellikle web ortamında paylaşılan bilgiler her zaman düzgün formata uymayabiliyorlar, böylesi metinler üzerinde yapılan çalışmalar son zamanlarda artış göstermiştir.

İngilizce tweet metinlerinde yapılan bir çalışmada Ritter ve diğ. (2011) başarımı artırmak için mevcuttaki doğal dil işleme araçlarını tweet metinlerinin doğasına uyarlamışlardır. Sözcük sınıfı etiketleme, isim tamlamalarının bulunması, varlık ismi bölütleme aşamalarında koşullu rastgele alanlar yöntemi kullanılmış, varlık ismi sınıflandırma aşamasında ise etiketlenmiş konu modelleme yöntemi kullanılmıştır. Sistem %51 başarım elde etmiştir.

İngilizce tweetler için yapılmış bir başka çalışmada Liu ve diğ. (2011) en yakın komşu algoritması ve koşullu rastgele alanlar yöntemi ile yarı denetimli bir sistem oluşturulmuştur. Metinlerin yazımsal, sözcüksel öznitelikleri ve ek sözlük yardımıyla %80.2 başarım elde edilmiştir.

Li ve diğ. (2012) İngilizce tweetlerde varlık isminin tespit eden denetimsiz öğrenme yöntemiyle bir sistem oluşturmuşlardır.

Oliveira ve diğ. (2013) İngilizce tweetler için filtre tabanlı bir varlık tanıma sistemi geliştirmişlerdir. Bu sistem koşullu rastgele alanlar yöntemine göre %3 daha fazla başarılı olmuştur.

Türkçe tweetlerde ise varlık ismi tanıma kapsamında iki çalışma mevcuttur. Bunlardan ilkinde Çelikkaya ve diğ. (2013) tweet verilerini normalizasyon aşamasından geçirdikten sonra haber metinleri ile eğitilmiş istatistiksel model üzerinde test etmişlerdir ve CoNLL metriğine göre %19.28' lik bir başarım elde edilmiştir. İstatistiksel modelin oluşturulmasında Şeker ve Eryiğit' in (2012) çalışması ile aynı teknikler kullanılmıştır.

Tweet metinleri için yapılmış diğer Türkçe çalışmada Küçük ve Steinberger (2014) daha önce Küçük ve Yazıcı (2009) tarafından geliştirilmiş kural tabanlı sistemi geliştirmişlerdir. Tweet metinleri kural tabanlı sistemle test edildiğinde %53.23' lük bir başarım elde edilmiştir. Sonrasında kural tabanlı sistem geliştirilerek başarım %61.17' ye yükseltilmiştir. Kural tabanlı sistemin geliştirilmesinde normalizasyon, kelimelerin büyük/küçük harf bilgileri, sözlük bilgisi kullanılmıştır.

2. VARLIK İSMİ TANIMA

2.1 Varlık İsmi Tanıma Nedir

Varlık ismi tanıma doğal dil işlemenin çalışma alanlarından bilgi çıkarımının bir alt dalı olup metinlerde geçen varlık isimlerinin tespiti ve snıflandırılması ile ilgilenir (Nadeau ve Sekine, 2007). Bilgi çıkarımı doğal dil işleme teknikleri kullanılarak herhangi bir formatı olmayan metinlerden bir bilgiyi elde etmeye yönelik çalışmaları kapsayan bir alandır. VİT uygulamalarında metindeki varlık isimleri tespit edilir ve önceden belirenen kategorilere göre sınıflandırılırlar. VİT uygulamasına örnek olarak Şekil 2.1 gösterilebilir.

Şekil 2.1 : Varlık ismi tanıma örneği.

1990' lı yıllarda başlayan ilk çalışmalar VİT problemini özel isimleri tanıma problemi olarak tanımlamışlar, genel olarak üç tür üzerinde çalışmıştır; kişi, yer, organizasyon (Nadeau ve Sekine, 2007). Bu kategoriler ilk başta kişi, yer, organizasyon isimleri iken daha sonra tarih, zaman, yüzde gibi sayısal birimleri de içerecek şekilde genişlemiştir. 1996 yılında 6. MUC (Message Understanding Conference) konferansında varlık ismi kavramı ortaya çıkmıştır. Daha sonraki çalışmaların çoğunda baskın olarak kullanılan varlık ismi kategorileri şu şekilde tanımlanmıştır;

- ENAMEX
    - Person (Kişi)

    - Location (Yer)
        - Ülkeler, şehirler, dağ, ırmak vb.
    - Organization (Organizasyon)
        - Devlet birimleri, şirketler, dernekler vb.
- TIMEX
    - Date (Tarih)
    - Time (Zaman)
- NUMEX
    - Money (Para)
    - Percent (Yüzde)

Daha sonraki yıllarda VİT problemi üzerinde yapılan çalışmalarda düzenlenen konferanslarda farklı kategorilere de yer verilmiştir. Bunlardan en popüler olanları CoNLL konferansında tanımlanan, kelime olarak "karışık" anlamına gelen "miscellaneous" kategorisidir. Bu kategori özel isim olan fakat enamex kategorisinde değerlendirilemeyen varlık isimleri için kullanılmaktadır. Başka çalışmalarda alt kategoriler de kullanılmıştır, örneğin "kişi" varlık ismi yerine "politikacı", "mühendis", "aktör" vb., "yer" varlık ismi yerine "şehir", "ülke", "köprü" vb. alt tanımlamalar kullanılabilmektedir.

Bunların dışında kimyasal madde, ilaç isimleri, protein isimleri, gen tespit etmek gibi alana özel çalışmalar da yapılmaktadır. Son zamanlarda popüler olan VİT uygulama alanlarından biri biyomedikal metinlerdeki protein ve gen isimlerinin tanınmasıdır.

Bu alanda İngilizce, Almanca, İspanyolca, Hintçe, Arapça, Çince vs. birçok dilde çalışmalar mevcuttur. Birçok dil için VİT sistemlerinin performansı %90 seviyesine çıkabilmiş hatta neredeyse insan etiketleme seviyesine ulaşmıştır. İnsanların %100 doğru olarak etiketleme yapamadıkları, kişiden kişiye etiketlemelerin farklılık gösterdiği görülmüştür, MUC-7 konferansındaki yorumcuların etiketlediği metinlerin başarımı %96-97 seviyesindedir (Marsh ve Perzanowski, 1998).

Türkçenin zengin ve karmaşık biçimbilimsel yapısı Türkçe için yapılan doğal dil işleme çalışmalarını daha zorlu bir hale getirmektedir. Buna rağmen Türkçe için de VİT sistemlerinin başarımı %90 üzerine çıkmayı başarmıştır.

Genellikle VİT sistemleri tek bir dil ve tek bir metin türü üzerinde verimli sonuçlar elde etmektedir dolayıyla tek bir dil ve metin türü odaklı çalışmalar çoğunluktadır, örneğin Türkçe çalışmaların çoğu haber metinleri üzerinedir. Birden çok metin türü, konusu üzerinde yapılan çalışmalar her metin türü için aynı performansı gösterememiştir. Bunun en temel sebebi farklı konulardaki metinlerde farklı terimler dolayısı ile farklı varlık isimleri olmasıdır.

Bunun yanında dilden bağımsız çalışmayı hedefleyen çalışmalar da mevcuttur, bu çalışmalarda tek bir sistemle farklı dillerdeki metinlerde varlık ismi tanıma gerçekleştirilebilmesi hedeflenmiştir. Farklı dillerin farklı dilbilgisi ve yazım kuralları olacağından böylesi bir sistem oluşturmak kolay değildir.

Son yıllarda gayri resmi, günlük konuşma diliyle yazılan metinlerdeki VİT çalışmaları hız kazanmıştır, çünkü web ortamının ve iletişim teknolojisinin gelişmesi ile birlikte bu tür metinler çoğalmıştır. Bu metinlere örnek olarak e-posta metinleri, sosyal paylaşım sitelerindeki kullanıcı yorumları ve paylaşımları, cep telefonlarında kullanılan kısa mesajlar, ses tanıma sistemlerinin çıktı metinleri verilebilir. Bu tür metinler yazıldığı dilin yazım ve imla kurallarına uymayabileceği gibi birden çok dil yada jargon kullanılarak yazılmış da olabilirler, bu durumlar bu tarz metinlerdeki tüm DDİ çalışmalarını zorlaştırmaktadır.

VİT sistemleri doğal dil işlemenin birçok uygulamasında yer bulabilmektedir. Bilgi çıkarımının bir alt dalı olup istenen bilgiye ulaşmakta önemli bir aşama olarak kullanılmaktadır. Örneğin soru-cevap sistemlerinde sorulan soru ya da cevabı bir kişi, mekan veya şirketle ilgili olabilir, ilgiliyi bilgiye erişim için bahsedilen varlığa ait varlık isimlerinin tespitinden yararlanmak kolaylık sağlamaktadır. Bunun yanında arama motorlarında, çoklu ortam indekslemede, makine çevirisi ve duygu analizi gibi DDİ uygulamalarında VİT sistemlerinden yararlanılmaktadır.

2.2 Kullanılan Yaklaşımlar

Varlık ismi tanımada kural tabanlı ve istatistiksel yaklaşım olmak üzere genel olarak iki yaklaşım mevcuttur, bu iki yaklaşımı birlikte kullanan melez sistemlerde mevcuttur.

- Kural tabanlı yaklaşım
- İstatistiksel yaklaşım

Kural tabanlı yaklaşımda dilin dilbilgisi özelliklerine dayalı kurallar kullanılarak varlık ismi tanıma yapılır, istatistiksel yaklaşımda ise makine öğrenmesi teknikleri kullanılarak eğitilmiş istatistiksel bir model kullanılır. İlk sistemler elle tanımlanmış kural tabanlı algoritmalar kullanırken, sonraki sistemler daha çok makine öğrenmesi teknikleri ile otomatik kural çıkartmaya veya sıralı etiketleme algoritmalarına dayanmaktadır (Nadeau ve Sekine, 2007).

Kural tabanlı yaklaşımda kurallar varlık isimlerinin ayırt edici özelliklerine dayanarak oluşturulur. Örneğin düzenli ifadelerden yararlanılarak tarih, saat, büyük harfle başlayan isimler tespit edilebilir.

Organizasyon ve yer isimlerinde genel olarak bir örüntü mevcuttur, Şekil 2.2' deki gibi örüntülerden belli kurallar çıkarılabilir. Örneğin büyük harfle başlayan bir kelimeden sonra "Üniversitesi" kelimesi geliyorsa bu bir organizasyon ismidir denilebilir (Küçük ve Yazıcı, 2009).

Şekil 2.2 : Örüntü örnekleri.

Makine öğrenmesine dayalı istatistiksel yöntemler ise geniş bir derlemin etiketlenerek makine öğrenmesi algoritmalarıyla eğitilerek bir model çıkarılmasına dayanır. Bu yöntemde eğitilecek derlemi oluşturmak ve etiketlemek ciddi emek istemektedir. Denetimli öğrenme teknikleri geniş bir etiketlenmiş derlem gerektirdiği ve bu masraflı olduğu için yarı-denetimli veya denetimsiz yöntemlere yönelimler olmuştur.

2.3 Kullanılan Öznitelikler

Varlık ismi tanıma sistemleri metinlerdeki kelimelerin özniteliklerini kullanarak tanıma işlemini gerçekleştirmektedirler. Bu öznitelikler kelimenin yazımsal, biçimbilimsel öznitelikleri olabilmektedir.

Üzerinde çalışılan dile ait bilindik kişi, yer, organizasyon isimlerini içeren geniş sözlüklerden yararlanmak varlık ismi tanıma çalışmalarına büyük fayda sağlamaktadır. Sözlüklerden yararlanmak varlık ismi tanıma problemini tam olarak çözemeyecektir çünkü zamanla dile yeni isimler eklenebilmekte ve girdi metninde sözlükte olmayan bir varlık ismi yer alabilmektedir. Zamanla dile yeni isimler eklenebildiği gibi bazı isimler kullanım sıklığı azalabilmektedir, dolayısıyla sözlüklerin sürekli güncellenmesi sistemin başarımını artıracaktır.

2.4 Varlık İsimlerinin Gösterilimi

Varlık isimleri ve çeşitli DDİ uygulamalarında kullanılan metin parçalarının etiketlenmesi ve etiketlerinin gösteriliminde çeşitli yöntemler kullanılmaktadır. Bu yöntemler aşağıdaki gibi listelenebilir (Tjong Kim Sang ve Veenstra, 1999).

- Sade gösterim
    - IO
- İçerde/Dışarda (Inside/Outside) gösterimi
    - IOB1
    - IOB2
    - IOE1
    - IOE2
- Başlangıç/Bitiş (Start/End) gösterimi
    - BILOU

Sade gösterimde sadece I ve O etiketleri kullanılır. Bir kelime varlık ismi ise I değilse O harfi ile etiketlenir.

İçerde/Dışarda gösteriminde söz konusu kelime bir varlık ismi değilse O etiketi ile gösterilir. Varlık ismini oluşturan kelimelerin alacağı etiketler ise alt kategorilere göre değişiklik gösterir;

- IOB1: Sadece kendisinden hemen önce başka bir varlık ismi yer alan varlık isimlerinin ilk kelimesi B, diğer kelimeleri I ön ekiyle etiketlerini alırlar. Söz konusu varlık isminin kendisinden önce bir varlık ismi yer almıyorsa tüm kelimeleri I ön ekiyle etiketlenir.
- IOB2: Varlık isminin ilk kelimesi B, diğer kelimeleri I ön ekiyle etiletlerini alırlar.
- IOE1: Sadece kendisinden hemen sonra başka bir varlık ismi gelen varlık isimlerinin son kelimesi E, diğer kelimeleri I ön ekiyle etiketlerini alırlar. Söz konusu varlık isminin kendisinden sonra bir varlık ismi gelmiyorsa tüm kelimeleri I ön ekiyle etiketlenir.
- IOE2: Varlık isminin son kelimesi E, diğer kelimeleri I ön ekiyle etiletlerini alırlar.

Başlangıç/Bitiş gösteriminde ise ilk yönteme ek olarak tek eleman simgesi (U) kullanılır. B varlık isminin başlangıcını, L bitişini, I etiketi de başlangıç ve bitiş arasındaki kelimeleri temsil eder. Eğer varlık ismi tek kelimeden oluşuyorsa B ve L etiketleri yerine U etiketi kullanılır. Varlık ismi haricindeki kelimeler için O etiketi kullanılır. Bazı çalışmalarda bu gösterim "O+C" gösterimi olarak da geçmektedir. Etiketlerde B, I, U, O harflerinden başka harflerin kullanımı da mevcuttur, örneğin tüm gösterimlerin olduğu Şekil 2.3'de BILOU gösterimi için L harfi yerine diğer gösterimlerle tutarlı olması açısından E harfi kullanılmıştır. Şekil 2.3' de "Etkinliğimiz haftaya Salı günü İstanbul Teknik Üniversitesi Bilgisayar ve Bilişim Fakültesinde Ayşe Hanım koordinatörlüğünde gerçekleştirilecektir." cümlesinin varlık isimleri ve farklı gösterilimleri verilmiştir.

Kelime	Etiket	IO	IOB1	IOB2	IOE1	IOE2	BILOU
Etkinliğimiz	-	O	O	O	O	O	O
haftaya	-	O	O	O	O	O	O
Salı	TARİH	I	I	B	I	E	U
günü	-	O	O	O	O	O	O
İstanbul	ORGANİZASYON	I	I	B	I	I	B
Teknik	ORGANİZASYON	I	I	I	I	I	I
Üniversitesi	ORGANİZASYON	I	I	I	E	E	E
Bilgisayar	ORGANİZASYON	I	B	B	I	I	B
ve	ORGANİZASYON	I	I	I	I	I	I
Bilişim	ORGANİZASYON	I	I	I	I	I	I
Fakültesinde	ORGANİZASYON	I	I	I	E	E	E
Ayşe	KİŞİ	I	B	B	I	E	U
Hanım	-	O	O	O	O	O	O
koordinatörlüğünde	-	O	O	O	O	O	O
gerçekleştirilecektir	-	O	O	O	O	O	O
.		O	O	O	O	O	O

Şekil 2.3 : Varlık ismi gösterilimleri.

2.5 Varlık İsmi Tanıma Sistemlerinin Başarımının Ölçülmesi

Varlık ismi tanıma sistemlerinin başarımı sistemin çıktısı olan metindeki tespit edilmiş varlık isimlerinin doğruluğu kontrol edilerek ölçülür. Çıktı metindeki doğru tespit edilmiş varlık ismi sayısının tespit edilmiş toplam varlık ismi sayısına oranı tutma oranı (precision) olarak adlandırılır **(2.1)**.

$$tutma = \frac{doğru\ tespit\ edilmiş\ varlık\ ismi\ sayısı}{tespit\ edilmiş\ varlık\ ismi\ sayısı} \quad \textbf{(2.1)}$$

Çıktıdaki doğru tespit edilmiş varlık ismi sayısının girdi metindeki tespit edilebilecek varlık isimlerinin sayısına oranı ise bulma oranı (recall) olarak adlandırılır **(2.2)**.

$$bulma = \frac{doğru\ tespit\ edilmiş\ varlık\ ismi\ sayısı}{girdi\ metindeki\ varlık\ ismi\ sayısı} \quad \textbf{(2.2)}$$

Tutma ve bulma oranlarının harmonik ortalaması olan f-ölçüt (f-measure) değeri aşağıdaki gibi hesaplanır **(2.3)**.

$$f - \text{ölçüt} = \frac{(\beta + 1) * kesinlik * çağrı}{\beta(kesinlik + çağrı)} \quad \textbf{(2.3)}$$

Eğer tutma ve bulma oranları aynı ağırlığa sahipse harmonik ortalama β=1 alınarak aşağıdaki gibi hesaplanır **(2.4)**.

$$f - \text{ölçüt} = \frac{2 * tutma * bulma}{tutma + bulma} \quad \textbf{(2.4)}$$

VİT sistemlerinin başarımlarının ölçülmesinde kullanılan farklı yöntemler mevcuttur. En çok kullanılan yöntemler CoNLL ve MUC metrikleridir.

MUC metriği MUC konferanslarında kullanılmıştır (Chinchor ve Marsh, 1998). Sistemin başarımını iki açıdan ele alır; doğru sınıf ve doğru sınırlar. Sistem tarafından tespit edilmiş ve sınıflandırılmış varlık isimlerinin sınırlarının doğruluğunu ve sınıflandırılmasının doğruluğunu ayrı ayrı değerlendirir.

CoNLL metriği CoNLL konferanslarında kullanılmıştır (Tjong Kim Sang, 2002; Tjong Kim Sang ve De Meulder, 2003). Sistemin varlık ismi tespitinin doğru kabul edilebilmesi için hem sınırların hem de sınıfın doğru tespit edilmiş olması gerekir. Sistem varlık isminin sınırlarını doğru tespit eder fakat varlık ismini yanlış sınıflandırır ise bu CoNLL metriğine göre yanlış bir çıktı olarak değerlendirilir.

Değerlendirme metriklerinin daha iyi anlaşılması için Şekil 2.4' te örnek bir VİT sistem girdisi, çıktısı ve sistem tarafından yapılan hatalar verilmiştir. Bu örnek sistem çıktısında CoNLL metriğine göre sadece 4. satırdaki çıktı doğrudur, varlık ismi sınırları doğru tespit edilmiş ve varlık ismi doğru sınıflandırılmıştır. Girdi metinde beş tane varlık ismi mevcuttur, çıktıya göre beş tane tespit edilmiş varlık ismi vardır. Bu sistemin CoNLL metriğine göre başarımı %20 olarak hesaplanır (**2.5)**.

$$tutma = {}^{1}/_{5} \qquad bulma = {}^{1}/_{5} \qquad f - \text{ölçüt} = {}^{1}/_{5} = \%20 \quad \textbf{(2.5)}$$

MUC metriğine göre ise sistemin iki adet sınıf kapsamında iki adet de sınırlar kapsamında olmak üzere toplam dört adet doğru çıktısı vardır. Girdi metinde beş tane varlık ismi mevcuttur, sınırlar ve sınıf için iki ayrı değerlendirme yapıldığından girdi metinde 10 tane varlık ismi olduğu varsayılır. Çıktıya göre beş tane tespit edilmiş varlık ismi vardır, bu da sınırlar ve sınıf için düşünüldüğünde toplam 10 varlık ismi olduğu varsayılır. Bu sistemin MUC metriğine göre başarımı %40 olarak hesaplanır (**2.6)**.

$$tutma = {}^{4}/_{10} \qquad bulma = {}^{4}/_{10} \qquad f - \text{ölçüt} = {}^{4}/_{10} = \%40 \quad \textbf{(2.6)}$$

Bu çalışmadaki başarımların ölçülmesinde CoNLL metriği kullanılmıştır. Metriğin sistem çıktılarına uygulanmasında CoNLL ağ sayfasında yer alan betik kullanılmıştır (Url-3, 2014).

	DOĞRU ÇÖZÜM	SİSTEM ÇIKTISI	HATA	HATA TİPİ
1	<b_enamex TYPE="PERSON">Bayram Bayraktar<e_enamex>	Bayram <b_enamex TYPE="PERSON">Bayraktar<e_enamex>	Varlık ismi doğru sınıflandırılmış fakat sınırları yanlış tespit edilmiş	SINIR
2	<b_enamex TYPE="LOCATION">Osmanlı<e_enamex>	<b_enamex TYPE="PERSON">Osmanlı<e_enamex>	Varlık ismi doğru tespit edilmiş fakat yanlış sınıflandırılmış	SINIF
3	Cumhuriyet	<b_enamex TYPE ="ORGANIZATION>Cumhuriyet<e_enamex>	Bir varlık ismi olmadığı halde sistem varlık ismi tespit etmiş	SINIF + SINIR
4	<b_enamex TYPE="LOCATION">Ayvalık<e_enamex>	<b_enamex TYPE="LOCATION">Ayvalık<e_enamex>	Varlık ismi doğru tespit edilmiş	-
5	<b_enamex TYPE="ORGANIZATION">Atatürk Araştırma Merkezi<e_enamex>	<b_enamex TYPE="LOCATION">Atatürk Araştırma Merkezi Yayınları<e_enamex>	Varlık ismi hem yanlış sınıflandırılmış hem de sınırları yanlış tespit edilmiş	SINIF + SINIR
6	<b_enamex TYPE="LOCATION">Ayvalık<e_enamex>	Ayvalık	Varlık ismi sistem tarafından tanınmamıştır	SINIF + SINIR

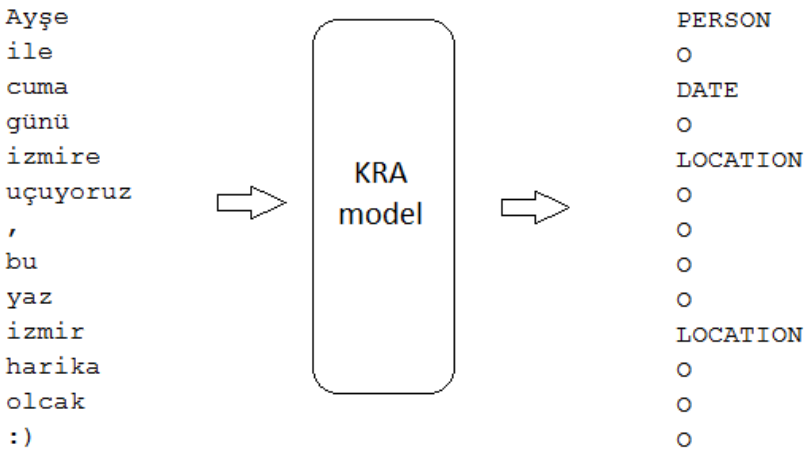
Şekil 2.4 : Örnek bir varlık tanıma sistemi için girdi ve çıktı.

3. KOŞULLU RASTGELE ALANLAR

Koşullu rastgele alanlar sıralı veri sınıflandırılmasında kullanılan bir makine öğrenmesi yöntemidir (Lafferty ve diğ, 2001). KRA sıralı veri sınıflandırmasında sıklıkla kullanılan saklı markov ve maksimum entropi markov modellerine göre daha yüksek başarım elde etmiştir.

Sözcük sınıflandırması, konuşma tanıma, varlık ismi tanıma gibi doğal dil işleme uygulamalarında çoğunlukla sıralı verilerin etiketlenmesine ihtiyaç duyulmaktadır. Bu sıralı veriler çalışmanın türüne göre cümleler, kelimeler, heceler vs. olabilmektedir ve bu veri birimlerine denk gelen etiketler vardır. Etiketleri veri birimlerinin sınıflarını temsil ederler, gözlemlenen verilerin ve etiketlerinin dizilişlerinin olasılıksal bir modeli vardır. Uygulamalardaki amaç bu verilerin ve etiketlerinin dizilimlerini modelleyerek yeni verilerin etiketlerini tahmin etmektir.

Varlık ismi tanıma problemi ele alınırsa, amaç girdi olarak verilen kelime dizisine denk gelen etiketleri tahmin etmektir. Bu etiketler “kişi”, “yer” gibi varlık isimlerine ek olarak varlık ismi olmayan kelimeleri temsil eden “O” etiketinden oluşmaktadır.



Şekil 3.1 : Koşullu rastgele alanlar ile varlık ismi tanıma

Çoğunlukla girdi kelimeler rastegele değil kullanılan dilin yazım kurallarına bağlı olarak dizilirler. Dolayısı ile gözlemlenen kelimelerin kendi arasında ve etiketleri ile

bağınıtıları vardır, örneğin bir kelime organizasyon ismi ise bir sonraki kelimenin de organizasyon ismi olması muhtemeldir.

Gözlemlenen kelimeler $\vec{x} = \{x_1, x_2, \dots, x_n\}$, etiketleri $\vec{y} = \{y_1, y_2, \dots, y_n\}$ ile temsil edilirse amaç gözlemlenen kelime dizisine denk gelen etiket dizilimlerinden olasılığı en yüksek olan etiket dizilimini bulmaktır. KRA $p = (\vec{y}|\vec{x})$ olasılıklarını hesaplar ve en büyüğünü seçer, formulü aşağıdaki gibidir **(3.1)**.

$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad \textbf{(3.2)}$$

$\vec{x}$ girdi olarak verilen kelime dizisini, $\vec{y}$ çıktı dizisini yani etiketleri, j gözlemlenen kelime dizisindeki pozisyonu temsil etmektedir. $f_i(y_{j-1}, y_j, \vec{x}, j)$ nitelik fonksiyonudur, λ_i ağırlık parametresidir. $Z_{\vec{\lambda}}(\vec{x})$ normalleştirme faktörüdür ve aşağıdaki gibi formülize edilebilir **(3.2)**.

$$Z_{\vec{\lambda}}(\vec{x}) = \sum_{\vec{y} \in Y} \exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad \textbf{(3.2)}$$

Bir KRA modeli oluşturmak için öncelikle öznitelik fonksiyonları tanımlanmalıdır, bunlar uygulamanın türüne göre belirlenir, kelimenin kendisine yakınındaki kelimelere, kendi etiketine, yakındaki etiketlere dayanarak oluşturulabilir.

Öznitelik fonksiyonları girdi olarak bir önceki ve şimdiki etiketleri, girdi kelime dizisini ve şimdiki kelime pozisyon bilgisini alırlar, çıktı olarak gerçek bir değer üretirler, genellikle 1 veya 0 olarak ikili değer üretirler. VİT uygulaması için aşağıdaki gibi öznitelik fonksiyon örnekleri verilebilir **(3.3)** ve **(3.4)**.

$$f_1(y_{j-1}, y_j, \vec{x}, j) = \begin{cases} 1, & y_j = LOC \ ve \ x_j = kemalpaşa \\ 0, & aksi \ durumda \end{cases} \quad \textbf{(3.3)}$$

$$f_2(y_{j-1}, y_j, \vec{x}, j) = \begin{cases} 1, & y_{j-1} = ORG \ ve \ y_j = ORG \ ise \\ 0, & aksi \ durumda \end{cases} \quad \textbf{(3.4)}$$

İlk fonksiyon kelime “kemalpaşa” ve etiketi “LOC” yani yer etiketi ise 1 üretmekte, aksi durumda 0 üretmektedir. İkinci fonksiyon ise aktif etiket ve bir önceki etiket “ORG” yani organizasyon etiketi ise 1, aksi durumda 0 üretmektedir.

Öznitelik fonksiyonları oluşturulduktan sonra fonskiyonlara ait ağırlık parametreleri eğitim verisinden hesaplanarak elde edilir, eğitim versisinde söz konusu özniteliğin görülme sayısıyla oranılı bir parametredir. Bu parametrenin pozitif bir değer alması aktif kelimeye ilgili etiketin atanması gerektiğini belirtmektedir, ağırlık parametresinin negatif bir değer alması aktif kelimeye söz konusu etiketin atanmaması yönünde bir işarettir. Örneğin denklem 3.3’ deki fonksiyonun ağırlık parametresi λ_1 0’ dan küçük bir değer almışsa bu demektir ki eğitilen KRA modeli “kemalpaşa” kelimesi gördüğünde “LOC” etiketi atamaktan kaçınacaktır. Örneğin denklem 3.4’ deki ağırlık parametresi λ_2 pozitif ve yüksek bir değer almışsa bu KRA modelinde arka arkaya “ORG” etiketi gelme eğilimindedir, λ_2 parametresi ne kadar yüksekse ilgili fonksiyonun gerçekleşme eğilimi de o kadar yüksektir. Fonkisyonun ağırlık parametresi 0 ise ilgili özniteliğin herhangi bir etkisi yoktur denilebilir.

SMM ve MEMM sıralı verilerin etiketlenmesinde sıkça kullanılmış ve başarımı ispatlanmış yöntemlerdir, KRA ise bu yöntemlerin bazı açıklarını kapatarak daha yüksek başarım elde etmiştir.

SMM yönteminde gözlemlenen durumların sadece bir önceki duruma bağlı olduğu varsayılmaktadır, oysaki gözlemlenen durumlar kendi aralarında farklı faktörlere de bağlı olabilmektedirler. SMM birleşik olasılıksal yapıdayken KRA bu durumun üstesinden koşullu olasılıksal yapısı ise gelmektedir.

4. GELİŞTİRİLEN YÖNTEM

Bu çalışmada tweet metinlerinde varlık isimlerini tespit edecek bir model geliştirilmiştir. Modelin oluşturulmasında koşullu rastgele alanlar yöntemi kullanılmıştır.

Geliştirilen sistem kişi, yer, organizasyon, para, tarih, saat, yüzde varlık tiplerinin tespitini yapabilmektedir.

Sistemin geliştirilmesinde haber metinleri ve tweet metinleri olmak üzere iki farklı veri kümesi kullanılmıştır. Tüm veri kümesi IOB2 formatı ile etiketlenmiş, eğitim ve testler bu formatta yapılmıştır.

4.1 Veri Kümesi Özellikleri

Kullanılan haber metinler Tür ve diğ. (2003) çalışmasında kullanıdığı veri kümesinden alınmıştır. Kullanılan tweet metinlerinin bir kısmı Çelikkaya ve diğ. (2013) çalışmasından alınmıştır, bir kısmı ise bu çalışmada kullanılmak üzere Twitter' dan toplanarak etiketlenmiştir.

Veri kümelerinin analizi Çizelde 4.1 ve 4.2' deki gibidir.

Çizelge 4.1 : Haber veri kümesinin analizi.

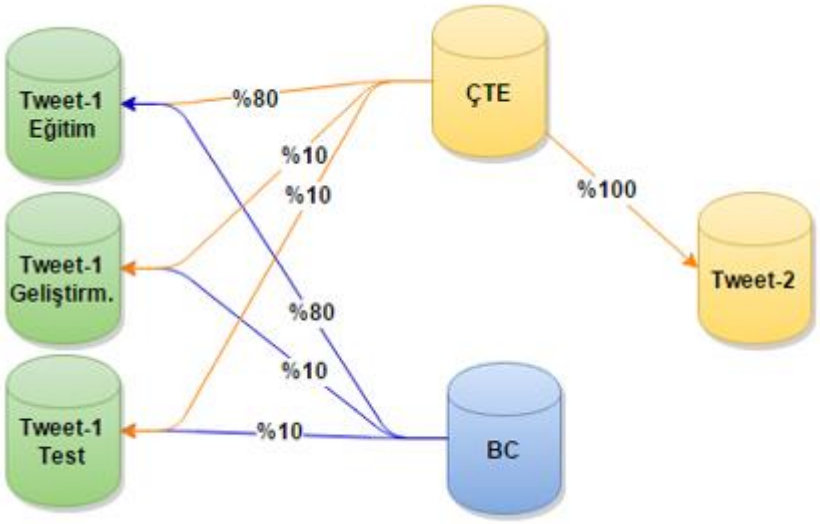
Eleman	Eğitim	Test	Toplam
Sözcük sayısı	444.475	47.343	491.818
Tekil sözcük sayısı	66.202	14.798	81.000
Toplam varlık ismi sayısı	38.388	3.579	36.967
Kişi	14.492	1.598	16.090
Yer	10.538	1.177	11.715
Organizasyon	8.358	804	9.162

Çizelge 4.1' de haber metinlerindeki varlık isimlerinin sayıları ve sözcük sayısı verilmiştir. Veri kümesi yaklaşık %90' nı eğitim amaçlı, geri kalanı test amaçlı kullanılmak üzere iki parçaya ayrılmıştır.

Çizelge 4.2 : Tweet veri kümesinin analizi.

Eleman	Eğitim	Geliştirme	Test	Toplam
Sözcük sayısı	99.459	21.374	21.300	142.133
Tekil sözcük sayısı	39.758	11.068	10.171	60.997
Tweet sayısı	10.268	2.203	1.930	14.101
Toplam varlık ismi sayısı	4.097	1.314	901	6.412
Kişi	1.614	476	429	2.519
Yer	906	266	172	1.344
Organizasyon	1.215	463	236	1.914
Zaman	66	12	21	99
Tarih	199	49	31	279
Para	47	38	10	95
Yüzde	49	10	2	61

Çizelge 4.2' de tweet veri kümesindeki tweet sayısı, varlık isimlerinin sayıları ve sözcük sayısı verilmiştir. Veri kümesi yaklaşık %80' nı eğitim amaçlı, geri kalanı ise %10 model geliştirme, %10 test amaçlı kullanılmak üzere toplam üç parçaya ayrılmıştır. Tweet veri kümesindeki 5043 tweet Çelikkaya ve diğ. (2013) çalışmasından alınmış, 9058 tweet bu çalışma kapsamında etiketlenmiştir. Tweet veri kümesinin kombinasyonlarının daha net anlaşılması için Şekil 4.1 incelenebilir.



Şekil 4.1 : Tweet veri kümesi.

ÇTE ile gösterilen veri kümesi Çelikkaya ve diğ. (2013) çalışmasından alınmış 5043 tweeti, BC ile gösterilen veri kümesi ise bu çalışma kapsamında etiketlenen 9058 tweeti temsil etmektedir.

Çalışmada üç adet sözlük kullanılmıştır, bunların içerdiği varlık sayıları ise Çizelge 4.3' de verilmiştir. Kişi sözlüğü internetten toplanan Türkçe isimleri içermektedir. Yer

sözlüğü dünya ülkelerinin isimlerini, Türkiye' nin il, ilçe, mahalle, köy gibi birimlerinin isimlerini içermektedir. Organizasyon sözlüğü ise "üniversite", "dernek", "bakanlık" gibi organizasyon isimlerinde kullanılan kelimeleri içermektedir.

Çizelge 4.3 : Sözlüklerin analizi.

Sözlük	Sözcük Sayısı
Kişi	8.674
Yer	24.709
Organizasyon	67

4.2 Çalışma Adımları

Çalışmanın ilk adımı olarak kullanılacak veriler toplanmış ve etiketlenmiştir. Tweetlerin elde edilmesi aşamasında Twitter API adlı platform (Url-4) kullanılmıştır. Verileri etiketleme aşamasında Küçük ve Yazıcının (2009) etiketleme aracı kullanılmıştır. Tweet metinleri etiketlendikten sonra %80' i eğitim, %10' u geliştirme ve %10' u test amaçlı olmak üzere üç parçaya ayrılmışlardır.

Modelin oluşturulmasında öncelikle veriler sözcüklerine ayrılmıştır ve daha sonra bu sözcüklere ait öznitelikler elde edilerek koşullu rastgele alanlar yöntemiyle modelin eğitilmesinde kullanılmıştır. Oluşturulan model ile test verisi test edilmiş ve sistem başarımı CoNLL metriği ile ölçülmüştür.

Model oluşturma ve sınama aşamasında verilerin gösteriliminde IOB2 gösterim yöntemi kullanılmıştır.

İki farklı metin türünde çalışıldığı için öznitelik çıkarımı açısından iki farklı yöntem izlenerek farklı modeller oluşturulmuştur. İlk model düzgün formatta yazılmış metinler, ikinci model ise kısa ve düzgün formatta olmayan tweet metinleri hedef alınarak geliştirilmiştir.

İlk model kelimelerin biçimbilimsel özelliklerine dayanarak eğitilmiştir. Bu yöntemde girdi metin sözcüklerine ayrıştırıldıktan sonra biçimbilimsel çözümleme ve belirsizlik giderme aşamalarına tabi tutulmuştur. Bu aşamalarda biçimbilimsel analiz, biçimbilimsel belirsizlik giderme ve sözdizimsel ayrıştırma fonksiyonlarını bünyesinde toplayan Eryiğit ve diğerleri (2008) tarafından geliştirilen araç kullanılmıştır (Eryiğit, 2012).

Bu aşamalar sonucunda metindeki kelimelere ait biçimbilimsel özellikler elde edilmiştir, bu elde edilen özellikler daha sonra KRA yöntemiyle model oluşturulmasında kullanılmıştır.

Geliştirilen yöntemde işlem adımları şu şekildedir;

- Girdi metnini sözcüklerine ayırma
- Biçimbilimsel açıdan çözümleme
- Biçimbilimsel belirsizlik giderme
- Diğer özniteliklerin biçimbirimsel özniteliklere eklenmesi
- KRA ile modelin eğitilmesi
- Hazırlanan özniteliklerin eğitilmiş model ile test edilmesi ve sonuçların elde edilmesi

Bu modelle haber metinlerinde yüksek başarım elde edilmesine rağmen tweet metinlerinde başarımı düşmektedir. Tweetler düzgün formatta yazılmadıkları için biçimbirimsel çözümleme ve belirsizlik giderme aşamalarında yanlış sonuçlar vermektedir. Başarımı yükseltmek için tweet metinlerinin normalizasyon aşamasından geçirilerek düzgün metin formatına getirilmesi kullanılan ve etkili bir yöntemdir fakat bu fazladan eklenen aşamalar zaman maliyetini de yükseltmektedir.

Bunun yerine ikinci bir yöntem olarak biçimbilimsel özniteliklere alternatif olabilecek öznitelliklerin seçilmesi düşünülmüştür. Örneğin kelimenin kökü yerine kelimenin ilk 3 veya 4 harfi kullanılmıştır. Kelimenin son harfleri de biçimbilimsel özniteliklere yakın bilgiler taşımakta olduğu için son 3 veya 4 harf bilgisi de öznitelik olarak seçilmitir. Bunun yanında bir kelimenin kesme işareti ile ayrılan bir ek içeriyor olması bu kelimenin bir özel isim dolayısı ile bir varlık ismi olduğu yönünde kuvvetli bir göstergedir.

Geliştirilen ikinci yöntemde ise işlem adımları şu şekilde olmaktadır;

- Girdi metnini sözcüklerine ayırma
- Özniteliklerin çıkarılması
- KRA ile modelin eğitilmesi

- Hazırlanan özniteliklerin eğitilmiş model ile test edilmesi ve sonuçların elde edilmesi

Bu yöntemde biçimbilimsel özelliklere yer verilmemiştir, dolayısıyla ilk yönteme göre işlem ve zaman maliyeti daha düşüktür. Bu yöntemdeki başarımın haber metinlerinde ilk yöntemdeki kadar yüksek olduğu gözlemlenmiştir. Tweet metinlerinde ise ilk modele göre daha yüksek olduğu gözlemlenmiştir.

4.3 Kullanılan Öznitelikler

Modelin oluşturulmasında verilerin özelliklerinden yararlanılır. Bu çalışmada metindeki kelimelerin biçimbilimsel özelliklerinden, kelimenin başlangıcındaki ve bitişindeki belli sayıdaki harflerden, harflerin küçük/büyük olma özelliklerinden ve sözlükte var olma olmama özelliklerinden yararlanılmıştır. Hepsi listelenecek olursa;

- Biçimbilimsel özellikler
    - Kelimenin kendisi
    - Kelimenin kökü
    - Kelimenin türü (zamir, isim, vs.)
    - İsim hali (yalın hali, -e hali, -de hali, -den hali, vs.)
    - Özel isim olup olmaması
    - Çekim ekleri
- Kelimenin kısımları
    - İlk 4 harfi
    - Son 4 harfi
- Kesme işareti içerip içermemesi
    - Kesme işaretinden sonraki ek
- Büyük küçük harf durumu
    - Kelimenin içerdiği harflerin büyük/küçük olma durumu
    - Bulunduğu cümlenin ilk kelimesi olması durumu
- Sözlük özellikleri

- Kişi isimleri sözlüğünün ilgili kelimeyi içermesi
- Yer isimleri sözlüğünün ilgili kelimeyi içermesi
- Organizasyon isimleri sözlüğünün ilgili kelimeyi içermesi

Biçimbilimsel özellikler biçmbilimsel çözümleme aşamasından elde edilmektedir. Kullanılan araç çıktı olarak her bir kelimeye ait kök, sözcük türü, isim hali, çekim eklerini verir. Örnek bir girdi ve çıktı Şekil 4.2' deki gibidir.

```
25              25+Num+Card
-               -+Punc
30              30+Num+Card
yaşında         yaş+Noun+A3sg+P3sg+Loc
olan            ol+Verb+Pos^DB+Adj+PresPart
bu              bu+Det
abide           abide+Noun+A3sg+Pnon+Nom
ağaçların       ağaç+Noun+A3pl+Pnon+Gen
yerine          yer+Noun+A3sg+P3sg+Dat
yenileri        yeni+Adj^DB+Noun+Zero+A3pl+P3pl+Nom
dikilse         dikil+Verb+Pos+Desr+A3sg
bile            bile+Adverb
pek             pek+Adverb
çoğumuz         çoğu+Pron+Quant+A1pl+P1pl+Nom
onların         o+Pron+Demons+A3pl+Pnon+Gen
büyüdüklerini   büyü+Verb+Pos^DB+Noun+PastPart+A3sg+P3pl+Acc
göremeyeceğiz   gör+Verb^DB+Verb+Able+Neg+Fut+A1pl
.               .+Punc
```

Şekil 4.2 : Örnek biçimbilimsel aşama çıktısı.

Elde edilen biçimbilimsel bilgiler modelin eğtim aşamasında öznitelik olarak kullanılmışlardır.

Kişi, yer, organizasyon gibi varlık isimleri özel isimlerden oluşmaktadır. Türkçede kelime büyük harfle başlıyorsa bu kelime bir özel isim demektir. Dilin bu özelliği varlık ismi tespiti için önemli bir ipucu olmaktadır fakat tek başına yeterli değildir çünkü özel isimlerin yanında ünvanlar, kısaltmalar, vs. başka kelimeler de büyük harfle başlayabilmektedir. Büyük harfle başlama özelliğinin en çok kullanıldığı başka bir yer özel isim olsun veya olmasın cümlelerin ilk kelimesidir. Daha başarılı bir model oluşturmak için eğitim aşamasında kelimelerin büyük/küçük harf bilgisi ve ilgili kelimenin cümlenin ilk kelimesi olması bilgisi birlikte kullanılmıştır.

Tweet metinleri gibi kısa metinler imla, dilbilgisi kurallarına dikkat edilmeden yazıldığı için biçimbilimsel öznitelikler doğru bir şekilde edilememektedir, dolayısıyla biçimbilimsel öznitelikler kullanılarak oluşturulan modelde yüksek bir başarım elde edilememektedir. Modelin oluşturulmasında kelimelerin biçimbilimsel özellikleri yerine kelime hakkında önemli bilgiler taşıyan ilk 3 veya 4 harfi, son 3 veya 4 harfi, kesme işareti içerip içermediği, içeriyorsa kesme işaretinden sonra gelen ek bilgisi kullanılmıştır. Kelimenin ilk 3 veya 4 harfi kelimenin kök bilgisine alternatif olarak kullanılmıştır. Yapılan sınamalarda ilk 3-4 harf ve kökün başarıma etkilerinin birbirine yakın olduğu gözlemlenmiştir.

Türkçede özel isimlere gelen ekler kesme işareti ile ayrılırlar, dolayısıyla bir kelimeden sonra kesme işareti ve bir ek geliyor olması da varlık ismi olması ile ilgili önemli bir ipucu içermektedir. Kelimenin kesme işareti içerip içermediği, içeriyorsa kesme işaretinden sonra aldığı ek ve kesme işaretinden önceki son 3-4 harf bilgisi öznitelik olarak kullanılmıştır.

Eğitim verisinde bulunmayan varlık isimlerinin tespitinde geniş isim sözlükleri önemli rol oynamaktadırlar. Girdi metindeki kelimenin varlık ismi olup olmadığının tespiti sözlüklere bakılarak yapılabilir, aktif kelime hangi sözlükte var ise o sözlük türüne ait bir varlık ismidir denilebilir.

Bu çalışmada üç tür sözlük kullanılmıştır;

- Kişi isimleri
- Yer isimleri
- Organizasyon isimleri ve organizasyon isimlernde kullanılan anahtar kelimeler (üniversite, bakanlık, dernek vb.)

Tweetler genellikle güncel konular üzerine yazıldıkları için bu sözlüklerin yeni türeyen kelimeleri içerecek şekilde genişletilmesi sistemin başarımına katkı sağlaycaktır.

Bunun yanında tweet metinleri kasıtlı ve kasıtsız yazım hatalarını çokça içerdiklerinden ötürü sözlük dosyaları ile birebir eşleşme yerine kısmi eşleşme yöntemi kullanılmıştır. Örneğin girdi metninde bulunan “mehmt” kelimesi aslında Türkçedeki “Mehmet” ismini kastetmektedir. “Mehmet” ile “mehmt” arasında sadece 1 harflik bir fark vardır, isim sözlüğünde “Mehmet” kelimesi vardır fakat “mehmt”

kelimesi sözlükte olmadığı için sistem tarafından isim olarak etiketlenme ihtimali düşmektedir. Bunun üstesinden gelmek amacıyla girdi kelime ile sözlükteki kelimeler arasındaki farklar ölçülmüş ve bu farklar arasından en küçük olanı öznitelik olarak model eğitiminde kullanılmıştır. Kelimeler arası uzaklıkların ölçümünde en bilindik algoritmalardan olan Levenshtein mesafe algoritması kullanılmıştır (Levenshtein, 1966). Bu algoritma iki kelime arasındaki benzerliği hesaplar, farklı bir deyişle kelimelerden birini diğerine dönüştürebilmek için gereken işlem sayısını bulur. İşlemden kastedilen harf ekleme, silme ve değiştirmektir. Ekleme ve silme 1 birim, değiştirme 2 birim mesafe olarak düşünülmüştür. Örneğin "Mehmet" ile "mehmt" kelimelerinin uzaklığı 1' dir, "İstanbul" ve "İstnbl" kelimelerinin uzaklığı ise 2' dir.

Girdi olarak verilen kelimenin üç sözlükteki her bir kelime ile uzaklığı hesaplanır ve aralarından en küçük olanları öznitelik olarak seçilir. Sözlüklerdeki her bir kelime ile karşılaştırma yapmanın zaman ve işlem maliyeti yüksek olduğu ve yapılan yazım hatalarında çoğunlukla ilk harf doğru yazıldığı için sadece ilk harfleri aynı olan kelimeler karşılaştırılır. Örneğin "mehmt" kelimesi sadece sözlüklerdeki "M" ile başlayan kelimelerle karşılaştırılır.

Tweet metinlerindeki bir diğer yazım kurallarına aykırı durum Türkçe karakterlerin kullanılmamasıdır. Örneğin "çatıdaki güvercin" yerine "catidaki guvercin" yazılabilmektedir. Böylesi durumların üstesinden gelebilmek için tüm veri kümesindeki Türkçe karakterler İngiliz alfabesindeki karşılıklarına çevrilmiştir.

Tweet verilerinde ayrıca nida, heyecan belirtme amacıyla kasıtlı olarak bazı harfler tekrarlı bir şekilde kullanılabilmektedir. Örneğin "geliyor" yerine "geliyooooooooo" yazılabilmektedir. Bunun üstesinden gelmek için kelime içinde ikiden fazla tekrarlanan karaterlerin sayısı ikiye indirilmiştir, bir harf yerine iki harfe indirilme sebebi ise dikkat, şiir, vb. yan yana aynı harfin kullanıldığı kelimeleri bozmamaktır.

4.4 Model eğitimi ve sınanması

Modelin oluşturulmasında verilerin özelliklerinden yararlanılır. Bu çalışmada model oluşturulmasında ve sınanmasında CRF++ aracı (Url-5) kullanılmıştır. CRF++ açık kaynak kodlu bir araç olup koşullu rastgele alanlar algoritmasının uyarlamasıdır. Model eğitimi ve sınanmasında kullanılmıştır.

Bu aracı kullanabilmek için sözcüklerine ayrılmış, öznitelikleri çıkarılmış eğitim ve sınama verileri formatları aynı olan iki dosya halinde hazırlanmalıdır, dosya formatına örnek olarak Şekil 4.3 verilebilir.

```
iyi iyi iyi 0 0 1 2 1 4 3 3 0 O
geceler gece eler 0 0 0 1 3 4 4 0 0 O
..:d.:d..:d.:d ..:d d.:d 0 4 0 1000 1000 1000 0 0 0 O

dacianın daci anın 0 2 1 3 4 5 0 0 0 O
roberto robe erto 0 0 0 4 4 7 0 0 0 B-PERSON
carloslu carl oslu 0 0 0 3 4 7 4 3 0 I-PERSON
reklamı rekl lamı 0 0 0 4 4 6 3 3 3 O
çok çok çok 0 0 0 1 2 1000 3 3 0 O
güzel güze üzel 0 0 0 0 1 3 10 3 0 O
:d :d :d 0 4 0 1000 1000 1000 0 0 0 O
#efsanerobertocarlos #efs rlos 0 4 0 1000 1000 1000 0 0 0 O

evcil evci vcil 0 2 1 1 2 5 5 4 0 O
katil kati atil 0 0 0 1 1 3 4 4 0 O
balinam. bali nam. 0 4 0 3 3 5 4 4 0 O
arabesk arab besk 0 2 0 2 2 4 4 3 0 O
rap rap rap 0 0 0 1 1 8 0 0 0 O
yapan yapa apan 0 0 0 1 1 5 4 3 0 O
kertenkeleyle kert eyle 0 0 0 6 7 9 5 3 0 O
takas taka akas 0 0 0 1 1 5 4 4 0 O
edilir edil ilir 0 0 0 2 3 5 4 3 0 O

karabükspor kara spor 0 2 1 4 6 8 5 5 0 B-ORGANIZATION
'dan 'dan 'dan 1 4 0 1000 1000 1000 0 0 0 O
bursaspor burs spor 0 2 0 4 5 7 5 3 0 B-ORGANIZATION
'a 'a 'a 1 4 0 1000 1000 1000 0 0 0 O
kınama kına nama 0 0 0 2 2 4 4 4 0 O
```

Şekil 4.3 : Örnek eğitim ve sınama dosyası içeriği.

Bu dosyalarda CRF++ formatına göre cümleler arasında birer satır boşluk bırakılmıştır. Her bir satırdaki ilk simge sözcüğün kendisini, son simge ise sözcüğe ait etiketi (kişi, yer, organizasyon) temsil etmektedir. Bunlar arasında kalan simgeler ise özniteliklerdir. Sözcük, öznitelikleri ve etiketi arasında format gereği birer boşluk bulunmalıdır.

Model oluşturulurken hangi özniteliklerin ne şekilde kullanılacağını CRF++ aracına tarif etmek gerekir, bunun için şablon dosyaları kullanılır. Model eğitilirken aktif sözcüğün öncesi ve sonrasındaki sözcükler ve onların özellikleri de dikkate alınır.

Örneğin bir sonraki sözcüğün kesme işareti içermesi o sözcüğün varlık ismi olma ihtimalini artırır ya da ardarda gelen sözcükleri aynı etiket türünden olma olasılığı fazladır. Dolayısı ile modelin oluşturulmasında sözcükten iki önceki ve iki sonraki sözcüklerin de özellikleri dikkate alınmıştır. Örnek bir şablo dosyası içeriği Şekil 4.4' deki gibidir. Şablonlar sözcüklerin bulunduğu veri dosyasındaki kolon ve satırları temsil eden iki boyutlu dizilerden oluşur.

```
# Unigram
U01:%x[-2,0]
U02:%x[-1,0]
U03:%x[0,0]
U04:%x[1,0]
U05:%x[2,0]
U08:%x[-2,0]/%x[-1,0]
U09:%x[-1,0]/%x[0,0]
U10:%x[0,0]/%x[1,0]
U11:%x[1,0]/%x[2,0]
U14:%x[-2,0]/%x[-1,0]/%x[0,0]
U15:%x[-1,0]/%x[0,0]/%x[1,0]
U16:%x[0,0]/%x[1,0]/%x[2,0]
U17:%x[-2,0]/%x[-1,0]/%x[0,0]/%x[1,0]
U18:%x[-1,0]/%x[0,0]/%x[1,0]/%x[2,0]
U19:%x[-2,1]
U20:%x[-1,1]
U21:%x[0,1]
U22:%x[1,1]
U23:%x[2,1]
U26:%x[-2,1]/%x[-1,1]
U27:%x[-1,1]/%x[0,1]
U28:%x[0,1]/%x[1,1]
U29:%x[1,1]/%x[2,1]
U32:%x[-2,1]/%x[-1,1]/%x[0,1]
U33:%x[-1,1]/%x[0,1]/%x[1,1]
U34:%x[0,1]/%x[1,1]/%x[2,1]
U35:%x[-2,1]/%x[-1,1]/%x[0,1]/%x[1,1]
U36:%x[-1,1]/%x[0,1]/%x[1,1]/%x[2,1]
U37:%x[-2,2]
                .
                .
                .
                .
                .
                .
U116:%x[-1,7]
U117:%x[0,7]
U118:%x[1,7]
U119:%x[2,7]
U129:%x[-2,8]
U130:%x[-1,8]
U131:%x[0,8]
U132:%x[1,8]
U133:%x[2,8]
U134:%x[-2,9]
U135:%x[-1,9]
U136:%x[0,9]
U137:%x[1,9]
U138:%x[2,9]
U139:%x[-2,10]
U140:%x[-1,10]
U141:%x[0,10]
U142:%x[1,10]
U143:%x[2,10]
U144:%x[-2,11]
U145:%x[-1,11]
U146:%x[0,11]
U147:%x[1,11]
U148:%x[2,11]
U149:%x[0,6]/[0,9]
U150:%x[0,7]/[0,10]
U151:%x[0,8]/[0,11]

# Bigram
B
```

Şekil 4.4 : Örnek şablon dosyası.

Bu dosyada her bir satır farklı bir şablondur, şablonlarda yer alan dizilerin ilk elemanı sütunu, ikinci elemanı ise öznitelik kolonunu temsil eder. Örneğin "U20:%x[-1,1]" şablonu aktif sözcükten bir önceki sözcüğe ait birinci özniteliği temsil eder. Şekil 4.5' te verilen örnek sistem girdisine göre 3. satırdaki "en" kelimesi aktif olduğunda "U20:%x[-1,1]" şablonu "'ün" özniteliğini gösterecektir. "U03:%x[0,0]" kelimenin

kendisini yani “en”, “U29:%x[1,1]/%x[2,1]” şablonu ise “güçl/10” şeklinde iki özniteliğin birleşimini gösterecektir.

Şablon dosyasına göre CRF++ aracı ile oluşturulan çıktıya örnek olarak Şekil 4.5 verilebilir. Bu dosyada her satırda kelime ve kelimeye ait öznitelikler yer almaktadır, eğitim ve sınamada kullanılabilirler.

```
2014     2014 2014 0 4 1 1000 1000 1000 B-DATE
'ün      'ün  'ün  1 4 0 1000 1000 1000 O
en       en   en   0 0 0 2    1    5    O
güçlü    güçl üçlü 0 0 0 0    0    4    O
10       10   10   0 4 0 1000 1000 1000 O
ordusu!  ordu usu! 0 4 0 2    3    5    O
türkiye  türk kiye 0 2 0 0    0    5    B-LOCATION
kaçıncı  kaçı ıncı 0 0 0 2    3    6    O
sırada?  sıra ada? 0 4 0 2    3    5    O
```

Şekil 4.5 : Örnek sistem girdisi.

Sınama aşamasının çıktısı, eğitilmiş modelin tahminlerinin ve doğru etiketlerin yer aldığı bir dosyadır, bu çıktıların başarımı CoNLL metriği (Url-3) ile ölçülmüştür.

5. DEĞERLENDİRME

Sistem alana özgü ve alan dışı olmak üzere iki açıdan değerlendirilmiştir. İlk olarak alana özgü değerlendirilme yapılmıştır. Haber metinleri haber metinleri ile eğitilmiş model üzerinde sınanmıştır, tweet metinleri ise tweet metinleri ile eğitilmiş model üzerinde sınanmıştır. İkinci olarak alan dışı değerlendirme olarak tweet metinler haber metinleri ile eğitilmiş model üzerinde sınanmıştır.

Asıl amaç tweet metinlerinde varlık ismi tanımayı geliştirmektir, fakat kullanılan yöntemlerin etksini ve sonuçları karşılaştırmak amacıyla çalışmada haber metinleri de kullanılmıştır.

Çalışmada iki yönteme dayanarak modeller oluşturulmuştur;

- Temel model
- Geliştirilen model

Temel model Şeker ve Eryiğit' in (2012) çalışmasında kullanılan yönteme dayanır. Geliştirilen model ise temel modelde kullanılan bazı yöntemlerin yerine tweet metinlerinin doğasına daha uygun olacak yeni yöntemlerin kullanıldığı modeldir.

Sistem performansı CoNLL metriği ile ölçülmüştür.

5.1 Alana Özgü Değerlendirme

İlk olarak haber metinleri ile temel model oluşturulmuş ve oluşturulan model sınanmıştır. Bu model kabaca biçimbilimsel özellikler, büyük/küçük harf bilgisi, cümle başlangıcı bilgisi ve sözlük bilgisi kullanılarak geliştirilmiştir.

İkinci olarak tweet metinleri ile temel yönteme dayalı bir model oluşturulmuştur. Tweetler biçimbilimsel aşamadan geçirilerek biçimbilimsel özellikleri elde edilmiş ve buna büyük/küçk harf bilgisi, cümle başlangıcı bilgisi, sözlük bilgileri eklenerek temel model oluşturulmuştur. Tweet metinleri ile oluşturulan bu temel modelin performansının düştüğü görülmüştür. Bunun en büyük sebebi tweet metinlerinin kuralsız doğasıdır. Tweet metinlerinin biçimbilimsel çözümleme aşamasında

performansı düşmektedir, dolayısı ile bu tüm sistemin başarımını da olumsuz yönde etkilemektedir. Örnek bir tweetin biçimbilimsel analiz aşama çıktısı ve olması gereken çıktı ile karşılaştırılması Şekil 5.1' de verilmiştir.

Sözcük	Biçimbilimsel analiz sonucu	Düzeltilmiş sözcük	Biçimbilimsel analiz sonucu
@pulmonerdamar	@pulmonerdamar+Noun+Prop +A3sg+Pnon+Nom	@pulmonerdamar	@pulmonerdamar+Noun+Prop +A3sg+Pnon+Nom
kanki	kanki+Noun+Prop+A3sg+Pnon +Nom	kanka	kanka+Noun+A3sg+Pnon+No m
ben	ben+Pron+Pers+A1sg+Pnon+N om	ben	ben+Pron+Pers+A1sg+Pnon+ Nom
de	de+Conj	de	de+Conj
senin	sen+Pron+Pers+A2sg+Pnon+G en	senin	sen+Pron+Pers+A2sg+Pnon+G en
tipini	tip+Noun+A3sg+P3sg+Acc	tipini	tip+Noun+A3sg+P3sg+Acc
coh	coh+Noun+Prop+A3sg+Pnon+ Nom	çok	çok+Adverb
seviyoom	seviyoom+Noun+Prop+A3sg+P non+Nom	seviyorum	sev+Verb+Pos+Prog1+A1sg
:D	:D+Noun+Prop+A3sg+Pnon+N om	:D	:D+Noun+Prop+A3sg+Pnon+N om

Şekil 5.1 : Bir tweetin ve düzeltilmiş versiyonunun biçimbilimsel aşama çıktısı.

Bir tweetin biçimbilimsel aşama çıktısının düzgün olabilmesi için düzeltilmesi yani normalize edilmesi gerekmektedir. Normalizasyon ciddi ve kompleks bir aşamadır, bu çalışmada kompleks bir normalizayon ve biçimbilimsel aşamanın kullanılmadığı bir yöntem geliştirilmiştir. Bu yöntemde biçimbilimsel aşama yerine kelimenin ilk ve son N harfi kullanılmış, veri kümesindeki Türkçe karakteler ASCII tablosu karşılığına dönüştürülmüş, sözlük kontollerinde farklı yöntemler kullanılmıştır.

Teme yöntem ve geliştirilen yöntem her iki veri türünde de denenmiştir. Temel yöntemde Şeker ve Eryiğit' in (2012) elde ettiği sonuçlara yakın sonuçlar elde edilirken tweet metinlerinin başarımı düşmüştür. Geliştirilen yöntemde haber metinlerinde başarım neredeyse aynı kalırken tweeet metinlerinde başarım temel yönteme göre yaklaşık %6 yükselmiştir. Alana özgü değerlendirme sonuçları Çizelge 5.1' de verilmiştir.

Çizelge 5.1 : Alana özgü değerlendirme sonuçları.

Model	Haber Test Seti	Tweet Test Seti
Temel model	90.38	55.29
Geliştirilen model	90.23	61.05
Başarım değişimi (%)	- 0.15	+ 5.76

Çizelge 5.1' göre geliştirilen model temel modelin başarımını yakalamıştır dolayısıyla temel modelde kullanılan biçimbilimsel araçlara ve aşamalara gerek olmadan aynı başarım elde edilebilmektedir.

Temel yöntemin haber metinlerinde alana özgü sonuçları Çizelge 5.2' de verilmiştir.

Çizelge 5.2 : Haber metinlerinde temel modelin başarımı.

Model	Haber Test Seti
Sadece kök	83.36
Sözcüğün kendisi	82.20
Sözcüğün kendisi (küçük harf)	82.93
+Kök	84.30
+Sözcük türü	84.85
+İsim hali	85.59
+Özel isim etiketi	86.91
+Çekim ekleri	87.14
+Büyük/Küçük harf	90.01
+Cümle başlangıcı mı	89.91
+Sözlük (tam eşleşme)	90.38
+Sözlük (kısmi eşleşme)	**90.47**

Çizelgedeki her bir satır farklı özniteliklerle eğitilmiş farklı modellerin başarımlarını göstermektedir. Özniteliklerin başarıma etkisini karşılaştırabilmek için öznitelikler sırayla eklenerek eğitim ve sınama aşamaları tekrarlanmıştır.

İlk satırda kelimenin sadece kökü kullanıılarak oluşturulan bir modeldeki başarımlar gösterilmektedir, ikinci satırda ise sadece sözcüğün kendisi kullanılarak oluşturulan bir başka modeldeki başarımlar gösterilmektedir. Üçüncü satırda metindeki tüm harfler küçük harfe çevrilerek ve sadece sözcüğün kendisi kullanılarak oluşturulmuş bir modeldeki modeldeki başarımlar gösterilmektedir. Bundan sonraki satırlarda yer alan öznitelikler bir öncekine eklenerek yeni modeller eğitilmektedir, satır başındaki "+" işareti bunu göstermektedir. En yüksek başarım tüm özniteliklerin kullanıldığı modelde elde edilmiştir ve en alttaki satırda kalın yazı tipi ile belirtilmiştir.

Biçimbilimsel öznitelikler yerine kelimenin ilk ve son 4 harflerinin kullanıldığı geliştirilen yöntemdeki modelin haber metinlerindeki başarımları Çizelge 5.3' dedir.

Çizelge 5.3 : Haber metinlerinde geliştirilen modelin başarımı.

Model	Haber Test Seti
Sözcüğün kendisi (küçük harf)	82.93
+İlk 4 karakter	84.07
+Son 4 karakter	85.34
+Kesme işareti	86.11
+Büyük/Küçük harf	89.41
+Cümle başlangıcı mı	89.50
+Sözlük (tam eşleşme)	90.17
+Sözlük (kısmi eşleşme)	**90.23**

Çizelge 5.2 ve 5.3' deki sonuçlar göstermektedir ki biçmbilimsel özellikler yerine kelimenin baştan ve sondan belli sayıda harfini kullanarak haber metinlerinde aynı başarım yakalanabilmektedir. Öyleyse VİT sisteminde düzgün kurallı yazılmış metinlerde biçimbilimsel analiz ve belirsizlik giderme aşamaları kullanılmadan da yüksek performans elde edilebilir.

Temel yöntemin tweet metinlerinde alana özgü sonuçları Çizelge 5.4' de verilmiştir.

Çizelge 5.4 : Tweet metinlerinde temel modelin başarımı.

Model	Tweet Geliştirme Seti	Tweet Test Seti
Sadece kök	45.88	44.12
Sözcüğün kendisi (küçük harf)	44.84	43.51
+Kök	48.73	47.98
+Sözcük türü	49.48	48.08
+İsim hali	49.08	47.82
+Özel isim etiketi	48.67	49.29
+Çekim ekleri	47.70	46.44
+Büyük/Küçük harf	50.85	51.80
+Cümle başlangıcı mı	50.21	52.39
+Sözlük (tam eşleşme)	**56.81**	**55.29**
+Sözlük (kısmi eşleşme)	56.49	55.25

Çizelge 5.4' de görüdüğü gibi tweet test setinde en yüksek başarıma sözlük bilgisinin tam eşleşme ile elde edildiği modelde ulaşılmıştır.

Geliştirilen yöntemin tweet metinlerindeki başarımları ise Çizelge 5.5' tedir.

Çizelge 5.5 : Tweet metinlerinde geliştirilen modelin başarımı.

	NORMAL		ASCIIFIED	
Model	Tweet Geliştirme Seti	Tweet Test Seti	Tweet Geliştirme Seti	Tweet Test Seti
Sözcüğün kendisi (küçük harf)	47.30	43.51	49.50	44.56
+İlk 4 karakter	52.90	49.42	50.22	49.70
+Son 4 karakter	50.37	50.65	51.06	51.03
+Kesme İşareti	52.96	51.94	52.64	52.88
+Büyük/Küçük harf	58.10	57.79	59.32	58.75
+Cümle başlangıcı mı	58.43	58.01	58.92	58.49
+Sözlük (tam eşleşme)	61.40	59.72	**60.92**	**61.05**
+Sözlük (kısmi eşleşme)	**61.72**	**62.15**	59.11	59.87

Çizelgedeki "Asciified" sütunları veri kümesindeki Türkçe karakterlerin ASCII tablosundaki karşılıklarına çevrilmiş haliyle oluşturulmuş modeldeki başarımı göstermektedir. Asciification terimi bu işlem için kullanılmaktadır. "Normal" sütunları ise verilerin asciification işlemi yapılmamış haliyle elde edilen sonuçları göstermektedir. Sonuçlara göre tüm metindeki Türkçe karakterleri ASCII tablosundaki karşılıklarına dönüştürmek başarımı bazı modellerde yükseltmiştir, daha çok özniteliğin eklendiği ve modelin kompleksleştiği durumlarda asciification işlemi başarımı düşürmüştür, örneğin asciification yapılmış metinlerde kısmi eşleşme yapılması başarımı düşürmüştür.

Alana özgü yapılan değerlendirmelerin sonuçlarına göre temel yöntemde kullanılan biçimbilimsel aşamalara gerek kalmadan VİT sisteminde yeterli başarım elde edilebilmektedir.

5.2 Alan Dışı Değerlendirme

Alan dışı değerlendirme eğitim ve test aşamalarında kullanılan veri türlerinin farklı olması demektir. Bu çalışmada haber metinlerinde eğitilmiş model üzerinde tweet metinlerinin test edilmesini kastetmektedir.

İlk olarak Çelikkaya ve diğ. (2013) çalışmasında uygulandığı gibi tweet metinleri haber metinlerinin temel yöntem ile eğitiminden oluşturulan modelde test edilmiştir. Daha sonra haber metinlerinin bu çalışmada geliştirilen yöntem ile eğitilmesinden oluşturulan modelde tweet metinleri test edilmiştir. Sonuçlar Çizelge 5.6 ve 5.7' da verilmiştir.

Çizelge 5.6 : Temel modelde tweet metinlerinin alan dışı başarımı.

Model	Tweet-1 Geliştirme Seti	Tweet-1 Test Seti	Tweet-2
Sadece kök	16.31	11.40	14.76
Sözcüğün kendisi (küçük harf)	32.30	31.59	14.01
+Kök	26.01	24.94	18.80
+Sözcük türü	26.22	25.02	15.16
+İsim hali	25.69	25.04	15.75
+Özel isim etiketi	15.14	19.02	17.55
+Çekim ekleri	16.46	14.37	9.58
+Büyük/Küçük harf	29.99	31.70	9.79
+Cümle başlangıcı mı	30.34	32.91	20.38
+Sözlük (tam eşleşme)	37.30	37.58	20.27
+Sözlük (kısmi eşleşme)	**42.43**	**38.20**	**25.45**

Çizelge 5.5 ve 5.6' daki Tweet-1 test ve geliştirme setleri bu çalışma kapsamında toplanan ve etiketlenen tweet verilerinin test ve geliştirme setleridir, Tweet-2 seti ise Çelikkaya ve diğ. (2013) çalışmasında kullanılan tweet verilerinin tümüdür.

Çizelge 5.7 : Geliştirilen modelde metinlerinin alan dışı başarımı.

Model	Tweet-1 Geliştirme Seti	Tweet-1 Test Seti	Tweet-2
Sözcüğün kendisi (küçük harf)	32.30	31.59	18.80
+İlk 4 karakter	34.75	32.59	19.12
+Son 4 karakter	36.39	34.72	20.80
+Kesme İşareti	37.43	36.09	22.70
+Büyük/Küçük harf	41.58	40.75	24.79
+Cümle başlangıcı mı	41.57	40.78	24.15
+Sözlük (tam eşleşme)	45.83	**43.67**	27.90
+Sözlük (kısmi eşleşme)	**46.61**	41.78	**28.53**

Çizelge 5.5 ve 5.6' da görüdüğü gibi alan dışı değerlendirmede tweet metinlerinin başarım düşmektedir.

6. SONUÇ VE ÖNERİLER

Bu çalışmada Türkçe tweetler için varlık ismi tanıma sistemi geliştirilmiştir. Tweet metinleri dilbilgisi, yazım ve imla kurallarına uymayan rahat bir dille yazılmış, 140 karakterle sınırlı, içeriği az olan kısa metinler oldukları için mevcuttaki doğal dil işleme yöntemlerinde yeterli performansa ulaşamamaktadırlar. Literatürdeki çalışmalarda bu tarz metinler normalizasyon yöntemleri ile mevcuttaki doğal dil işleme araçlarına uygun olacak formata dönüştürülmüşlerdir ya da bu tarz metinler için yeni yöntemler geliştirilmiştir.

Normalizasyon tekniklerinde kuralsız formatta yazılmış metinlerin imla ve yazım kuralları açısından düzeltilmesi hedeflenir. Böylece mevcuttaki doğal dil işleme araçlarında tweet metinlerinin performans düşmesi bir nebze engellenmiş olur.

Bu çalışmada ise normalizasyon ile tweetleri düzeltmek yerine, tweetlerin doğasına uygun yeni bir sistem geliştirilmesi hedeflenmiştir. Geliştirilen sistemde biçimbilimsel aşamalar yerine kelimenin ilk ve son 4 harfi kullanılmış, büyük/küçük harf özelliklerinden yararlanılmış, tweetlerdeki Türkçe karakterler temizlenmiş ve isim sözlükleri ile kısmi eşleşme yaparak hatalı yazılan sözcüklerin elenmesinin önüne geçilmesi hedeflenmiştir.

Karşılaştırma yapabilmek amacıyla hem kurallı formattaki haber metinleri hem de kuralsız formattaki tweet metinleriyle modeller oluşturulmuş ve sınanmıştır. Modellerin oluşturulmasında daha önceki varlık ismi tanıma çalışmalarında yüksek başarım elde ettiği gösterilmiş olan koşullu rastgele alanlar yöntemi kullanılmıştır.

Tweet metinlerindeki varlık ismi tanıma performansını yükseltmek amaçlı yapılan bu çalışmada biçimbilimsel çözümleme araçları kullanılmadan tweetlerin doğasına uygun seçilen öznitelikler ile alan içi değerlendirmede yaklaşık %6 performans artışı gözlemlenmiştir.

KAYNAKLAR

Bayraktar, Ö., Temizel, T. T. (2008). Person Name Extraction From Turkish Financial News Text Using Local Grammar Based Approach. In 23rd International Symposium on Computer and Information Sciences. ICSIS'08. İstanbul.

Chinchor, N. A., Marsh, E. (1998). MUC-7 Information Extraction Task Definition. In Proceedings of the Seventh Message Understanding Conference.

Cucerzan, S., Yarowsky, D. (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In Proceedings Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Çelikkaya, G., Torunoğlu, D. ve Eryiğit, G. (2013). Named Entity Recognition on Real Data. Application of Information and Communication Technologies. Sf. 1-5.

Eryiğit, G. (2012). The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Mayıs 2012.

Eryiğit, G., Nivre, J., Oflazer, K. (2008). Dependency Parsing of Turkish**,** Computational Linguistics, 34 no.3, 2008.

Grishman, R., Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. COLING'96. Sf. 466-471.

Küçük, D., Steinberger, R. (2014). Experiments to Improve Named Entity Recognition on Turkish Tweets. Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM). EACL 2014. Sf. 71-78. Gothenburg, İsviçre.

Küçük, D., Yazıcı, A. (2009). Named Entity Recognition Experiments on Turkish Texts. In Proceedings of the 8. International Conference on Flexible Query Answering Systems. FQAS'09. Sf. 524-535. Berlin, Almanya.

Küçük, D., Yazıcı, A. (2012). A Hybrid Named Entity Recognizer for Turkish. Expert Systems With Applications. 39(3):2733-2742.

Lafferty, J. D., McCallum, A., Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML. Sf. 282-289.

Levenshtein, V. (1966). Binar Codes Capable of Correcting Deletions, Insertions, and Revelsals. Soviet Physics Doklady. Sf. 707- 710.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., Lee, B. (2012). TwiNER: Named Entity Recognition in Targeted Twitter Stream. SIGIR 2012.

Liu, X., Zhang, S., Wei, F., Zhou, M. (2011). Recognizing Named Entities in Tweets. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Sf. 359- 367.

Marsh, E., Perzanowski, D. (1998). MUC-7 Evaluation of IE Technology: Oveeview of Results.

Nadeau, D., Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes. 30(1):3-26. John Benjamins Publisher Company.

Özkaya, S., Diri, B. (2011). Named Entity Recognition by Conditional Random Fields from Turkish informal texts.

Oflazer, K. (1994). Two-Level Description of Turkish Morphology. Literary and Linguistic Computing, 9(2):137-148.

Oliveira, D., Laender, A., Veloso, A., Silva, A. (2013). FS-NER: A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. International World Wide Web Conference Committe.

Ratinov, L., Roth, D. (2009). Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Sf. 147-155.

Rau, L. F. (1991). Extracting Company Names From Text. In Proceedings Conference on Artificial Intelligence Applications of IEEE.

Ritter, A., Clark, S., Etzioni, M., Etzioni O. (2011). Named Entity Recognition in Tweets An Experimental Study. Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP'11. Sf. 1524-1534.

Sak, H., Güngör, T. ve Saraçlar, M. (2008). Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. In GoTAL 2008: Vol 5221 of LNCS. p. 417-427. Springer.

Şeker, G. A., Eryiğit, G. (2012). Initial Explorations on Using CRFs for Turkish Named Entity Recognition. In Proceedings of the 24th International Conference on Compuatational Linguistics. COLING 2012. Mumbai, Hindistan.

Tatar, S., Çiçekli, İ. (2011). Automatic Rule Learning Exploiting Morphological Features for Named Entity Recogition in Turkish. Journal of Information Sciences. 37(2):137-151.

Tjong Kim Sang, E. F., Veenstra, J. (1999). Representing Text Chunks.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In Procedings of the CoNLL 2002. Sf. 155-158.

Tjong Kim Sang, E. F. ve De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In Procedings of the CoNLL 2003. p. 142-147.

Tür, G., Tür, D. ve Oflazer, K. (2003). A Statistical Information Extraction System for Turkish. Natural Language Engineering. 9:181-210.

Yeniterzi, R. (2011). Exploiting Morphology in Turkish Named Entity Recognition System. In Procedings of the ACL 2011 Student Session. p. 105-110. Portland, USA.

Url-1 <*http://www.internetlivestats.com/total-number-of-websites*>, alındığı tarih: 29.12.2014

Url-2 <*http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html*>, alındığı tarih: 29.12.2014

Url-3 <*http://www.cnts.ua.ac.be/conll2000/chunking/output.html*>, alındığı tarih: 30.12.2014

Url-4 <*https://dev.twitter.com/*>, alındığı tarih: 30.12.2014

Url-5 <*http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar*>, alındığı tarih: 15.01.2015

ÖZGEÇMİŞ

Ad Soyad:	Beyza EKEN
Doğum Yeri ve Tarihi:	İzmir, 1989
E-Posta:	beyzaeken@itu.edu.tr
Lisans:	Sakarya Üniversitesi, Bilgisayar Mühendisliği
Yüksek Lisans :	İstanbul Teknik Üniversitesi, Bilgisayar Mühendisliği

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- Eken B., and Tantuğ C., 2015: Recognnzing Named Entities in Turkish Tweets. *International Conference on Artificial Intelligence and Applications*, January 23-24, 2015 Dubai, UAE.