İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERİN ETİKETLENMESİ

YÜKSEK LİSANS TEZİ

Seda KAZKILINÇ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Eşref ADALI

EKİM 2012

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE METİNLERİN ETİKETLENMESİ

YÜKSEK LİSANS TEZİ

Seda KAZKILINÇ
(504081555)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Eşref ADALI

EKİM 2012

İTÜ, Fen Bilimleri Enstitüsü'nün 504081555 numaralı Yüksek Lisans Öğrencisi **Seda KAZKILINÇ**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"TÜRKÇE METİNLERİN ETİKETLENMESİ"** başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Prof. Dr. Eşref ADALI** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Yrd.Doç. Dr. Banu Diri** Yıldız Teknik Üniversitesi	.....
	Yrd.Doç. Dr. Ahmet Cüneyt TANTUĞ İstanbul Teknik Üniversitesi	.....
		.....

Teslim Tarihi : **05 Ekim 2012**
Savunma Tarihi : **11 Ekim 2012**

Babanneme ve hocama,

ÖNSÖZ

Gittikçe sayısı artan elektronik metinlerde, istenilen veriye daha kolay ulaşmak için bilgi çıkarımı yöntemlerinden faydalanılmaktadır. Metni en iyi şekilde temsil eden söz öbeklerini seçmek, metnin içeriğini bir kaç kelime ile özetlemek açısından çok önemlidir ve konu çıkarımı, anlamsal ağ gibi çeşitli alanlarda kullanılabilir.

Bu çalışmadaki asıl amaç metinden çıkarılan ve metin hakkında yüksek derecede bilgi taşıyan bu söz öbeklerini özne, yüklem, yer ve zaman unsurlarıyla etiketlemekdir. Bu konuyu bana öneren ve çalışmam boyunca beni yönlendiren danışman hocam Sayın Prof. Dr. Eşref ADALI'ya, beni destekleyen aileme, tezimi bitirmem konusunda beni motive eden ve bana yardımcı olan arkadaşlarıma, teşekkürü bir borç bilirim.

EKİM 2012 Seda KAZKILINÇ

İÇİNDEKİLER

Sayfa

KISALTMALAR

DDİ : Doğal Dil İşleme
VİT : Varlık İsmi Tanımlama
KEA : Anahtar Sözcük Çıkarma Algoritması
TF : Terim Sıklığı
IDF : Ters Döküman İndeksi
DVM : Destek Vektör Makinesi
SSM : Saklı Markov Modeli
MEMM : Maksimum Entropi Markov Modeli
CRF : Koşullu Rastgele Alanlar

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

TÜRKÇE METİNLERİN ETİKETLENMESİ

ÖZET

Her geçen gün belge sayısı artan Web'in tam potansiyeliyle kullanılması için anlamsal ağ alanındaki çalışmaların Web'in geleceğini oluşturacağı düşünülmektedir. Belge sayısındaki bu artışa bağlı olarak istenilen metne erişebilmek için bu metni en iyi temsil eden söz öbeklerinin bulunması doğru bir yaklaşım olacaktır. Tüm metni okumadan o metni en iyi ifade edecek söz öbeklerine erişmek hem kullanıcı açısından hem de tarayıcı açısından büyük önem taşımaktadır.

Bu çalışmanın amacı haber metinlerinde, haber metninin öznesi, yüklemi, yer ve zamanını belirtecek söz öbeklerinin metinde bulunup, metnin etiketlenmesidir. Bu amaçla, metinde geçen cümleler içerisinden seçilen en baskın özne, yüklem, yer ve zaman bilgilerinin çıkarılması hedeflenmektedir. Elde edilen bu etiket bilgileri sayesinde metnin konusu temsil edilmektedir. Bu sayede anlamsal ağda etiket olarak kullanılabilir ve arama motorlarında istenilen veriye ulaşabilmek için kullanılabilir.

Hedefimiz doğrultusunda ilk olarak, metindeki cümleler biçimbilimsel çözümleyicide analiz edilmektedir. Bunun nedeni eklemeli bir dil olan Türkçe'de sözcüklerin gövdelerine erişmektir. Biçimbilimsel çözümleyicinin sonucunda, her sözcük için birden fazla çözüm üretilmektedir.. Bu nedenle bulunan çözümlerden en yüksek olasılıklı olanı bulmak için belirsizlik gidericiye ihtiyaç vardır. Sözdizimsel çözümelere erişmek için de sözdizimsel çözümleme işlemi yapmak gerekmektedir.

Çalışmamızda bir metin ilk olarak yukarıda sıralanan üç aşamalı çözümleme işleminden geçirilmiştir. Tez çalışmasının ilk kısmında biçimbilimsel ve sözdizimsel çözümü çıkarılmış olan metinlerden kurallar çıkarılarak etiketleme işlemi yapılmaya çalışdıysak da yeterli başarımı elde edemedik. Bu nedenle, çıkaramadığımız bazı kuralları çıkarabileceğini düşünerek makine öğrenmesi yöntemleri üzerinde çalışılmıştır. Makine öğrenmesi yöntemi olarak bir dizilim sınıflandırıcısı olan Koşullu Rastgele Alanlar (CRF) üzerinde çalışılmıştır. Kural tabanlı yaklaşımda elde ettiğimiz bazı kuralları kullanarak ve çözümleyi çıktılarını kullanarak metindeki her bir sözcüğe ait nitelikler belirlenmiştir. Önceden elle işaretlediğimiz metinleri ve belirlenen nitelikleri kullanarak, CRF modelimizi oluşturulmuştur. Daha sonra önceden etiketlenmemiş metinleri, bu model sayesinde etiketleme işlemini geliştirilmiştir.

Bu çalışmanın bilimsel ve teknik katkısını ortaya çıkarabilmek için, sınama kümesindeki elle etiketlediğimiz metinlerin etiketlerini CRF'in ürettiği etiketler ile karşılaştırıp başarımımızı tutturma ve bulma olasılıkları ve bunlardan türeyen F-ölçüm oranı cinsinden ölçülmüştür.

LABELING TURKISH DOCUMENTS

SUMMARY

Most current significant word extraction from a document uses keyphrase extraction features. In this thesis, a new approach that is labeling the main subject, main predicate, main location and main date of a electronic document is introduced. The main subject label tells whom or what the document about. The main predicate label tells what the subject is or does. The main location label tells where the document passed and the main date label tells when the document passed. With the help of this new methodology, extraction of not only high level description of the content, but also the attribute of a phrase in a document are provided. As an experiment set Turkish news are selected. To use as a training and test set, manual labeling is made by human annotators. Then, different models for each label are implemented to extract the labels automatically and they are compared to manually labeled results.

As internet grows dramatically, the number of electronic text documents increases considerably. By means of increasing number of documents, the information extraction grows in importance. On this account, there are several researches to reach the information needed. This thesis introduces a new approach to information extraction, which provides extraction of the main subject, main predicate, main location and main date of a text document and label it to use for semantic web applications. This approach is a new field of study, which aims to short summary of a text with the help of labeled entities. The most pronounced difference between keyphrase extraction studies and labeling study presented in this thesis is that this study extract the most significant phrases with their functions in the document.

The news in Turkish language is selected as an experiment set in this study. Labeling the main subject, main predicate,main location and main date of news which are gathered from web, is totally new field of study which is introduced in this thesis. As a literature survey it is gathered that the best similar studies are focused on the keyphrase extraction.

To use as a training and test set, 200 raw news are gathered from RSS feeds of Turkish news distributors from internet. All these news are converted to the XML file.Then all labels are manually annotated. If they cannot find the label in the document,they should enter the label tag with dash punctuation. This means there is not any proper label in the document.

150 news are used as a training set to obtain best model of extraction each label. 50 news are used as test set to compare manually annotated results with automatically extracted results.

In order to decide whether to label a phrases as a keyphrase, the words in the document must be distinguished by using specified features and also the properties of keyphrases

have to be identified. The first possible feature that comes into mind is the frequency, which is the number of times a keyphrase appears in the text. It is obvious that the more important phrases will be more used in a text. Second one is the first place the phrase occurs in a document has more priority for labeling. To extract the phrases in a document, several models can be used as named entity recognition or collocation extraction models.

Subject label indicates what or whom about the document. Due to the experiment set of this study is news, main subject and main location of the text should be proper noun phrases. This assumption is obtained after inspected all manually annotated subject labels. In order to obtain proper name phrases in Turkish language, firstly all words start with capital letter are gathered. However, this assumption is not correct at all because some other words may start with capital letter, such as first word of sentence, titles, month or day names in dates etc. First of all, all words starts with capital letter and conjunctions between them are gathered. Because of some of proper name phareses sequentially present at the document. Some of Turkish language rules defined.

For example, If the word is first word in a sentence and it is a proper name, it is a possible candidate of proper noun phrase. If a word starts with capital letter and not the first word of sentence, select it as a possible candidate of proper name phrase. If a conjunction is between two possible candidates of proper name phrases, select this word.

But all these rules are not enough to divide all these words into proper noun phrases. For instance, “Mustafa Kemal Atatürk" Ankara’ya gitti.”is a sample Turkish sentence. In this sentence Mustafa Kemal Atatürk ”and “Ankara”are two different proper noun phrases. However, the rules explained above selects the proper name phrase as “Mustafa Kemal Atatürk Ankara’ya”. So new boundary rules are defined.For instance, if a possible candidate of proper noun etc, this word is the last name of proper noun phrase. If a possible candidate has the suffix as ”P3sg”, this word is a last word of proper noun phrases.

However all these rules are not adequate to select main subject of text. So, Conditional Random Fields are used as a machine learning classifier.

Due to the Turkish is an agglutinative language, input file is converted to the file includes the information of stems, inflectional suffixes and parser results of the raw new. Input file is converted to the file includes the information of stems, inflectional suffixes and parser results of the raw new. The reason why we need stems and inflectional suffixes is Turkish is an agglutinative language. Turkish language has few prefixes and many suffixes. Labelling of a Turkish text is not so easy as an English text.

The processing steps of our algorithms are given as follows:

- Morphogical analysis
- Morphogical disambugation
- Dependancy parser

After preprocessing document, to devolop CRF system, features are selected as following categories:

- Rule based features
- Morphological features
- Syntactic Features
- Structural Features

When the feature selection is completed, realted fatures are assigned to the training set. As using feature assigned training set, CRF model is trained. Then, this model is ready to label any unlabeled news.

During evaluation, test set is used to compare the annotator's tags with CRF tags. For each label, this comparison is made. If a phrase in annotator's tag is exactly the same as the phrase in program tag or one tag contains the other tag , it is assumed as labeling is correct for this phrase.

In this study, the main concern is the precision and the recall that is how many of the suggested keywords are correct (precision), and how many of the manually assigned labels that are found (recall). We measure the performance of the algorithm in relation to the labels assigned by the annotators.

The problem caused by errors in automatic morphological analysis, disambiguation and dependency parser should be taking into account during evaluation of the results. Another important effect is that the use of spell checker can increase parsing accuracy substantially. By combining the linguistic rules approach with statistical approaches,we have been able to achieve the highest accuracy of labeling documents.

1. GİRİŞ

İnternet ortamındaki belge sayısındaki hızlı artışa bağlı olarak kullanıcıların aradığı belgelere daha hızlı erişmesini sağlayacak tekniklere duyulan ihtiyaç gün geçtikçe artmaktadır. Her ne kadar İnternet, aranan kaynağa ulaşmakta önemli olanaklar yaratmaktaysa da artan belge sayısı beraberinde bilgi kirliliğine ve belirsizliğine neden olmaktadır. Web'de bulunan kaynak sayısındaki hızlı artış, asıl belgeye ulaşmak işlemini zorlaştırmaktadır. Günüzmüzde yaygın olarak kullanılan arama motorları çok basit yöntemlerle arama yapmaktadırlar. Bunun sonucu olarak, kullanıcının karşısına ilgili ve ilgilsiz çok sayıda belge getirilmektedir. Bu belgeler içinde gerçekten aranan belgeye ulaşmak kullanıcının yeteneğine ve sabrına bırakılmaktadır.

Bilim ve teknoloji alanındaki tüm gelişmelere karşın günümüzde kullanıcı aradığı belgeye ulaşmak isterken, bu tarayıcıların çıkardığı sonuçları kendi yorumuyla elemek durumundadır. Bu nedenle tarayıcıların belge aramada daha etkili kullanılabilmesi için çeşitli çalışmalar yapılmaktadır. Bu hedefe yönelik ufak bir katkı bile büyük miktarda zaman kazanımına neden olacaktır.

Doğal dil işleme alanındaki çeşitli çalışmalar bu gereksinimi gidermek için yeni olanaklar sunmaktadır. Kaynak taramamızın sonucu, incelenen kaynakların büyük bir çoğunluğunun bu işlemi anahtar sözcük yakalama işlemini kullanarak çözme üzerinde yoğunlaştığını göstermektedir. Bu çalışmalarda temel amaç bir belgenin diğer belgelerden farkını ortaya koyarak, belgeye ayırt edici özellik kazandıran ve belgeyi en iyi şekilde tanımlayan anahtar sözcük öbeklerini otomatik olarak çıkarma olarak açıklanabilir.

Bizim çalışmamızda ise amaç, anahtar sözcük öbeği bulma çalışmalarından farklı olarak, metni en iyi tanımlayan özne, yüklem, yer ve zaman etiketlerini çıkartmaktır. Bu açıklamamızdaki özne metnin tümünün öznesi, yüklem metnin tümünü açıklayan yüklem olarak düşünülmektedir. Benzer yorum yer ve zaman etiketleri içinde geçerlidir. Çalışmamızı anahtar sözcük bulma çalışmalarından ayıran en belirgin

fark, sadece anahtar sözcük bulmak yerine, bulunan sözcüklerin niteliklerini de bularak anlamlaştırmak olarak açıklanabilir. Çalışmamız Web üzerinde yayımlanan haber nitelikli belgeler üzerinde yoğunlaşmıştır. Bu nedenle çalışmamızın bu gözle değerlendirilmesi gerekmektedir. Özne, yüklem, yer ve zaman etiketlerinin anlamları Çizelge 1.1'de gösterilmiştir.

Çizelge 1.1: Haber metni etiketleri ve anlamları

Özne	Haberdeki ana karakter, yüklemin bildirdiği durumu üzerine alan kimse veya şey, fail, süje veya kümeleri
Yüklem	Oluş, iş ve hareket bildiren sözcük veya sözcük kümesi
Yer	Haberin geçtiği veya belirttiği yer
Zaman	Haberin geçtiği veya belirttiği zaman

Web'de bulunan veya web'e yeni konulacak belgelerin anlamsal web uygulamalarında kullanılmak üzere özne, yüklem, yer ve zaman etiketleriyle etiketlenmesi, erişilmek istenen belgeye ulaşmada zamanı kısaltacak ve gerçekten aranan belgeyi bulmada önemli bir katkı sağlayacaktır.

Çalışmamızın çerçevesi şöyledir:

- Ele alının belgelerin dili Türkçe'dir.
- Belgeler, haber niteliklidir.
- Belgelerin boyutları ortalama 50-300 sözcüktür.

Çalışmamızın başarım ölçümünü yapabilmek için 200 kadar haber metni tarafımızdan değerlendirilmiş ve elle etiketlenmiştir.

Haber metinlerinin başlıkları çoğu zaman insanlar için bile yanıltıcı olmaktadır. Bu nedenle çalışmamızı metnin asıl gövdesi üzerinde yapmamız gerekmiştir.

Bir metnin etiketlenmesi sırasında izlediğimiz yöntem şöyledir:

- Ele aldığımız metinlerin Türkçe olması nedeniyle dilimize özel yöntemler geliştirilmiştir.

- Metinde geçen sözcüklere ait nitelikleri çıkarmak için önce sözcüklerin biçimbilimsel çözümlemlenmesi yapılmış, ardından tüm sözcüklerin biçimbilimsel belirsizliği giderilmiş, daha sonra tümcelerin sözdizimsel çözümlemeleri yapılmıştır. Bu çalışmamız sayesinde metinde geçen tüm sözcüklerin, dilbilimsel nitelikleri elde edilmiş olmaktadır.

- İlk aşamada kural tabanlı yaklaşımlar ile etiketleme işlemi gerçekleştirilmeye çalışılmış, ancak başarı oranın düşük olması nedeniyle yeni yöntem arayışına gidilmiştir. Bu kapsamda makine öğrenmesine dayalı yaklaşımlar ile problem çözülmeye çalışılmıştır. Kural tabanlı çalışmamızda etiketleme için etkin olan nitelikler makine öğrenmesine dayalı yöntemde etkin niteilkler olarak kullanılmıştır.

- Yukarıda anlatılan adımların gerçekleştirilmesinin ardından sözcük, cümle ve belgeye ilişkin nitelikler çıkarılmış ve bir makine öğrenmesi yöntemi olan CRF ile modellenmek üzere önceden elle etiketlenmiş belgeler yardımıyla sistem eğitilmiştir.

- Etiketlenmemiş belgeler tarafımızdan geliştirilen yöntem uyarınca çalışan CRF temelli yazılım kullanarak etiketlenmeye çalışılmıştır.

- Geliştirdiğimiz yöntemin başarımını ölçmek için ise, daha önce insanlar tarafından elle etiketlenmiş belgeler tarafımızdan geliştirilmiş olan etiketleme yazılımına girdi olarak verilmiş; yazılımın ürettiği etiketler gerçek doğru etiketlerle karşılaştırılarak çalışmamızın başarımı ölçülmüştür. Yöntemimizin başarımı şu an için

Metinlerin etiketlenmesinde metinlerde bulunan asıl başlık ve alt başlıkları kullanılmamıştır. Bunun nedeni metin başlıklarıyla metnin ana etiketinde büyük farklılıklar oluştururmasındandır.

Çalışmamızın ilk aşamasında sıklık analizi ve sözcüğün ilk bulunduğu yeri önemseyen yöntemler denenmiştir. Türkçe sondan eklemeli bir dil olduğu için sözcüğün sıklığını bulmada sözcüklerin gövdelerini bulmak gerekir. Metin etiketleme işlemi Hint-Avrupa dillerindeki belgeler için yapılmaya çalışıldığında, çözüm Türkçe metinlere göre daha kolay olmaktadır. Bunun nedeni Hint-Avrupa Dilleri'nde peş peşe eklenen

eklerin hem sayısı az, hem de bir sözcüğe ard arda eklenen eklerin sayısı bir, ikiyi aşmaz. Dolayısıyla sözcükerin köküne ulaşma zorunluluğu yokyur; sözcük sıklığına bakılarak önemli sonuçlar alınabilir. Türkçe sözcüklerin gövdelerine ulaşmak için ilk aşamada biçimbilimsel çözümleyiciye gerek vardır. Ancak bilinen bir başka gerçek ise Türkçe sözcüklerin ortalama iki biçimbilimsel çözümü bulunmaktadır. Dolayısıyla biçimbilimsel çözümleyiciden elde edilen sonuçların belirsizliklerini gidermeye ihtiyaç duyulmaktadır. Biçimbilimsel çözümleyide Oflazer'in biçimbilimsel çözümleyicisi [1] çalışmasından ve belirsizlik giderici olarakSak ve arkadaşları'nın [2] belirsizlik giderici çalışmasından faydalanılmıştır.

Sadece sözcüklerin metin içerinde sıklığından ve sözcüğün metin içerisinden bulunduğu yeri hesaba katan yöntemlerden faydalanmak yeterli ve anlamlı bir sonuç vermemektedir. Türkçe'nin sözdizimsel kurallarını da çözüm sırasında göz önüne almak gerekmektedir. Bu konuda Eryiğit'in [3] çalışmasından faydalanılarak metinde geçen cümlelerin sözdizimsel analizi ve sözcüklerin buna göre etiketlenmesi gerçekleştirilmiştir.

Bir tümce içinde özne, yüklem ve tümleçlerin tek sözcükten oluşmadığı da bilinen bir gerçektir. Bu nedenle sözcük kümelerinde oluşan öznelerin, yüklemlerin ve tümleçlerin değerlendirilmesi gerekmektedir. Varlık İsmi Tanımlama (VİT; Named Entity Recognition; NER) ve birliktelikler konusunda Türkçe üzerine yapılan çalışmaların belli bir başarıya ulaşamaış olması nedeniyle tarafımızdan bu amaçla kurallar oluşturulmuştur. Bu yöntemde belirli varsayımlar yaparak sözcük öbekleri elde edilmiştir. Bu kurallarla ilgili br örnek aşağıda verilmiştir.

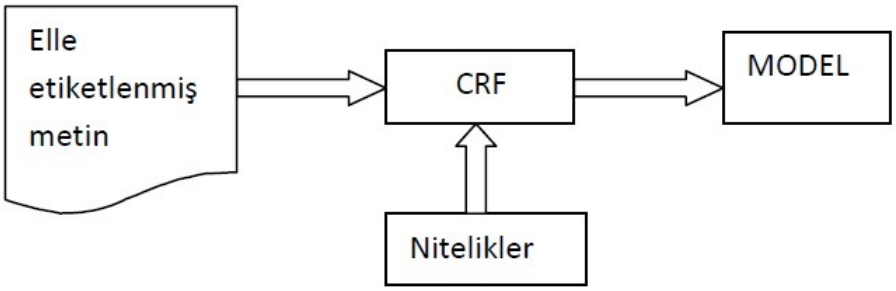
Elle etiketlenmiş haber metinlerinin tamamına yakınının öznesi özel isimdir ve büyük harfle başlar. Özne ve yer etiketlerini çıkarmada kullanılmak üzere metinde geçen özel isim öbekleri kural tabanlı yaklaşımlarla çıkarılmaya çalışılmış ve olası öbekler makine öğrenmesi uygulamalarında kullanılmak üzere nitelik olarak kaydedilmiştir.

Tez çalışmasının ilk bölümünde kural tabanlı yaklaşımlar ile etiketleme sorunu çözülmeye çalışılırken, tüm kuralları ve ilişkileri çıkarmanın zor olduğu görülmüştür. Bu nedenle soruna sınıflandırma sorunu olarak yaklaşmak ve makine öğrenmesi yöntemleriyle sorunu çözmek için çeşitli araştırmalar ve çalışmalar yapılmıştır. Konu

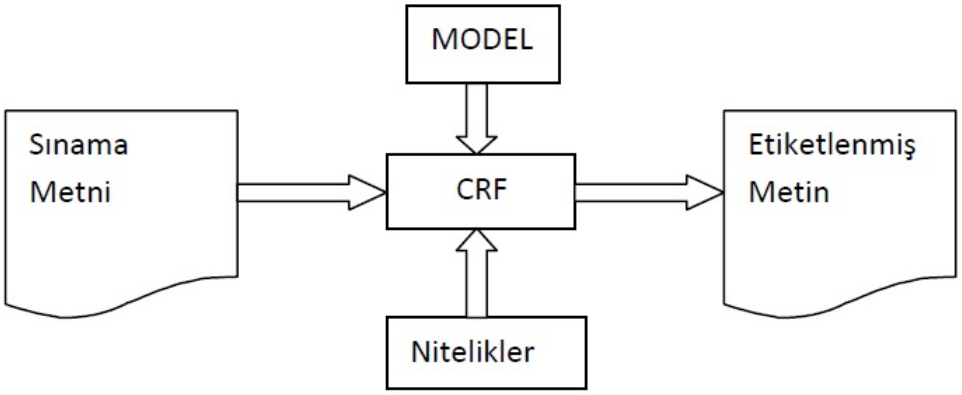
bir dizilim sınıflandırma sorunu olduğundan, benzer konularda verimli sonuçlar verdiği bilinen Koşullu Rastgele Alanlar (KRA; Conditional Random Fields; CRF) yöntemi denenmiş ve amacımıza yönelik olarak geliştirilmiştir.

Biçimbilimsel analiz ve belirsizlik giderici çalışmaları sonunda biçimbilimsel nitelikler elde edilmiş; sözdizimsel çözümlemenin sonunda sözdizimsel nitelikleri elde edilen sözcükler, belgedeki çeşitli yapısal nitelikleri ve kural tabanlı yaklaşımlarla elde edilen nitelikler yardımıyla makine öğrenmesinde kullanılmak üzere kaydedilmiştir.

Elle etiketlenmiş sözcükler nitelikleri ile birlikte kaydedilerek, CRF modellenmesinde eğitim kümesi olarak kullanılmıştır. Böylece eğittiğimiz CRF yapısına hiç etiketlenmemiş belgeler girdi olarak verilmiş ve etiketlenmeleri sağlanmıştır. Şekil 1.1 ve Şekil 1.2'de tarafımızca tasarlanan eğitim ve sınama aşamaları gösterilmiştir.



Şekil 1.1: Sistemin Eğitimi



Şekil 1.2: Sistemin Sınanması

Son olarak öğrenme kümesi kullanılarak eğittiğimiz CRF yapısına elle işaretlenmiş belgeler, giriş olarak uygulanmış geliştirdiğimiz yöntemin bulma ve tuturma olasılıklarıyla birlikte başarımı ölçülmüştür.

1.1 Tezin Katkısı

Tezin katkılarını aşağıdaki gibi sıralayabiliriz:

- Geliştirdiğimiz yöntem metnin özne, yüklem ve tümlecini bulmaya yönelik olması nedeniyle anahtar sözcük çıkarımı veya özetleme çalışmalarından farklıdır.
- Etiketlerin niteliklerinin bulunmasında kural tabanlı yaklaşımlardan faydalanılmış ve etkin nitelikler çıkarılmıştır.
- Türkçe'nin eklemeli bir dil yapısı olduğu göz önüne alınarak sözcük tiplelerinin ve ek tiplerinin etkisini bulabilmek için biçimbilimsel çözümleme ve belirsizlik giderme işlemleri yapılmıştır. Buna bağlı olarak ekleri nitelik olarak kullanarak Türkçe'de eklerin kelimeye kazandırdığı bağlılık analizi yapılmıştır.
- İncelediğimiz kaynak taraması sonucu benzer bir çalışmaya rastlanmamıştır.

1.2 Tezin Amacı

Bu çalışmanın asıl hedeflerden biri anlamsal web için kullanılmak üzere, metinde en çok bahsi geçen özneyi, metni en iyi açıklayan eylemi, metnin yeri ve metinde bahsi geçen zaman bilgisine ulaşmaktır. Bu sayede gittikçe sayısı artan web kaynaklarında kullanıcının erişmek istediği asıl kaynağa daha kolay ve hızlı erişmesi hedeflenmektedir.

Başka bir yararı ise arama motorları belgenin etiketlerine bakarak arama yapacağı için, işi kolaykaşır ve hızı artar. Bu çalışma sayesinde arama motorlarına ek bir özellik vererek haber metinlerine ulaşmaya çalışan kullanıcılara, girdikleri anahtar sözcükler ile etiketler karşılaştırılarak daha nitelikli sonuçlar sunulabilir.

Bu çalışmadaki hedef her ne kadar anlamsal web için destekleyici bir çalışma yapmak olsa da, bir haber metninin öznesi, yüklemi, yeri ve zamanını görmek okuyucuya bir bakışta konunun ile ilgi alanı olup olmadığı hakkında fikir verir.

Çalışmanın başka bir kullanım alanı ise metnin konusunu özetleyecek bir cümlenin çıkarılması olabilir. Elimizde metni en iyi tasvir eden öğelerin bulunduğunu hesaba katarsak, bu öğelerden metni özetleyecek bir cümle oluşturmak çok da zor olmayacaktır ve bu konuda Türkçe dili için geliştirilebilecek cümle oluşturma yöntemleri ile birlikte kullanılabilir.

Aynı zamanda çalışma, belge sınıflandırmada da kullanılabilinir. Ayrıca çalışmanın çıktısı olan metinin öznesi, yüklemi, yeri ve zamana ilişkin bilgiler, belge sınıflandırması işleminde daha ayrıntılılı sonuç verecek bir araç olarak kullanılabilir.

1.3 Tezin Yapısı

Bu tezde ilk olarak literatür araştırılması yapılmış ve teze referans olabilecek yakın çalışmalar incelenmiştir.

İkinci bölümde ise kuramsal altyapı çalışmaları detaylı olarak anlatılmıştır. Tezin anlaşılabilirliğini ve bütünlüğünü sağlamak amacıyla bazı altyapı bilgileri bu kısımda verilmiştir. Tezin geliştirilmesi kısmında bu kısımda anlatılan kurallar ve yöntemler kullanılmıştır.

Üçüncü bölüm kurallar ve yöntemlerin belirlenmesi kısmıdır. Tez çalışmasında bu kural ve yöntemler kullanılarak geliştirme yapılmıştır.

Dördüncü bölümde bu tezde tasarlanan ortaya konan yöntemin nasıl geliştirildiğii ayrıntılı olarak anlatılmıştır. Geliştirme kısmında hangi tür teknolojilerin kullanıldığı, ve oluşturulan algoritmalar anlatılmıştır.

Beşinci bölüm ise ölçme ve değerlendirme kısmıdır. Bu bölümde sistemin başarımını ölçmek için kullanılan yöntemler ve sonuçlar tartışılmıştır.

Son bölümde çalışma genel olarak çalışma değerlendirilmiş ve bu konuda neler yapılabileceği özet olarak açıklanmıştır.

1.4 Benzer Çalışmalar

Çalışmamıza yön gösteren bildiri ve makaleler incelenmesi sonunda ve bunlar dört kısımda kümelenmiştir.

1. Varlık İsmi Tanımlama
2. Birliktelikler
3. Anahtar sözcük öbeği çıkarma
4. Diğer çalışmalar

Bu kısımlarda yer alan akademik çalışmalar ile ilgili değerlendirmeler aşağıda anlatılmıştır.

1.4.1 Varlık ismi tanımlama

Varlık İsmi Tanımlama (VİT) bilgi çıkarımının bir alt dalı olup, metinlerde daha önceden çıkarılmış veya elde var olan bilgileri kullanarak kişi, kurum, kuruluş, yer isimleri, zaman ifadeleri, para birimleri gibi varlıkları tanımlama işlemidir [4]. Örnek olarak Şekil 1.3'deki gibi bir çıkarım yapılabilir.

<KİŞİ> Seda Kazkılınç </KİŞİ> <ZAMAN> 17 Ekim 1985'de </ZAMAN> <YER> Kırşehir 'de </YER> dünyaya geldi.

<KURUM> İstanbul Teknik Üniversitesi </KURUM>, <YER> İstanbul'da</YER> yer alan <ZAMAN>1773 yılında </ZAMAN> kurulmuş devlet üniversitesidir.

Şekil 1.3: Varlık ismi tanımaya bir örnek

VİT çalışmaları kural tabanlı, gözetimli makine öğrenmesi ve melez yaklaşımlar olarak üç ana başlıkta incelenebilir.

- Kural tabanlı çalışmalar
- Makine öğrenmesi temelli yaklaşımlar
- Melez yaklaşımlar

Genel olarak incelendiğinde ilk kümedeki çalışmalar daha çok kural tabanlı iken, daha güncel olan çalışmalar istatistiksel yöntemlere ağırlık vermektedir. İstatistiksel yöntemlerin ve makine öğrenmesine dayalı yöntemlerin başarım oranları eğitim kümesi ile doğru orantılıdır. Ancak çoğu zaman büyük boyuttaki eğitim verisini hazırlamak zahmetli bir işlemdir. Bunun önüne geçmek için yarı güdümlü

makine öğrenmesi yöntemlerinden olan önyükleme algoritmalarına başvurumaktadır. Güdümsüz yöntemler genellikle demetleme algoritmalarını kullanır. Eğitilmemiş bir derlem kullanılarak istatistiksel yöntemler sayesinde kümeleme işlemi yapılır.

1.4.1.1 Kural tabanlı çalışmalar

Kural tabanlı yaklaşımlar genellikle doğal dil işleme (DDİ; Natural Language Processing; NLP) yöntemlerini kullanırlar.

Kural tabanlı yaklaşımlara bir örnek olarak İngilizce dili için yapılmış olan Crystal [5] çalışması verilebilir. Bu çalışma dilden örüntüler çıkarılarak oluşturulmuş bir sözlükte benzer sözcüklerin çıkarılması için kullanılabilir. Bunun için kavramlar sözlüğünün otomatik olarak oluşturulmasını sağlamaya çalışır. Makine öğrenmesi yöntemleriyle eğitim kümesinin sistemi eğitmesiyle oluşturulur.

Diğer bir örnek olarak Nymble [6] ise varlık isimlerini metinlerden çıkarmak için Saklı Markov Modeli'ni kullarak eğitilmiş bir modeldir. Eğitim kümesinin istatistiksel yöntemlerde başarı oranını doğrudan etkilemesinden dolayı başarısı yüksek bir yöntemdir. İngilizce ve İspayolca için uygulanmıştır.

Diğer bir önemli çalışma ise NetOwl'dur [7]. İleri dil işleme yöntemlerini kullanarak anahtar kavramları çıkarıp sınıflandırmayı hedefler.

Küçük tarafından yapılan çalışma da [8] kural tabanlı bir yaklaşımdır. Kişi isimleri, tanınmış kişiler, tanınmış organizasyon isimleri gibi sözlükleri bulmaktadır. Ayrıca Türkçe için belirli örüntüler çıkarılır. Bunlara bağlı olarak haber metinlerinde varlık isimlerini çıkarmaktadır ve

Bayraktar ve arkadaşları tarafından yapılan "Finansal Haber Metinlerinde Kişi İsmi Etiketleme" isimli çalışma [9] ise yerel dilbilgisi yaklaşımı üzerine yoğunlaşmıştır. Yerel dilbilgisi yaklaşımı varlık tanıma esnasında diğer varlık tanıma sistemlerinin aksine hiç bir genel sözlük, isim, organizasyon ya da yer sözlüğüne ihtiyaç duymamaktadır. Sonuç olarak yerel dilbilgisi yaklaşımı daha önce görülmemiş metinlerde varlıkları tanımakta ve sınıflandırmaktadır. Diğer varlık tanıma sistemleri yerel dilbilgisi yaklaşımının aksine örüntü oluşturmadan önce bazı anlamsal ve yapısal analizlere ihtiyaç duymaktadır. Kişi isimlerini çıkarmada kullanılan bu yöntem ile

yerel dilbilgisi yaklaşımının sıklık analizi, uygunluk analizi ve eşdizimlilik analizi kullanarak Türkçe'ye uygulanabilirliğini araştırılmıştır.

1.4.1.2 Makine öğrenmesine dayalı çalışmalar

Güdümlü makine öğrenmesi temelli yaklaşımlar DDİ yöntemlerini kullanmadan kendi modellerini çıkarmayı hedeflerler.

Bu alandaki öncü çalışmalardan biri olan Cucerzan ve arkadaşlarının çalışması [4] kişi, yer, kuruluş ve diğer önemli isimleri metinden çıkarmayı hedefler. Dilden bağımsız geliştirilen bu çalışma tekrarlı öğrenmeye dayanan ve biçimbilimsel örüntüleri kullanarak ve bağlama bağlı olarak hiyerarşik bir model oluşturur. Sadece dilden bağımsız olarak elle etiketlenmiş bir veri kümesini model oluşturmak için kullanır. Bu veriler sayesinde o dile bağlı örüntüler çıkarır. Bu yöntem önyükleme algoritması izlenerek oluşturulmuş bir yöntemdir. Bir çok dil için uygulanan bu yöntem Türkçe için de uygulanmıştır [4].

1.4.1.3 Melez çalışmalar

Melez yöntemler DDİ çalışmalarının ve istatistiksel yaklaşımların bir arada kullanılması ile yapılan çalışmalardır.

Oflazer ve arkadaşları tarafından yapılan "Türkçe için İstatistiksel Bilgi Çıkarım Sistemleri" isimli çalışamada [10], Saklı Markov Modeli içinde gömülü n-gram dil modelini kullanılmıştır. Sözlük modeli ve biçimbilimsel modelin birlikte uygulanması sonucu ortaya çıkan bu yeni model ile % 91.56 oranında başarı elde edilmiştir.

1.4.1.4 Diğer çalışmalar

Yapılan kaynak taramasında bu çalışamada hedeflenen etiketlemeye benzer sadece bir çalışmaya rastlanmıştır. Nallapati ve arkadaşlarının yaptığı [11] haber metinlerinden anahtar sözcük çıkarımı çalışması anahtar kişiler, anahtar yerler, anahtar isimler ve anahtar eylemeri haber metinlerinen çıkarmayı hedefler. Buna bağlı olarak bu sorunu sınıflandırma problemi olarak görür. Öncelikli olarak anahtar sözcükleri çıkarır ve anahtar sözcükleri Naive Bayes, Saklı Markov modeli ve Maksimum Entropi Model'i ile anahtar sözcükleri sınıflandırır. Arama motorlarınca dikkate alınmayan ve çok

tekrarlanan ve sıralama hesaplarına dahil edilmeyen sözcüklerin ayıklanmasıyla elde edilen anahtar sözcüklerin Maksimum Entropi Model'i ile sınıflandırılması sonucu en iyi sonuçlar elde edilmiştir. Bizim çalışmamızın İngilizce dili için yapılmış bu çalışmadan farkı, bu işlemi Türkçe gibi eklemeli bir dil ile yapmasının yanında çıkartılan etiketlerin cümlelerin öğeleri gibi metnin öğelerini çıkaran bir yaklaşım izlemesi ve bu amaca yönelik bilgi çıkarma yöntemine gitmesidir.

Her ne kadar başarılı sistemler geliştirilmiş olsa da VİT sistemleri hâlâ bir çok ismi düzgün biçimde çıkaramamaktadır. Ard arda gelen varlık isimlerini çıkarmada hâlâ bir çok sorun bulunmaktadır. Örneğin yer isminden sonra gelen kişi isimleri buna bir örnektir. Diğer bir zorluk ise bir varlık isminin diğer bir varlık ismini içinde barındırmasıdır. Örneğin içinde kişi ismi barıdıran bir organizasyon isminin bulunmasından dolayı problem yaşarlar.

Bu nedenlerden dolayı VİT sistemleri çalışmamızda kullanılmamıştır. Çünkü haber metinleri bol miktarda özel isim öbekleri içeren metinlerdir ve özne, yer ismi çıkarmada özel isimlerin çıkarımı başarıyı büyük oranda etkilemektedir. Aynı zamanda öznenin organizasyon ismi mi kişi ismi mi olduğunu bilmeye bu çalışmada gerek yoktur. Önemli olan öznenin düzgün etiketlenmesidir ve bu amaca yönelik yöntemler geliştirilmiştir.

1.4.2 Birliktelikler

Bu kümede incelenen çalışmalarda Derlem Dilbilimi'nde, birliktelik şans eseri olması umulandan çok daha fazla sıklıkta bir arada görülen sözcük veya terim dizilimini tanımlamaktadır [12].

Birlikteliklerde sözcükler anlamsal olarak birbirlerine bağlıdırlar bu nedenle bağımsız iki ayrı sözcük gibi değerlendirilmeleri anlamsal açıdan yanlış olacaktır. Örnek olarak "Yanlış kişi olduğunu fark ettim." Cümlesindeki yüklem "fark etmek"'dir.

Birliktelik analizinde, bir veya birden fazla sözcüğün bir derlem içinde ne sıklıkta beraber bulunduğu önemli bir değerdir, ancak tek başına yeterli değildir.

Diğer bir önemli nokta ise, her ne kadar birliktelikler n-gramlardan oluşmuşsa da, en çok incelenenler 2-gram ve 3-gramlardır. Oflazer ve arkadaşları çalışmasında [13] biçimbilimsel örüntüler çıkararak birliktelikler bulma yoluna gitmişlerdir.

Birliktelik analizinde en sık kullanılan yöntemlerden biri ki-kare yöntemidir. Karaoğlan ve Metin'in yaptığı çalışmada da bu yöntem ile ortak bilgi yöntemini Türkçe için kullanılmıştır. [14]

Ki-kare yöntemi t-testinden esinlenen bir yöntemdir; t-testinde yer alan normal dağılım varsayımının aksine olasılıkların rasgele dağıldığı yaklaşımını getirerek bu varsayımdan kaynaklanan hatayı giderir [12]. Böylece birlikteliklerin bulunmasında daha sağlıklı sonuç veren bir yöntem ortaya çıkar.

Sınama en basit olarak 2x2 tablolar üzerinde uygulanır. Sınamanın esası gözlenen sıklıklarla tabloda bulunması beklenen sıklıkların karşılaştırılmasıdır. Eğer bu karşılaştırma sonucunda gözlenen ve beklenen sıklıklar arasındaki fark çok fazla ise sıfır-hipotezi reddedilir. Beklenen ve gözlenen sıklıklar arasındaki farkı hesaplamak için kullanılan ki-kare formulü aşağıda verilmiştir.

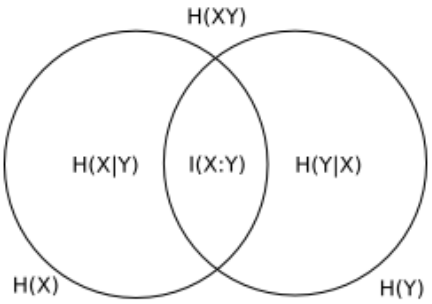
$$X^2 = \sum_{ij} \frac{(Q_{ij} - E_{ij})^2}{E_{ij}} \tag{1.1}$$

Ancak 2x2 lik tablolar için bu formülü aşağıdaki gibi basitleştirebiliriz.

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{1.2}$$

Bouma çalışmasında [15] ortak bilgi çıkarım yöntemini kullanmıştır. Ortak bilgi yöntemi X ve Y rasgele olaylarının birbirine ne denli bağımlı olduklarını ölçen istatistiksel bir yöntemdir. [12]

Bu durum (1.3) bağıntısı ile ifade edilebilir. İki sözcüğün bir birliktelik oluşturup oluşturmadığı konusunda ortak bilgi yönteminde faydalanılabilir. Birlikteliklerde sözcükler ortak bir bilgi taşır. Oysa birliktelik oluşturmuyorlarsa ortak taşıdıkları bilgi sıfır veya sıfıra çok yakın olur.



Şekil 1.4: Ortak bilgi şeması

$$I(XY) = \log_2 \frac{P(XY)}{P(X)P(Y)} \tag{1.3}$$

1.4.3 Anahtar sözcük öbeği çıkarımı

Bilgi çıkarımı konusu, genellikle bir metin üzerinde DDi kullanılarak anahtar bilgileri elde etmeyi hedefler. Bu işlem sırasında örneğin bir kalıba uygun olan verilerin çıkarılması istenebilir. Amaç çok miktardaki veriyi otomatik olarak işleyen bir yazılım üreterek insan katkısını en az seviyeye indirmektir.

Elle anahtar sözcük veya sözcük öbeği çıkarma zahmetli ve zaman alan bir işlemdir. Ayrıca bir çok hataya da neden olabilir. Bunun için bir çok anahtar sözcük öbeği çıkarma algoritması geliştirilmiştir. Bu yöntemler dört başlık altında incelenebilir.

1.4.3.1 İstatistiksel yaklaşımlar

Diğer yöntemlere göre daha basit bir yapıya sahiptirler ve eğitim verisine ihtiyaç duymazlar. Çoğunlukla en sık geçen sözcükleri bulma, TF*IDF ve sözcüğün ilk gözlemlendiği yeri dikkate alarak geliştirilirler. Cohen'in çalışması [16] ve Matsuo ile Ishizuka'nın çalışması [17] örnek verilebilir.

1.4.3.2 Kural tabanlı yaklaşımlar

Bu yaklaşımlar sözcüğün, cümlenin ve metnin özelliklerini kullanırlar. Dile ait biçimbilimsel, sözdizimsel ve anlamsal özelllikler kullanılarak kurallar elde edilir.

Plas ve arkadaşları [18] WORDNET kullanarak konuşma dilinde bulunan anahtar sözcük çıkarımını kullanmıştır. Hulth [19] ise çalışmasında sözdizimsel kurallara ek olarak NP yığınları ve n-gram metodlarını uygulamış ve başarılı sonuçlar elde etmştir.

1.4.3.3 Makine öğrenmesine dayalı yaklaşımlar

İnsanlar tarafından elle seçilmiş veri kümesi eğitim ve sınama kümesi olarak kullanılır. Makine öğrenmesine dayalı anahtar sözcük öbeği çıkarımlarında en önemli çalışmalardan biri Anahtar Sözcük Öbeği Çıkarım Algoritması (Keyword Extraction Algorithm; KEA)'dır. [20]

KEA algoritması eğitim ve çıkarım işlemi olmak üzere iki ana adımdan oluşan gözetimli öğrenme algortimasıdır. Önceden elle işaretlenmiş sözcük öbekleri yardımıyla Naive Bayes Algoritması ile bir model oluşturulur. Bu model sayesinden sınama aşamasında aday sözcük öbeklerinden anahtar sözcük öbekleri seçilir. Aday sözcük öbeklerinin seçimi işlemi girişin temizlenmesi, sözcük öbeklerinin çıkarılması ve sözcüğün gövdesini bulma yani biçimbilimsel analiz işlemlerinden oluşur. Modelin çıkarılmasında TF*IDF değeri ve sözcüğün belgede ilk bulunduğu yer özellik olarak kullanılır.

TF*IDF ağırlıklandırmasında her bir belgedeki sözcüklerin sıklığı rol oynamaktadır. Böylece belgede daha fazla görülen sözcükler varsa (TF, terim sıklığı yüksek) o belge için daha değerli olduğu anlaşılır. Ayrıca IDF tüm belgelerde seyrek görülen sözcükler ile ilgili bir ölçü verir. Bu değer tüm eğitim belgelerinde hesaplanılır. Bu yüzden eğer bir sözcük belgede sık geçiyorsa belge için belirleyici olmadığı düşünülebilir. Eğer sözcük belgelerde çok sık geçmiyorsa o sözcüğün o belge için belirleyici özelliği vardır diyebiliriz.TF*IDF genel olarak sorgu vektörü ile eğitim dokümanı vektörü arasındaki benzerlik oranını bulmak için kullanılır.

Sonuç olarak elle işaretlenen sözcük öbekleri ve KEA tarafında bulunan sözcük öbekleri karşılaştırılır. KEA yönteminin Türkçe metinler için bir uygulaması Pala ve Çiçekli tarafından uygulanmıştır [21]. Biçimbilimsel analiz kısmı Türkçe için değiştirilmiş, arama motorları tarafından da kullanılan etkisiz sözcükler listesi Türkçe dili için seçilmiş ve yeni bir kaç özellik eklenerek model Türkçe dili için uygun hale getirilmiştir. Bu yeni özellikler sayesinde İngilizce için uygulanan KEA'ya yakın başarımında sonuçlar elde edilmiştir.

1.4.3.4 Melez yöntemler

Yukarıda bahsedilen yöntemlerin bir arada kullanılmasıyla oluşan yöntemlerdir. Anahtar sözcük öbeği çıkarım algoritmalarına bir örnek de yapay sinir ağları kullanılarak oluşturulmuş bir modeldir. Wang ve arkadaşları [22] bu yöntemde TF*IDF özelliği kullanılmış bir ağırlık değeri olarak seçilmiştir. Bunun yanında ise başlık ve alt başlıklar, sözcük öbeğinin bulunduğu paragraf sayısı ağırlık değeri olarak kullanılmıştır. Bu ağırlık değerleri yardımıyla oluşturulan Yapay Sinir Ağı algoritması eğitim ve sınama süreçlerine sahiptir. Bu uygulamanın duyarlık ve doğruluk yönteminin başarısı

Bir başka sözcük öbeği çıkarım algoritması ise C4.5 ve GenEx algoritmalarıdır. [23] [24] Her ikisi de güdümlü öğrenme algoritmalarıdır. Öncelikle tüm olası sözcük öbekleri metinden çıkarılır. Sözcük öbeğinin geçme sıklığı, metinde ilk kullanıldığı yer, özel isim olup olmadığı gibi özellikler yardımıyla bir model oluşturulur. Her ne kadar C4.5 algoritmasının başarı oranı az da olsa da sonuçlar tatmin edici değildir.

Kalaycılar ve Çiçekli tarafından önerilen TurkeyX [25] ise anahtar sözcük öbeği çıkarımında kullanılan bir güdümsüz öğrenme modelidir. Bir metinde istatistiksel olarak isim öbeklerinin bulunma sıklığına bakar. KEA ve GenEx'den bazı özelliklerini kullanan bu yöntem ilk olarak tüm aday sözcük öbeği listesini çıkarır. Bu kısımda biçimbilimsel analizi yapılmış sözcükler kullanılır. Daha sonra başka bir sözcük öbeğinin içinde geçen öbeklerden az sözcüklü olanı elenir. Daha sonra en çok geçen sözcük öbekleri anahtar sözcük öbeği olarak adlandırılır. Genel başarı oranı % 25 civarındadır.

2. KURAMSAL ALTYAPI

Bu çalışmada öncelikle, Türkçe dilinin bazı kuralları ve özellikleri çalışmamızın gerektirdiği kadarıyla tanıtılmıştır. Ardından çalışmamızın geliştirme sırasında faydalanılan istatistiksel DDİ araçları açıklanmıştır.

2.1 Türkçe Dili Kuramsal Altyapısı

Her dilin yapısı ve kuralları farklı olduğundan dolayı DDİ çalışmalarında o dile özgü özellikleri bilmek, başarıyı arttıran önemli etkenlerdendir. Bu nedenden DDİ çalışmalarına başlamadan önce çalışmada kullanılacak dilin yapısı ve kurallarını iyi kavramak önemlidir. Bu gerekçelerle Türkçe dilini özellikleri ve kuralları açısından incelemek çalışmamızdaki ilk önceliğimizdir. Çünkü dilimizin yapısını ve kurallarını ne kadar iyi bilirsek, çalışmamızdaki başarı oranı o kadar yüksek olacaktır.

2.1.1 Türkçe dilinde doğal dil işleme

İnsanoğlu çevresini yorumlar ve anadili sayesinde bu yorumu dış dünyayla paylaşır. Her dilin kendine özgü yapısı sayesinde dili kullanarak anlama ve anlatma süreçleri gerçekleşir. DDİ sayesinde doğal diller ile makineler arasında etkileşim sağlanabilir. DDİ, doğal dillerin anlaşılması bilgisayar ortamına taşınması, bu ortamda yaşatılması ve belirli işlerin bu yolla gerçekleştirilmesine yönelik konular ile uğraşır.

Türkçe Ural Altay dil ailesine ait bir dildir. Türkçe yapısı ve üretkenliği açısından bitişken bir dildir. Türkçe bitişken yapısı ile beraber, kuraları ve ses düzeni ile dikkat çeken bir dildir.

Türkçe'nin hemen her DDİ ile ilgilenen tarafından incelenmesinin nedeni, dilin bir takım dilbilimsel olgularda tipik bir örnek oluşturmasıdır, örnek olarak ses uyumu, bitişken sözcük yapısı, sözdizimsel serbestlik, ve öbek yapılarında tamlayanların her zaman tamlananlardan önce gelmesi verilebilir [26].

2.1.2 Türkçe eklemeli bir dildir

Türkçeyi kaynaşık dillerden ayıran en temel özelliklerden biri, eklemeli dil yapısıdır. Türkçe'de yaklaşık olarak 200 adet ek bulunduğu bunlardan 70 tanesininde sıklıkla kullanıldığı bilinmektedir. Bir kök sözcüğüne eklenen çeşitli ekler yardımıyla sözcükler elde edilir. Ancak bu esnada sözcükler ünlü düşmesi, ünsüz yumuşaması, ünlü uyumu gibi nedenlerle değişikliğe uğrar. Köklere eklenen ekler sayesinde oluşan yeni ve farklı anlamda sözcükler türetilir.

2.1.3 Türkçe'nin zor yanları

Köklere yapım ve çekim ekleri eklenerek yeni oluşturulan sözcükler çeşitli özellikler taşır. Türkçe bir çok çekim ekine sahip olamanın yanında, çok üretken şekilde yapım eklerine de sahiptir. DDİ açısından Türkçe'yi işlenmesi zor kılan yapılardan biri sözcüklerin çözümlenmesindeki bu zorluktur. Diğer bir önemli nokta ise bu çözümlemedeki zorluktan kaynaklanan belirsizlik giderme işlemi, bazen öbek ve cümle düzeyinde incelemeyle giderilebilir.

Biçimbilimsel incelemedeki zorluk dışında, Türkçe'de sözlüksel belirsizlikler ve öbek yapısındaki belirsizlikler, Türkçe dil işleme uygulamalarındaki zorluklardan biridir.

Türkçe'nin diğer bir özelliği de sözdizim sırasının oldukça esnek olmasıdır. Dilimizin bu özelliği, özellikle cümlenin öğelerine ayırma işleminde karşılaşılan bir sorundur.

2.1.4 Büyük harflerin kullanılması

Bu bölümde, Türkçe dili için büyük harflerin nasıl kullanıldığı TDK kaynaklarından yaralanarak açıklanmıştır [27].

1. Cümle büyük harfle başlar:

   Örnek: Hayatta en hakiki mürşit ilimdir, fendir. (Atatürk)

2. Cümle içinde tırnak veya yay ayraç içine alınan cümleler büyük harfle başlar ve sonlarına uygun noktalama işareti (nokta, soru, ünlem vb.) konur:

   Örnek: Atatürk "Muhtaç olduğun kudret, damarlarındaki asil kanda mevcuttur!" diyor.

3. İki noktadan sonra gelen cümleler büyük harfle başlar:

   Örnek: Menfaat sandalyeye benzer: Başında taşırsan seni küçültür, ayağının altına alırsan yükseltir. (Cenap Şahabettin)

4. Dizeler büyük harfle başlar.

5. Özel İsimler büyük harfle başlar. Bütün özel isimler (özel ismi oluşturan her sözcük ve onları niteleyen, tanıtan unvanlar) büyük harfle başlar. Büyük harfle başlamazsa cins ismi zannedilebilirler.

6. Takma adlar da büyük harfle başlar: Muhibbi (Kanuni Sultan Süleyman), Demirtaş (Ziya Gökalp), Tarhan (Ömer Seyfettin)

7. Belirli bir tarih bildiren ay ve gün adları büyük harfle başlar:

   Örnek: 29 Mayıs 1453 Salı günü, 29 Ekim 1923, 28 Aralık 1982'de göreve başladı. Lale Festivali 25 Haziran'da başlayacak.

8. Tabela, levha ve levha niteliğindeki yazılarda geçen sözcükler büyük harfle başlar: Örnek: Giriş, Çıkış, Müdür, Vezne, Başkan, Doktor, Otobüs Durağı, Dolmuş Durağı, Şehirler Arası Telefon, 3. Kat, 4. Sınıf, 1. Blok vb.

9. Kitap, bildiri, makale vb.nde ana başlıktaki sözcüklerin tamamı, alt başlıktaki sözcüklerin ise yalnızca ilk harfleri büyük olarak yazılır.

10. Kitap, dergi vb.nde bulunan resim, çizelge, tablo vb.nin altında yer alan açıklayıcı yazılar büyük harfle başlar. Açıklayıcı yazı, cümle niteliğinde değilse sonuna nokta konmaz.

11. Kişi adlarıyla soyadları özel isimdir: Mustafa Kemal Atatürk, İsmet İnönü, Eşref Adalı, Ahmet Haşim, Sait Faik Abasıyanık, Yunus Emre, Karacaoğlan, Âşık Ömer, Wolfgang von Goethe, Vilhelm Thomsen vb.

12. Kişi adlarından önce ve sonra gelen unvanlar, saygı sözleri, rütbe adları ve lâkaplar büyük harfle başlar: Cumhurbaşkanı Mustafa Kemal Atatürk, Kaymakam Erol Bey, Dr. Alâaddin Yavaşça; Sayın Prof. Dr. Hasan Eren; Mustafa Efendi, Zeynep Hanım, Bay Ali Çiçekçi; Mareşal Fevzi Çakmak, Yüzbaşı Cengiz Topel; Mimar Sinan, Fatih Sultan Mehmet, Genç Osman, Deli Petro vb.

13. Akrabalık adı olup lakap veya unvan olarak kullanılan sözcükler büyük harfle başlar: Baba Gündüz, Dayı Kemal, Hala Sultan, Nene Hatun; Gül Baba, Susuz Dede, Telli Baba vb.

14. Cümle içinde özel adın yerine kullanılan makam veya unvan sözleri büyük harfle başlar: Uzak Doğu'dan gelen heyeti Vali dün kabul etti.

15. Saygı bildiren sözlerden sonra gelen ve makam, mevki, unvan bildiren sözcükler büyük harfle başlar: Sayın Bakan, Sayın Başkan, Mektuplarda ve resmî yazışmalarda hitaplar büyük harfle başlar: Sevgili Kardeşim, Aziz Dostum, Değerli Dinleyiciler,

16. Hayvanlara verilen özel adlar büyük harfle başlar: Boncuk, Fındık, Minnoş, Pamuk vb.

17. Millet, boy, oymak adları büyük harfle başlar: Alman, Arap, İngiliz, Japon, Rus, Türk; Kazak, Kırgız, Oğuz, Özbek, Tatar; Hacımusalı, Karakeçili vb.

18. Dil ve lehçe adları büyük harfle başlar: Türkçe, Almanca, İngilizce, Rusça, Arapça; Oğuzca, Kazakça, Kırgızca, Özbekçe, Tatarca vb.

19. Devlet adları büyük harfle başlar: Türkiye Cumhuriyeti, Kuzey Kıbrıs Türk Cumhuriyeti, Amerika Birleşik Devletleri, Suudi Arabistan, Azerbaycan, Kırım Özerk Cumhuriyeti vb.

20. Din ve mezhep adları ile bunların mensuplarını bildiren sözler büyük harfle başlar: Müslümanlık, Müslüman; Hristiyanlık, Hristiyan; Musevilik, Musevi; Budizm, Budist; Hanefilik, Hanefi; Katoliklik, Katolik vb.

21. Din ve mitoloji ile ilgili özel adlar büyük harfle başlar: Tanrı, Allah, İlah, Cebrail, Zeus, Osiris, Kibele vb.

22. Gezegen ve yıldız adları büyük harfle başlar: Merkür, Neptün, Satürn; Halley vb.

23. Düşünce, hayat tarzı, politika vb. anlamlar bildirdiğinde doğu ve batı sözlerinin ilk harfleri büyük yazılır: Batı medeniyeti, Doğu mistisizmi vb.

24. Yer adları (kıta, bölge, il, ilçe, köy, semt vb.) büyük harfle başlar: Afrika, Asya; Güneydoğu Anadolu, İç Anadolu; İstanbul, Taşkent; Turgutlu, Ürgüp; Akçaköy, Çayırbağı; Bahçelievler, Kızılay, Sarıyer vb.

25. Yer adlarında ilk isimden sonra gelen ve deniz, nehir, göl, dağ, boğaz vb. tür bildiren ikinci isimler büyük harfle başlar: Ağrı Dağı, Aral Gölü, Asya Yakası, Çanakkale Boğazı, Dicle Irmağı, Ege Denizi, Erciyes Dağı, Fırat Nehri, Süveyş Kanalı, Tuna Nehri, Van Gölü, Zigana Geçidi vb.

26. 15. Mahalle, meydan, bulvar, cadde, sokak adlarında geçen mahalle, meydan, bulvar, cadde, sokak sözcükleri büyük harfle başlar: Halit Rifat Paşa Mahallesi, Yunus Emre

27. Mahallesi, Karaköy Meydanı, Zafer Meydanı, Gazi Mustafa Kemal Bulvarı, Ziya Gökalp Bulvarı, Nene Hatun Caddesi, Cemal Nadir Sokağı, İnkılap Sokağı vb.

28. 16. Saray, köşk, han, kale, köprü, kule, anıt vb. yapı adlarının bütün sözcükleri büyük harfle başlar: Dolmabahçe Sarayı, İshakpaşa Sarayı, Çankaya Köşkü, Horozlu Han, Ankara Kalesi, Alanya Kalesi, Galata Köprüsü, Mostar Köprüsü, Beyazıt Kulesi, Zafer Abidesi, Bilge Kağan Anıtı vb.

29. Yer bildiren özel isimlerde kısaltmalı söyleyiş söz konusu olduğunda, yer adının ilk harfi büyük yazılır: Hisar'dan, Boğaz'dan, Köşk'e vb.

30. Kurum, kuruluş ve kurul adlarının her sözcüğü büyük harfle başlar: Türkiye Büyük Millet Meclisi, Türk Dil Kurumu, Dil ve Tarih-Coğrafya Fakültesi, Devlet Malzeme Ofisi, Millî Kütüphane, Çocuk Esirgeme Kurumu, Atatürk Orman Çiftliği, Çankaya Lisesi; Anadolu Kulübü, Mavi Köşe Bakkaliyesi; Türk Ocağı, Yeşilay Derneği, Muharip Gaziler Derneği, Emek İnşaat; Bakanlar Kurulu, Türk Dili Dergisi Yayın Danışma Kurulu, Talim ve Terbiye Kurulu Başkanlığı; Türk Dili ve Edebiyatı Bölümü vb.

31. Kanun, tüzük, yönetmelik, yönerge, genelge adlarının her sözcüğü büyük harfle başlar: Medeni Kanun, Türk Bayrağı Tüzüğü, Telif Hakkı Yayın ve Satış Yönetmeliği vb.

32. Kurum, kuruluş, kurul, merkez, bakanlık, üniversite, fakülte, bölüm, kanun, tüzük, yönetmelik ve makam sözleri asılları kastedildiğinde büyük harfle başlar:Türkiye Büyük Millet Meclisi her yıl 1 Ekim'de toplanır. Bu yıl ise Meclis, yeni döneme erken başlayacak.

33. Kitap, dergi, gazete ve sanat eserlerinin (tablo, heykel, beste vb.) her sözcüğü büyük harfle başlar: Nutuk, Safahat, Kendi Gök Kubbemiz, Anadolu Notları, Sinekli Bakkal; Türk Dili, Türk Kültürü, Varlık; Resmî Gazete, Hürriyet, Milliyet, Türkiye, Yeni Asır; Kaplumbağa Terbiyecisi; Yorgun Herkül; Saraydan Kız Kaçırma, Onuncu Yıl Marşı vb.

34. Ulusal, resmî ve dinî bayramlarla anma ve kutlama günlerinin adları büyük harfle başlar: Cumhuriyet Bayramı, Ulusal Egemenlik ve Çocuk Bayramı, 19 Mayıs Atatürk'ü Anma Gençlik ve Spor Bayramı, Ramazan Bayramı, Kurban Bayramı, Nevruz Bayramı, Miraç Kandili; Anneler Günü, Öğretmenler Günü, Dünya Tiyatro Günü, 14 Mart Tıp Bayramı, Hıdırellez vb.

35. Kurultay, bilgi şöleni, çalıştay, açık oturum vb. toplantıların adlarında her sözcüğün ilk harfi büyük yazılır: VI. Uluslararası Türk Dili Kurultayı, Kitle İletişim Araçlarında Türkçenin Kullanımı Bilgi Şöleni, Karamanlı Türkçesi Araştırmaları Çalıştayı vb.

36. Tarihî olay, çağ ve dönem adları büyük harfle başlar: Kurtuluş Savaşı, Millî Mücadele, Cilalı Taş Devri, İlk Çağ, Lale Devri, Cahiliye Dönemi, Buzul Dönemi, Millî Edebiyat Dönemi, Servetifünun Dönemi'nin, Tanzimat Dönemi'nde vb.

37. Özel adlardan türetilen bütün sözcükler büyük harfle başlar: Türklük, Türkleşmek, Türkçü, Türkçülük, Türkçe, Avrupalı, Avrupalılaşmak, Asyalılık, Darvinci, Konyalı, Bursalı vb.

38. Yer, millet ve kişi adlarıyla kurulan birleşik sözcüklerde sadece özel adlar büyük harfle başlar: Antep fıstığı, Brüksel lahanası, Frenk gömleği, Hindistan cevizi, İngiliz anahtarı, Japon gülü, Maraş dondurması, Van kedisi vb.

2.1.5 Özel isimler

Kâinatta tek olan, tam bir benzeri bulunmayan varlıkları karşılayan sözcüklere özel isim denir [28].

Bu varlıklar zaten özel oldukları için adlarına da "özel" denir. "Mehmet" sözcüğü milyonlarca insana ait olabilir, ama bütün "Mehmet"ler tek tek özel oldukları için adları da özeldir.

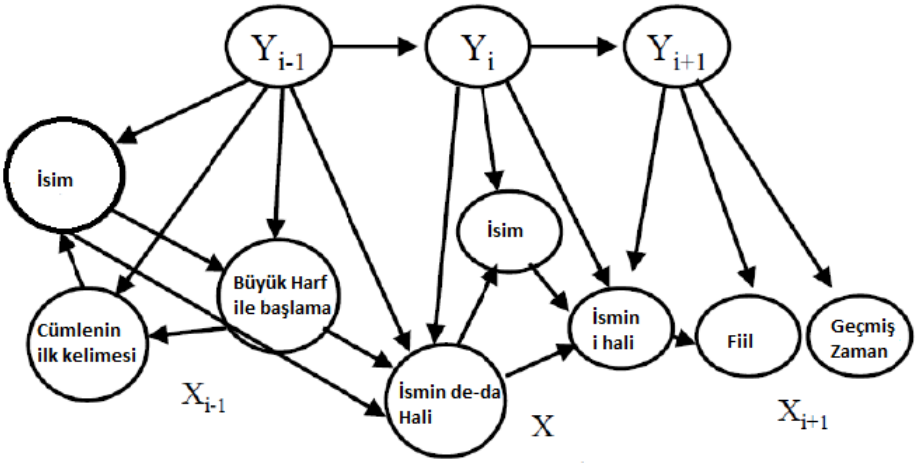
Özel isimler, etiket isimlerdir; varlıklara sonradan takılmış hususî adlardır. Cins isimlerdeki gibi nesne ile sözcük arasında tam bir ilişki yoktur. Özel isimlerin sahipleri tanınmazsa zihinde bir varlık, kavram oluşmaz [28] .

2.1.6 Koşullu rastgele alanlar

Klasik sınıflandırma yöntemleri etiketleme sorunlarını çözmek için sadece o anki durumu göz önüne alır. Oysa dizilim etiketleme sorunlarında o anki durumun olasılığı, çevresindeki durumlardan etkilenir. Bu nedenle dizilim sınıflandırma sorunlarında kullanılmak üzere komşu durumları da hesaba katan yöntemler geliştirilmiştir.

Markov varsayımına göre şu anki durumun olasılığı, sadece bir önceki duruma ve şu anki duruma bağlıdır [29]. Bu varsayım sayesinde dizilimleri sınıflandırmak için bir önceki durumun olaslığını da hesaba katan dizilim sınıflandırıcıları ortaya çıkmıştır. Bunlardan en çok bilineni Saklı Markov Modeli (SMM; Hidden Markov Model; HMM)'dir. Her bir durumu ve her bir geçişi hesaba katan SMM'i gerçeklemek kolay değildir; çünkü Şekil 2.1'de de görüleceği üzere tüm durumlar geçişlerle birbirine dolaylı da olsa bağlantılıdır.

Yukarıda anlatıldığı üzere çok sayıda bağlantı olması sorununu çözmek için, bazı bağlantılar hesaba katılmaz. Bağlantıları hesaba katmamak ise gerçek bir çözüm değildir. SMM'den farklı olarak CRF birleşik olasılık P(X,Y) yerine koşullu olasılık P(Y|X) olarak dizilim sınıflandırma sorununu ele alır. Verilen bir giriş kümesine en uygun etiket dizilimini koşullu olasılık bağıntısıyla çözmeye çalışır.



Şekil 2.1: Saklı Markov Model'de İlişkier

CRF, Lafferty ve arkadaşları [30] tarafından önerilen istatistiksel dizilim sınıflandırmasına dayanan bir makine öğrenmesi yöntemidir. Dizilim sınıflandırıcıları bir dizilim içerisindeki her birime bir etiket atamaya çalışırlar. Olası etiketler üzerinde bir olasılık dağılımı hesaplar ve en olası etiket dizilimini seçerler.

Grafiksel olarak modeli ele alırsak, ard arda gelen düğümlerin birbirlerinin meydana gelme olaslıklarını etkilediği düşünülür. SMM ve Maksimum Entropi Markov Modeli (MEMM; Maximum Entropy Markov Model) gibi dizilim sınıflandırıcısı olan CRF, bir dizilim içerisindeki her bir birime etiket atamaya çalışır [31] . Olası etiketler üzerinden en olası etiket dizilimini seçer. CRF, sözcük sınıfı etiketleme, Varlık İsmi Tanıma ve Gen Tanıma gibi problemlerde sıklıkla başvurulan bir yöntemdir.

Etiket dizisine $y = t_1 ... t_n$, sözcük dizisine $x = c_1 ... c_n$ diyecek olursak:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\} \tag{2.1}$$

Burada Z_x tüm olası etiket dizileri için normalizasyon faktörüdür ve aşağıdaki şekilde tanımlanır:

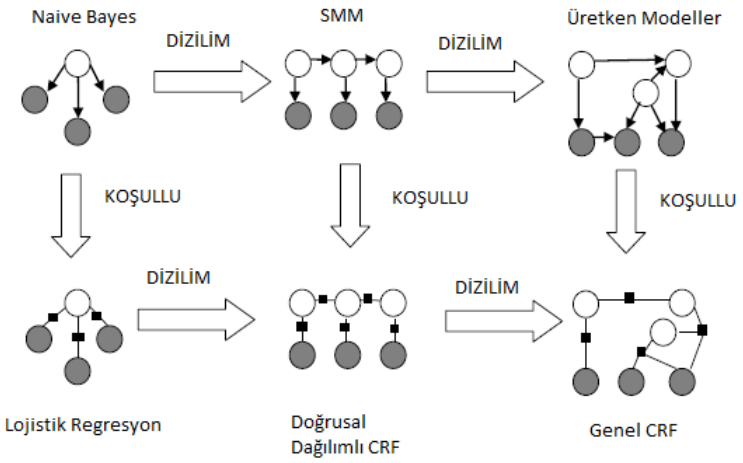
$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y} \in Y^T} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t) \right\}. \tag{2.2}$$

Burada, denklem 2.1'de de görüleceği üzere nitelik fonksiyonu parametreleri t. etiket y_t ve t-1. etiket $y_t - 1$ ve sözcük dizilimi x olan bir fonksiyonudur.

Sonuç olarak en olası etiket dizilimine Y^* dersek. Her bir sözcük dizilimi (x) için en yüksek olaslıklı etiket dizilimini denklem 2.3'de verildiği gibi bulunabilir.

$$Y^* = argmaxP(\mathbf{y}|\mathbf{x}) \tag{2.3}$$

Şekil 2.2'de sınıflandırma yöntemlerinin gelişimi ve bir birleri ile olan ilişkileri gösterilmiştir. Burdan da görüleceği üzere Doğrusal Bağlantılı Koşullu Rastgele Alanlar sınıflandırma yöntemi SMM'in koşullu şekli, Doğrusal Regresyon'un ise dizilim sınıflandırmasına uyarlanmış şeklidir.



Şekil 2.2: Sınıflandırma Yöntemleri Arasındaki İlişki

2.1.6.1 CRF'in etiketlemede kullanımı

Sorunumuz dizilimlere etiket atama sorunu olduğundan, sınıflandırıcı olarak dizilim sınıflandırıcısı kullanmak uygun olacaktır. Yukarıda bahsedildiği üzere SMM gibi birleşik olasılık yerine koşullu olaslık yöntemiyle çalışmamızı modellemek istedik. Sorunu MEMM ile çözmeye çalışabilirdik; ancak CRF, MEMM'den farklı olarak etiket eğilim sorununu (label bias problem) çözmektedir. Yani önceki sözcüklerden az bilgi taşıyanları hesaba katılmaz. MEMM'de nitelik fonsiyonlarının ağırlık değerleri normalize edilmezken, CRF'de nitelik fonksiyonlarının ağırlık değerleri normallize edilir ve bu sayede çok düşük ağırlıklı değerlerle uğraşılmamış olunur.

3. GELİŞTİRİLEN YÖNTEMLER

Tezin amacı Türkçe metnin genelinin özne, yüklem ile zaman ve yer bildiren tümleçlerinin ortaya çıkarılması; bu bilgilerin belgeye etiket olarak işlenmesidir. Bu tür etiketlenmiş olan belgelerin internet üzerinde yapılan sorgulamalarda daha hızlı ve sağlıklı taranacağı açıktır.

Bir metnin etiketini oluşturan bilgilerin çıkarılmasında kullanılacak olan yöntemlerin metnin yazıldığı dile bağlı olacağı açıktır. Bu tez kapsamında ele alınan metinler Türçe olacaktır. Dolasıyla Türkçe'ye özgü bir çözüm geliştirilecektir. En temel yaklaşımda metin içinde geçen sözcüklerin sıklıklarına, birlikte olmalarına bakılarak metnin öznesi, yüklemi, tümleci olmasına karar verilebilir.

Türkçe eklemeli bir dil olması nedeniyle sözcük kökleri çok sayıda ek alabilmektedirler. Bu nedenle öncelikle metin içerisindeki sözcüklerin köklerinin veya gövdelerinin bulunması gerekmektedir. Sözcüklerin kök ve gövdelerini bulabilmek için sözcüklerin bir biçimbilimsel çözümleyiciden geçirilmeleri gerekmektedir. Biçimbilimsel çözümleyiciden geçirilmiş Türkçe sözcüklerin birden çok sonucunun olduğu bilinen bir gerçektir. Bu nedenle çözümlenmiş her bir sözcüğün çözüm sonuçları içinde en doğru sonucu ortaya çıkarmak üzere belirsizliğin giderilmesi gerekmektedir. Üçüncü aşama olarak cümle içerisindeki sözcüklerin nitelikleri çıkarılmalıdır. Bunun için de sözdizimsel çözümleyiciden faydalanılmaktadır.

Bütün bu ön çalışmalar yapıldıktan sonra metnin öznesini, yüklemini, yer ve zaman bilgisini bulma çalışmalarına başlanmıştır. Bu bölüm içinde bu tez kapsamında geliştirilen, bir metnin öznesini, yüklemini, yer ve zaman bilgisini bulmaya yönelik çalışmalar anlatılmıştır.

3.1 Çözümleme Çalışmaları

Bu bölümde faydalanılan biçimbilimsel çözümleme, belirsizlik giderme ve sözdizimsel analiz çalışmaları anlatılmıştır. Bu işlemleri bir paket halinde sunan Eryiğit'in çalışmasından faydalanılmıştır [32].

3.1.1 Biçimbilimsel çözümleme

Sondan eklemeli dillerde bir sözcüğün kökünün türetilmesiyle çok sayıda sözcük türetildiği için biçimbilimsel çözümleyici kullanmak gerekir. Biçimbilimsel analiz işlemi sonucu sözcüklerin kökleri ve ekleri yanında isim, fiil, zarf ve edat gibi tipleri de belirlenmektedir. Bu işlem için Oflazer'in biçimbilimsel çözümleyicisi kullanılmıştır [1]. Oflazer, Türkçe dili için geliştirdiği bu çalışmada iki seviyeli bir yaklaşımı kullanmaktadır. Bu yaklaşımı üç başlık altında incelersek:

1. **Sözlük:** Türkçe dili için köklerin tutulduğu bir yapıdır. Yaklaşık 23.000 kökün bulunduğu bir sözlükten faydalanılmaktadır.
2. **Biçimbilimsel Kurallar:** Türkçe'nin sondan eklemeli bir dil olması, yani köke eklenen eklerle yeni sözcüklerin oluşması, sonlu durum makineleriyle modellenmektedir.
3. **İmla Kuralları:** Sesli uyumu, sesli düşmesi, sesli daralması gibi Türkçe diline özgü kurallar da kullanılmaktadır.

Örnek bir cümle için biçimbilimsel çözümleyici işlemine girmeden önceki cümle ile, çözümleyicinin çıktısı aşağıdaki Çizelge 3.1'de gösterilmiştir. Burada dikkat edilmesi gereken önemli noktalardan biri ise biçimbilimsel çözümleyiciye giren veriler yazım ve noktalama hatası içeriyorsa biçimbilimsel çözümleyicinin sonucu güvenilir olmaz. Oysa haber metinlerinde yazım ve noktalama hatası bulunabilir. Bu nedenle şimdilik yazım ve noktalama hatasından dolayı etiketlerde çıkabilecek hatalar ihmal edilmiştir.

Burada elde edilen biçimbilimsel çözümleyicinin çıktısı belirsizlik gidericinin girdisi olarak kullanılacaktır.

Çizelge 3.1: Biçimbilimsel çözümleyiciye bir örnek

Çözümleyici Girdisi	Çözümleyici Çıktısı
Her	Her Her Noun+Prop+A3sg+Pnon+Nom
şey	şey şey+Noun+A3sg+Pnon+Nom
çok	çok çok+Adverb çok çok+Det çok çok+Adj çok çok+Postp+PCAbl
güzel	güzel güzel+Adj
olacak	olacak ol +Verb+Pos+Fut+A3sg olacak ol +Verb+PosDB+Adj+FutPart+Pnon
.	. . +Punc

3.1.2 Belirsizlik giderme

Türkçe'de herbir sözcüğün ortalama iki biçimbilimsel çözümünün olduğu bilinmektedir. Dolayısıyla biçimbilimsel çözümlemesi yapılmış her bir sözcüğün, içinde bulunduğu cümleye göre en doğru çözümünün bulunması işlemine belirsiz giderme diyoruz.

Sak tarafından hazırlanan [2] belirsiz giderici yardımıyla biçimbilimsel analiz sonucu üretilen sonuçlardan cümleye en uygunu seçilmiştir. Sak tarafından gerçeklenmiş bu çalışmanın başarımı %98'dir. Sak'ın sözdizimsel çözümleyicisine bir örnek Çizelge 3.2'de verilmiştir.

Çizelge 3.2: Belirsizlik gidericiye bir örnek

Belirsizlik Giderici Girdisi	Belirsizlik Giderici Çıktısı
Her Her Noun+Prop+A3sg+Pnon+Nom	Her Noun+Prop+A3sg+Pnon+Nom
şey şey+Noun+A3sg+Pnon+Nom	şey şey+Noun+A3sg+Pnon+Nom
çok çok+Adverb çok çok+Det çok çok+Adj çok çok+Postp+PCAbl	çok çok+Adverb
güzel güzel+Adj	güzel güzel+Adj
olacak ol +Verb+Pos+Fut+A3sg olacak ol +Verb+PosDB+Adj+FutPart+Pnon	olacak ol+Verb+Pos+Fut+A3sg
.	. . +Punc

3.1.3 Sözdizimsel çözümleme

Son olarak belirsizlik gidericiden çıkan sonuçlar CONLL formatına dönüştürülerek, Eryiğit'in [33], [3] hazırlamış olduğu sözdizimsel çözümleyicisi yardımıyla etiketlenmiştir. Eryiğit çalışmasında elle etiketlenmiş bilgilerden faydalanarak sistemi, Destek Vektör Makinesi (DVM) yardımıyla eğitmiştir. Sözdizimsel çözümleyici örnek Çizelge 3.3'de verilmiştir.

Çizelge 3.3: Sözdizimsel çözümleyiciye bir örnek

Etiketleyici Girdisi
1 Her Noun Prop Prop A3sg|Pnon|Nom _ _ _ _
2 şey şey Noun Noun A3sg|Pnon|Nom _ _ _ _
3 çok çok Adv Adv _ _ _ _ _
4 güzel güzel Adj Adj _ _ _ _ _
5 olacak ol Verb Verb Pos|Fut|A3sg _ _ _ _
6 . . Punc Punc _ _ _ _ _
Etiketleyici Çıktısı
1 Her Noun Prop Prop A3sg|Pnon|Nom 5 SUBJECT _ _
2 şey şey Noun Noun A3sg|Pnon|Nom 5 SUBJECT _ _
3 çok çok Adv Adv _ 4 MODIFIER _ _
4 güzel güzel Adj Adj _ 5 MODIFIER _ _
5 olacak ol Verb Verb Pos|Fut|A3sg 6 SENTENCE _ _
6 . . Punc Punc _ 0 ROOT _ _

3.2 Geliştirilen Etiketleme Modeli

Bölüm 3.1'de anlatılmış olanlar, başka araştırmacılar tarafından hazırlanmış olan biçimbilimsel çözümleyici, belirsizlik giderici ve sözdizimsel çözümleyicilerdir. Bu bölümde ise bu tez tarafımızdan yapılan çalışmalara yer verilmiştir.

Bölüm 3.1'de anlatılan çalışmalardan faydalanılarak önce her bir cümlenin öznesi, yüklemi ve tümleçleri bulunmaya çalışılmıştır. Ardından bir haber metninin geneline ilişkin özne, yüklem, yer ve zaman bilgisi çıkarılmaya çalışılmıştır. Bu çalışmada amacımız bir metnin bütününün ifade ettiği fikrin öznesi, yüklemi ile yer ve zaman bilgilerini çıkarmaktır.

Tez çalışmasının ilk kısmında yaptığımız çalışma ile Kural Tabanlı yaklaşımlar yardımıyla hedefimize ulaşmak istenmiştir. Ancak başarı oranın düşük olmasından

dolayı makine öğrenmesi yöntemlerine başvurulmuştur. Türkçe gibi kurallı bir dilde, hedefe ulaşmak için bir çok kural koymak gerekir. Bir çok kural ise gözden kaçabilecek niteliktedir. Bu nedenle probleme makine öğrenmesi problemi şeklinde yaklaşmak ve gerekli ilişkileri makinenin bulması için sistemi eğitmek farklı bir yaklaşım olacaktır.

Tezimizde amaç bir dökümana ait özne, yüklem, yer ve zaman etiketlerini bulmaktır. Bu bakış açısı ile problem dizilimlerden oluşan, bir sınıflandırma problemidir. Her bir sözcük ya belgenin öznesinidir, ya yüklemidir, ya yeridir, ya zamanıdır veya bunlardan hiçbiridir. Çizelge 3.4'de etiketler ve anlamları gösterilmiştir.

Çizelge 3.4: Etiketler ve Anlamları

Çözümleyici Girdisi	**Çözümleyici Çıktısı**
SUBJ	Metnin Öznesi
PRED	Metnin Yüklemi
LOC	Metnin Yeri
DATE	Metnin Zamanı
O	Yukarıdakilerden hiç biri

Problemimizi dizilim sınıflandırma problemi olarak ele aldığımızdan ve sistemi eğitmedeki veriminden dolayı koşullu rastgele alanlar yöntemi ile sistemi modellemeyi tercih ettik. Bu çalışmada, doğrusal zincir koşullu rastgele alanlar kullanılmıştır.

3.2.1 Niteliklerin belirlenmesi

Koşullu Rastgele Alanlarda sistemi eğitmek için nitelikleri belirlemek gerekir. Kural tabanlı çalışmalarımızda elde ettiğimiz deneyimlerimizin sonucu olarak, CRF için gerekli olan nitelikleri belirledik.

Kullanılan nitelikleri dört ana başlıkta sıralayabiliriz:

- Biçimbilimsel nitelikler
- Sözdizimsel nitelikler
- Yapısal nitelikler
- Kural tabanlı nitelikler

3.2.2 Kural Tabanlı Nitelikler

Bu kısımda bahsedilen nitelikler, sorunu kural tabanlı olarak çözmek için yapılan araştırmalardan sonra çıkarılmıştır. Sadece kural tabanlı olarak sorunu çözmeye çalışarak, istenilen başarım oranına erişemedik. Fakat bu kurallardan bazılarının başarım oranını olumlu etkiledeği gözlemlenmiş ve CRF için nitelik olarak kullanılmıştır.

3.2.2.1 Özne ve yer etiketleri için kural tabanlı nitelikler

Tanım olarak bir metnin öznesinden bahsedilemez. Bizim çalışmamızda bir metnin bütünün öznesi olarak kabul edilebilecek sözcüğe metnin öznesi adını vermekteyiz. Çalışmamızda kullanılan metin kümesi, haber nitelikli olduğu için, özne etiketleri ile yer etiketleri genelde özel isimlerden oluşmaktadır. Daha önce değinildiği gibi özne tek bir sözcük olabileceği gibi sözcük öbeği de olabilir.

Aşağıda ayrıntıları anlatılan kurallar ile sözcüğün bir özel isim kümesine dahil olup olmadığı olasılığı bir nitelik olarak alınmıştır. Buna göre eğer bir sözcük özel isim kümesinin bir üyesi ise ve bu sözcük cümlede özne olarak kullanılmış ise 'POSSUB' isimli bir nitelikle tanımlanır. Eğer sözcük özel isim kümesinin bir üyesi ise ve bu sözcük cümlede dolaylı tümleç olarak kullanılmış ise bu sözcük 'POSLOC' olacak şekilde nitelik olarak tanımlanır.

Özel isim öbekleri

Türkçe'de yazım kuralları gereği özel isimler büyük harfle başlar. Ancak 2. Bölümde anlatıldığı gibi büyük harfle başlayan her sözcük özel isim değildir. Buna en iyi örnek cümle başındaki sözcüktür. Cümlelerin ilk sözcükleri özel isim olup olmadığına bakılmaksızın büyük harfle başlarlar.

Yan yana geçen büyük harfli sözcükler özel isim öbeği olabilir. Ancak bu kural tek başına yeterli değildir; çünkü bazı özel isim öbekleri içinde bağlaç bulunabilir. Örnek olarak "Çalışma ve Sosyal Güvenlik Bakanlığı", "Türk Dil ve Tarih Kurumu" örnekleri verilebilir.

Buna göre aşağıdaki kurallar çıkarılabilir.

Kural 1: Eğer bir sözcüğün ilk harfi büyük harf ise ve cümle başında değil ise özel isimdir.

Bu üç kurala özgü örnekler Çizelge 3.4, Çizelge 3.5 ve Çizelge 3.6'da verilmiştir.

Çizelge 3.5: Kural 1 ile çıkarılan özel isim grupları

Cümleler	**Özel İsim Grupları**
Söz konusu anlaşmaya Obama ile Afganistan Devlet Başkanı Hamid Karzai imza atacak.	Obama Afganistan Devlet Başkanı Hamid Karzai
Yetkililer, her gün birkaç yabancı şirketin Türkiye'deki bu alanlar için ortaklık teklifi yaptığını söyledi.	Türkiye
İlk kez dolaylı bir vergiden vazgeçiliyor. Maliye Bakanı, Avrupa Kültür Vergisi'ni kaldıracaklarını belirtti.	Maliye Bakanı Avrupa Kültür Vergisi

Kural 2: Eğer bir sözcük cümlenin ilk sözcüğü ise ve özel isim türü ile etiketlenmiş ise (Prop etiketli) özel isimdir.

Çizelge 3.6: Kural 2 ile çıkarılan özel isim grupları

Cümleler	**Özel İsim Grupları**
İstanbul Teknik Üniversitesi'nde gerçekleşen konferansa yaklaşık 200 kişi katıldı.	İstanbul Teknik Üniversitesi
Celal Gündüz annesini Kısıklı'daki evinde ziyaret etti.	Celal Gündüz Kısıklı
Fenerbahçe'de şike tartışmaları yaşanıyor.	Fenerbahçe

Kural 3: Eğer iki özel isim arasında bağlaç var ise bu bağlaç özel isim öbeğine aittir. Örnek olarak, "Çalışma ve Sosyal Güvenlik Bakanlığı", "Gençlik ve Spor Bayramı" verilebilir.

Ancak bu üç kural tek başına yeterli değildir; çünkü iki özel isim grubu yan yana görülebilir. Örnek olarak "Ali ve Ayşe", "Ayşe ile Ankara'ya" verilebilir.

Özel isim öbekleri için sınır kuralları

Özel isim öbeklerini birbirinden ayıran bir takım sınır kuralları belirlenmiştir.

Sınır Kuralı 1: Eğer özel bir isim –i halini almış bir sözcük ise bu sözcük özel isim öbeğinin son sözcüğüdür. Sadece son sözcüğü –i hali almış özel isim

Çizelge 3.7: Kural 3 ile çıkarılan özel isim grupları

Cümleler	**Özel İsim Grupları**
Daniel Craig İstanbul'daki ilk akşamında Nişantaşı Köşebaşı Restaurant'daydı.	Daniel Craig İstanbul Nişantaşı Köşebaşı Restaurant
Celal Gündüz Kısıklı'ya gitti.	Celal Gündüz Kısıklı
Milli Savunma Bakanı Çankaya Köşkü'ne Cumhurbaşkanı Gül ile görüşmeye gitti.	Milli Savunma Bakanı Çankaya Köşkü Cumhurbaşkanı Gül

öbekleri içerisinde bağlaç barındırabilir. Örnek olarak, "Savunma ve Sosyal Güvenlik Bakanlığı", "Öğrenci Seçme ve Yerleştirme Sınavı"

Sınır Kuralı 2: Herhangi bir noktalama işareti o sözcük öbeğinin bittiğine işaret eder. Bu sayede virgün "," ve kesme "'" işareti sadece sözcük öbeğindeki son sözcükte bulunabilir.

3.2.3 Biçimbilimsel nitelikler

Biçimbilimsel nitelikler biçimbilimsel çözümleyici ve sonrasında belirsizliği giderilmiş çözümler içerisinden seçilir ve dokuz ana başlık halinde aşağıdaki gibi sınıflandırılabilir. Temel olarak sözcük sınıfları ve ek sınıflarından oluşan bu nitelikler her bir sözcüğün biçimbilimsel belirsizlik gidericiden çıkmış hali işlenerek çıkartılır.

3.2.4 Sözdizimsel nitelikler

Sözdizimsel çözümleyiciden çıkan çözümlerden o sözcüğe ait sözdizimsel niteliği çıkarılır. Sözdizimsel etiketler aşağıda sıralanmıştır.

3.2.5 Yapısal nitelikler

Yapısal nitelikler, dilin ve işlenen metnin yapısını ortaya koyan niteliklerdir.

3.2.5.1 Metnin sırası

İlgili metnin ismini temsil eder. Bu sayede her bir sözcüğe metin bazlı sınıflandırma niteliği katılmış olunur.

Çizelge 3.8: Biçimbilimsel Çözümleyici Nitelikleri

Nitelik Sınıfı	**Nitelikler**
Ana Sözcük Sınıfı	Adj, Adv, Conj, Det, Dup, Interj, Noun, Num, Postp Pron,Punc, Verb
Alt Sözcük Sınıfı	Able, Acquire, ActOf, Adamantly, AfterDoingSo, Agt Almost, As, AsIf, AsLongAs, Become, ByDoingSo, Card, Caus, DemonsP, Dim, Distrib, EverSince, FeelLike, FitFor, FutPart, Hastily, InBetween, Inf, Inf1, Inf2, Inf3, JustLike, Ly, Ness, NotState, Ord, Pass, PastPart, PCAbl, PCAcc, PCDat, PCGen, PCIns, PCNom, Percent, PersP, PresPart, Prop, Quant, QuesP, Range, Ratio,Real, Recip, ReflexP, Rel, Related, Repeat, Since, SinceDoingSo, Start, Stay, Time, When, While, With, Without, Zero
Kişi Ekleri	A1pl, A1sg, A2pl, A2sg, A3pl, A3sg
İyelik Ekleri	P1pl, P1sg, P2pl, P2sg, P3pl, P3sg, Pnon
Ad Durum Ekleri	Abl, Acc, Dat, Equ, Gen, Ins, Loc, Nom
Zaman Ekleri	Aor, Desr, Fut, Imp, Neces, Opt, Pres, Prog1, Prog2, Cop Cond, Past, Narr
Çatı Ekleri	CompCond, CompNarr, CompPast
Diğer Ekleri	Cop Neg, Pos

Çizelge 3.9: Sözdizimsel Nitelikler

Sözdizimsel Nitelikler
MODIFIER, OBJECT, SUBJECT,. DATIVE.ADJUNCT, LOCATIVE.ADJUNCT, SENTENCE, DERIV , CLASSIFIER, DETERMINER, INTENSIFIER POSSESSOR, COORDINATION

3.2.5.2 Cümle sırası

Metin içerisindeki cümle sırasını temsil eder. Bu sayede cümle bazlı olarak sınıflandırma niteliği katılmış olunur.

3.2.5.3 Sıklık

İlk olarak biçimbilimsel çözümleme, ardından belirsizlik giderici ve daha sonra sözdizimsel çözümleme yapılmıştır. Bu işlemlerin sonunda biçimbilimsel ve sözdizimsel nitelikler ile birlikte sözcük köklerine ve gövdelerine ulaşılmıştır. Bu sözcük köklerinin sayısını, dökümandaki tüm sözcüklerin sayısına bölerek sıklığı hesaplanmıştır. Ancak makine öğrenmesi açısından anlamlı olabilmesi için bu sıklık 0

ile 9 arasındaki doğal sayı değerleriyle ifade edilmiştir. Bu sayede 10 adet sınıf olacak şekilde sıklık değeri nitelendirilmiştir.

$$\text{Sıklık Sınıfı} = (\text{Dogal Sayı})\frac{\text{SayıToplam(İlgili Sözcük)*100)}}{\text{SayıToplam(Tüm Sözcükler)}} \quad (\mathbf{3.1})$$

3.2.5.4 İlk gözlemlendiği yer

Bir sözcüğün ilgili metinde ilk gözlemlendiği yeri ifade etmek için kullanılır. Makine öğrenmesi açısından anlamlı olabilmesi için bu sıklık 0 ile 9 arasındaki doğal sayı değerleriyle ifade edilir.

$$\text{İlk Görülme Sınıfı} = (\text{Doğal Sayı})\frac{\text{(İlkGörüldüğüSözcükNumarası(İlgili Sözcük)+1)*10)}}{\text{SayıToplam(Tüm Sözcükler)}} \quad (\mathbf{3.2})$$

3.2.5.5 Büyük harfle başlama

Bir sözcüğün büyük harfle başlayıp başlamadığını ifade eder.

3.2.6 Nitelik Seçimi ve Performans İlişkisi

Nitelikleri seçerken koşullu rastgele alanlar yönteminin bağıntısından da anlaşılacağı üzere nitelik fonksiyonlarının ağırlıkları normalize edilmiş şekildedir. Uygulamanın gerçeklenmesi kısmında da bahsedileceği üzere fazla nitelik seçmenin sistemin performansına etkisi ihmal edilebilir düzeydedir. Bu nedenle MrMr gibi nitelik seçme işlemine veya en çok etki eden nitelikleri bulma gibi bir yönteme başvurulmamıştır.

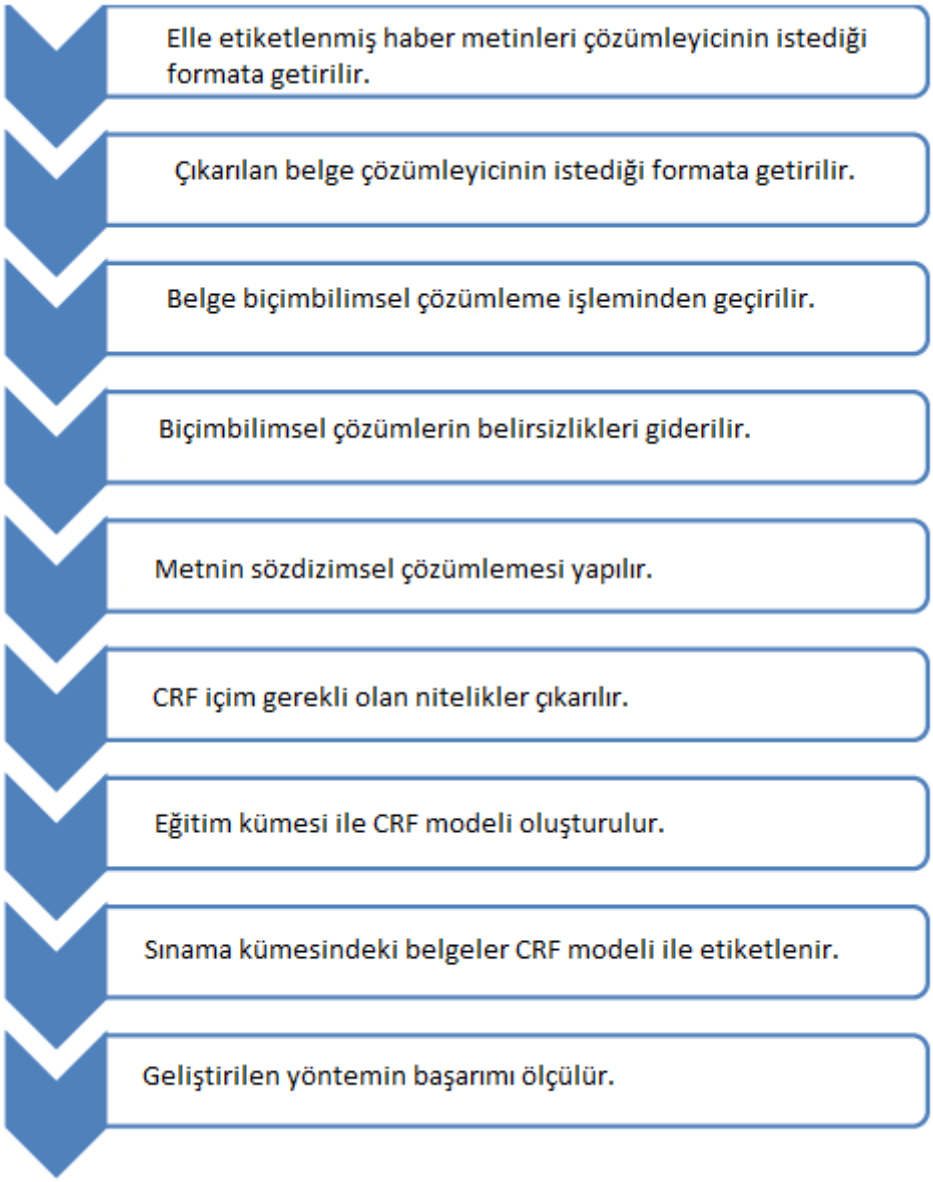
4. UYGULAMANIN GELİŞTİRİLMESİ

Bölüm 3'de bu tez için geliştirilmiş olan yöntemler tanıtılmıştır. Bu bölümde ise bu bilgiler ışığında yapılan geliştirme çalışmaları anlatılacaktır. Geliştirme kısmını altı ana başlıkla inceleyebiliriz:

1. Metinlerinin Toplanması
2. Metinlerinin Elle Etiketlenmesi
3. Metinlerinin Önişlenmesi
4. Metinlerinin Toplanması
5. Metinlerinin Etiketlenmesi
6. Etiketlerin Karşılaştırılması

Bu aşamalar Şekil 4.1'de özetlendiği şekilde program geliştirmiştir.

Çalışmalar JAVA programlama dili kullanılarak Eclipse Helios JavaEE geliştirme ortamında geliştirilmiştir.



Şekil 4.1: Geliştirilen yöntemin aşamaları

4.1 Metinlerinin Toplanması

Eğitim ve sınama kümesinde kullanılmak üzere internet üzerinden haber dağıtımcılarının RSS bağlantıları üzerinden çekilmiştir. Bu işlem için Java kütüphanelerinden faydalanılmıştır. Çekilen haber metinlerinin başlıkları haricindeki metnin gövdesi elle işlenmek üzere XML kalıbına uygun olarak dosyaya kaydedilmiştir. Bir örnek metin Şekil 4.2'de verilmiştir.

Veriler RSS ile çekilen haber metinleridir. Bu metinler XML kalıbına uygun olarak kaydedilmiştir. Bu metinerin başına özne, yüklem, yer ve zaman bilgileri XML kurallarına uygun olarak girilebilmesi için belgenin başlık kısımları eklenmiştir. Tüm bu belgede UTF-8 alfabesi kullanılmıştır.

Böylece hazırlanmış olan belgerin elle etiketlenmesini sağlamak üzere bir program geliştirilmiş ve bu program sayesinde de eğitim ve sınama kümesinde yer alacak belgeler etiketlenmiştir. Dosya isimleri elle etikenmenin yapıldığını temsil etmek için anlaşılır şekilde kaydedilir. Örneğin birinci metin "a1.xml" şeklinde kaydedilir. n'inci metin "an.xml" şeklinde kaydedilir.

Şekil 4.2'de elle etiketlenmeye hazır haldeki dosyalara bir örnek verilmiştir.

```
<DOCUMENT>
<SUBJECT></SUBJECT>
<PREDICATE></PREDICATE>
<LOCATION></LOCATION>
<DATE></DATE>
<NEWS>
Adli Tıp Kurumu hırsızı sabıkalı çıktı.
Alınan bilgiye göre, Adli Tıp Kurumu Kimya İhtisas Dairesi'nde 2 personelin
bazı şahsi eşyalarının çalınmasına ilişkin çalışma başlatan Asayiş Şube
Müdürlüğü Hırsızlık Büro Amirliği ekipleri, şüphelinin kullandığı plakanın
kiralık bir  otomobile ait olduğunu ve aracın Avcılar'dan kiralandığını belirledi.
Polis, daha sonra da Cihan A. adına sahte kimlik düzenlemiş olan Gürsoy Erinci'yi,
araçta bulunan CPS cihazı sayesinde, Avcılar'da, olayda kullandığı otomobilde
yakaladı.
Asayiş Şube Müdürlüğüne götürülen şüphelinin, daha önce Adli Tıp Kurumuna
su veren şirkette çalıştığı, 'hırsızlık' suçundan poliste 25 kaydı bulunduğu,
2006'da Bahçelievler'de karıştığı hırsızlık suçundan arandığı ve yine aynı suçtan
bulunduğu Ankara Sincan Açık Cezaevinden, 17 Ekim'de izin ihlali yaparak kaçtığını
belirledi.
Emniyetteki sorgusunda suçunu itiraf eden Erinci'nin, 'Cezaevine girmeden önce Adli
Tıp Kurumuna su veren firmada çalışıyordum. Ara sıra da su götürdüğüm oluyordu. Kendimi
su firması görevlisi olarak tanıtıp girdim. Çaldığım kredi kartlarıyla, kontör ve takım
elbise aldım. Ziynet eşyalarını da bozdurdum. Pişmanım'' dediği öğrenildi.
Adli Tıp Kurumu çalışanları tarafından da teşhis edilen Gürsoy Erinci, Emniyetteki işlem
ardından Bakırköy Adliyesine sevk edildi.
</NEWS>
</DOCUMENT>
```

Şekil 4.2: Elle etiketlenmeye hazır haber belgesi örneği

4.2 Metinlerinin Elle Etiketlenmesi

Haberlerin elle etiketlenmesi için kullanıcılara hazırlamış olduğumuz dosyalar verilmiştir. Kullacılardan beklenen haber metnini okuduktan sonra, metni en iyi şekilde temsil eden özne, yüklem, yer ve zaman etiketlerinin atanması işlemini gerçekleştirmeleridir. Kullanıcılar eğer bir etiket tipi için metinden bir varlık

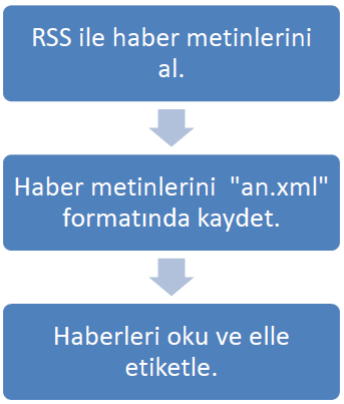
çıkaramazlarsa "-" işaretini girmelidirler. Burada "-" şeklinde etiketlenmesinden amaç kullanıcının bu etiket tipine uygun bir varlığın metinden çıkarılmadığının kullanıcı tarafından karar verildiğini anlamaktır.

Şekil 4.3'de elle etiketlendikten sonraki haline bir örnek gösterilmiştir.

```
<DOCUMENT>
<SUBJECT>Adli Tıp Kurumu,Gürsoy Erinci</SUBJECT>
<PREDICATE>sabıkalı_çık</PREDICATE>
<LOCATION>Adli Tıp Kurumu Kimya İhtisas Dairesi</LOCATION>
<DATE>17 Ekim</DATE>
<NEWS>
Adli Tıp Kurumu hırsızı sabıkalı çıktı.
Alınan bilgiye göre, Adli Tıp Kurumu Kimya İhtisas Dairesi'nde 2 personelin
bazı şahsi eşyalarının çalınmasına ilişkin çalışma başlatan Asayiş Şube
Müdürlüğü Hırsızlık Büro Amirliği ekipleri, şüphelinin kullandığı plakanın
kiralık bir  otomobile ait olduğunu ve aracın Avcılar'dan kiralandığını belirledi.
Polis, daha sonra da Cihan A. adına sahte kimlik düzenlemiş olan Gürsoy Erinci'yi,
araçta bulunan CPS cihazı sayesinde, Avcılar'da, olayda kullandığı otomobilde
yakaladı.
Asayiş Şube Müdürlüğüne götürülen şüphelinin, daha önce Adli Tıp Kurumuna
su veren şirkette çalıştığı, 'hırsızlık' suçundan poliste 25 kaydı bulunduğu,
2006'da Bahçelievler'de karıştığı hırsızlık suçundan arandığı ve yine aynı suçtan
bulunduğu Ankara Sincan Açık Cezaevinden, 17 Ekim'de izin ihlali yaparak kaçtığını
belirledi.
Emniyetteki sorgusunda suçunu itiraf eden Erinci'nin, 'Cezaevine girmeden önce Adli
Tıp Kurumuna su veren firmada çalışıyordum. Ara sıra da su götürdüğüm oluyordu. Kendimi
su firması görevlisi olarak tanıtıp girdim. Çaldığım kredi kartlarıyla, kontör ve takım
elbise aldım. Ziynet eşyalarını da bozdurdum. Pişmanım'' dediği öğrenildi.
Adli Tıp Kurumu çalışanları tarafından da teşhis edilen Gürsoy Erinci, Emniyetteki işlem
ardından Bakırköy Adliyesine sevk edildi.
</NEWS>
</DOCUMENT>
```

Şekil 4.3: Elle etiketlenmiş haber belgesi örneği

Metinlerin elle işaretlenme süreci Şekil 4.4'de gösterilmiştir

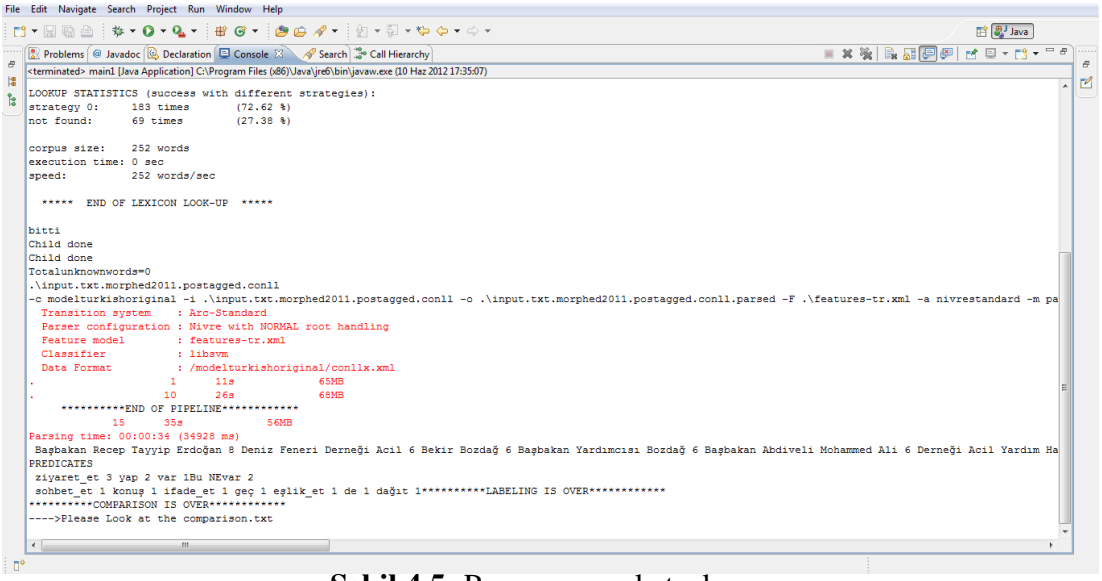


Şekil 4.4: Haber belgelerinin elle etiketlenme süreci

4.3 Metinlerin Önişlenmesi

Bu kısımda cümleler ve sözcükler için kullandığımız biçimbilimsel çözümleyici, belirsizlik giderici ve sözdizimsel çözümleyici işlemleri tanıtılmaktadır. Bu işlem, Bölüm 3'de anlatıldığı üzere, Eryiğit [33] tarafından biraraya getirilip tek bir paket halinde sunulan çalışmadır. Bu paket sayesinde metindeki tüm sözcüklerin biçimbilimsel çözümleri ve sözdizimsel karşılıklarına erişilmiştir. Bu paket Java programla dilinde hazırlandığı için programımız tarafından kolaylıkla kullanılabilir durumdadır.

Şekil 4.5'de de görüldüğü üzere "main.java" dosyasının Eclipse Helios JavaEE geliştirme ortamında çalıştırılmasıyla konsolda programın ilerlemesi gösterilmiştir.



Şekil 4.5: Programın çalıştırılması

4.4 Metinlerinin Etiketlenmesi

Haber metinlerinden özne, yüklem, yer ve zaman etiketletleri çıkarılması işlemleri, bir Bölüm 3'de çeşitli nitelikler çıkarılarak, bu nitelikleri kullanıp sistemin eğitilmesinden oluşur.

4.4.1 Niteliklerin belirlenmesi

Niteliklerin belirlenmesi aşaması, makine öğrenmesinin başarımını doğrudan etkileyen bir aşamadır. Çalışmamızda sınıflandırılmak istenilen varlıklar sözcüklerdir. Sözcüklerin nitelikleri Çizelge 4.1'de görülebileceği üzere dört ana sınıf altında nitelikler gösterilmiştir. Niteliklerin seçiminde kural tabanlı yaklaşım ile yaptığımız çalışmadan faydalanılmıştır.

Nitelik dosyasına bir örnek Şekil 4.6'da verilmiştir.

```
80 1 Zirve zirve Noun Noun SUBJECT CAPITAL NONLASTWORD POSSUB 0 0 A3sg Pnon Nom A3sg Pnon Nom
80 1 Yayınevi'nde Yayınevi'nde Prop Prop CLASSIFIER CAPITAL LASTWORD NONPOS 0 0 A3sg Pnon Gen
80 1 3 3 Num Card MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Ins
80 1 kişinin kişi Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 With
80 1 katledilmesiyle  Noun NInf MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg Pnon Gen
80 1 ilgili ilgi Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 Rel
80 1 davanın dava Noun Noun POSSESSOR NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Loc
80 1 bugünkü bugün Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom
80 1 duruşmasında duruşma Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom
80 1 tutuklu tutuklu Noun Noun CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom A3sg Pnon Nom
80 1 sanık sanık Noun Noun CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg Pnon Nom
80 1 Emre Emre Noun Prop CLASSIFIER CAPITAL NONLASTWORD POSSUB 0 1 A3sg Pnon Nom
80 1 Günaydın günaydın Noun Noun SUBJECT CAPITAL NONLASTWORD POSSUB 0 1 A3sg Pnon Nom
80 1 "5 "5 Prop Prop SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Nom
80 1 saf saf Noun Noun CLASSIFIER NONCAPITAL NONLASTWORD POSSENT 0 0 A1pl P1pl Acc
80 1 genç genç Noun Noun OBJECT NONCAPITAL NONLASTWORD POSSENT 0 0 A3sg Pnon Nom
80 1 kendi kendi Pron ReflexP MODIFIER NONCAPITAL NONLASTWORD POSSENT 0 0 ByDoingSo
80 1 kendimizi kendi Pron ReflexP OBJECT NONCAPITAL NONLASTWORD POSSENT 0 0 A3sg Pnon Acc
80 1 gaza gaza Noun Noun OBJECT NONCAPITAL NONLASTWORD POSSENT 0 0 A3sg Pnon Nom
80 1 getirerek getir Adv Adv MODIFIER NONCAPITAL NONLASTWORD POSSENT 0 0 Pos Past A3sg
80 1 olayı olay Noun Noun OBJECT NONCAPITAL NONLASTWORD POSSENT 0 1
80 1 gerçekleştirdik" gerçekleştirdik" Prop Prop OBJECT NONCAPITAL NONLASTWORD POSSENT 0 0
80 1 dedi de Verb Verb SENTENCE NONCAPITAL NONLASTWORD POSSENT 0 2 A3sg Pnon Nom
80 2 MALATYA Malatya Noun Prop SENTENCE CAPITAL NONLASTWORD POSSENT 0 0 A3sg Pnon Nom A3sg P3sg Loc A3sg
80 2 Zirve Zirve Noun Prop CLASSIFIER CAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom With
80 2 Yayınevi'nde Yayınevi Noun Prop LOCATIVE.ADJUNCT CAPITAL LASTWORD POSLOC 0 0
80 2 biri biri Pron Pron SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Gen
80 2 Alman alman Noun Noun OBJECT CAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Gen
80 2 uyruklu uyruk Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 ByDoingSo
80 2 3 3 Num Card MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Nom
80 2 kişinin kişi Noun Noun POSSESSOR NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Dat
80 2 boğazının boğaz Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 PCDat
80 2 kesilerek  Adv Adv MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg Pnon Gen
80 2 öldürülmesi  Noun NInf CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 0 1
80 2 olayına olay Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 0 1
80 2 ilişkin ilişkin Postp Postp MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0
80 2 davanın dava Noun Noun S.MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg P3sg Nom
```

Şekil 4.6: Örnek nitelik dosyası

4.4.1.1 Kural tabanlı niteliklerin belirlenmesi

Özne ve yer etiketleri için kural tabanlı niteliklerin çıkarılmasında Bölüm 3'de anlatılan kurallar kullanılmıştır. Bu kurallar kullanılarak Şekil 4.6'da gösterildiği gibi bir algoritma oluşturulmuştur.

4.4.1.2 Diğer niteliklerin belirlenmesi

Sırasıyla biçimbilimsel çözümleyici, belirsizlik giderici ve sözdizimsel çözümleyiciye giren metnin, her bir sözcüğü için biçimbilimsel çözümünden sözcük türü ve eklerine ait nitelikler, sözdizimsel çözümlenmesinden ise cümledeki tipine ait nitelikler çıkarılabilir. Bu niteliklerin yanında yapısal nitelikler olarak adlandırdığımız,

Çizelge 4.1: Tüm Nitelikler

Nitelik Türü	**Nitelik İsmi**	**Nitelik Değerleri**
Biçimbilimsel Nitelikler	Ana Sözcük Sınıfı	Adj, Adv, Conj, Det, Dup, Interj Noun, Num, Postp Pron,Punc,Verb
	Alt Sözcük Sınıfı	Able, Acquire, ActOf, Adamantly, AfterDoingSo, Agt Almost, As AsIf, AsLongAs, Become,ByDoingSo, Card, Caus, DemonsP, Dim,Distrib, EverSince, FeelLike, FitFor, FutPart, InBetween, Inf, Inf1, Inf2, Inf3, JustLikeLy, Ness, NotState, Ord, Pass, PastPart, PCAbl, PCAcc, PCDat, PCGen, PCIns, PCNom, Percent, PersP, PresPart, Prop, Quant, QuesP, Range, Ratio,Real, Recip, ReflexP, Rel, Related, Repeat, Since, SinceDoingSo, Start, Stay, Time, When, While, With, Without, Zero
	Ekler ve Diğer Sözcük Sınıfları	A1pl, A1sg, A2pl, A2sg, A3pl, A3sg P1pl, P1sg, P2pl, P2sg, P3pl, P3sg, Pnon, Abl, Acc, Dat, Equ, Gen, Ins, Loc, Nom, Aor, Desr, Fut, Imp, Neces, Opt, Pres, Prog1, Prog2, Cop Cond, Past, Narr CompCond, CompNarr, CompPast Cop Neg, Pos
Sözdizimsel Nitelikler		MODIFIER, OBJECT, SUBJECT, DATIVE.ADJUNCT, LOCATIVE.ADJUNCT, DETERMINER, SENTENCE, DERIV, CLASSIFIER, INTENSIFIER, POSSESSOR,COORDINATION
Yapısal Nitelikler	Metnin Sırası	1,2,3,...,(Metin adeti)
	Cümle Sırası	1,2,3,...,(Metindeki cümle adeti)
	Sıklık	1,2,3,...,10
	İlk Gözlemlendiği Yer	1,2,3,...,10
	Büyük Harfle Yer	CAPITAL,NONCAPITAL
Kural Tabanlı Nitelikler		POSSUB,POSLOC,POSDATE, POSSENT

sözcüklerin metin içerisinde görülme sıklığı, sözcüğün metinde ilk gözlemlendiği yer, cümlenin metindeki sırası gibi nitelikler metinden çıkarılmıştır.

4.4.2 Koşullu Rastgele Alanlar Yönteminin Geliştirilmesi

Koşullu Rastgele Alanlar yönteminin eğitim ve sınama aşamaları için MALLET aracı kullanılmıştır. MALLET, istatiksel DDİ uygulamaları için hazırlanan, makine öğrenmesi yöntemlerini içeren Java tabanlı bir pakettir. MALLET içerisinde CRF ile ilgili bir kısım da bulunmaktadır. CRF'in gerçekleyebilmek için eğitim sırasında L-BFG, sınama sırasında da Viterbi dinamik algoritmalarını kullanmaktadır.

Tarafımızdan etiketlenmiş belgerin bulunduğu eğitim kümesi uygun biçime getirildikten sonra CRF'i eğitmek için kullanılmıştır. Şekil 4.8'de görüldüğü gibi her bir satır belgedeki bir sözcüğü temsil eder. Her bir satırda sözcüğün belge ve cümle içerisindeki nitelikleri verilir. Her bir satırdaki en son parametre ise çıkarılmak istenilen etikettir.

CRF'in eğitilmesi sonucu bir model oluşmaktadır. Model oluşurken her bir nitelik fonksiyonunun ağırlık değeri ve etiketlerin ardarda gelme olasılıkları gibi parametreler hesaplanmakta ve model dosyasına kaydedilmektedir. Bu yapı daha sonra etiketlenmemiş bir metnin etiketlenmesi için kullanılacaktır. Sınama aşmasında,eğitilmiş CRF modeline bu kez etiketlenmemiş bir metin girilir. Bu metin de her bir satır bir sözcüğü temsil eder. Şekil 4.9'da örnek bir sınama belgesi gösterilmiştir.

Eğitilmiş CRF modeline etiketlenmemiş metin girildiğinde çıkış olarak Şekil 4.10'da örnek verildiği gibi etiketlenmiş bir çıkış üretmektedir.

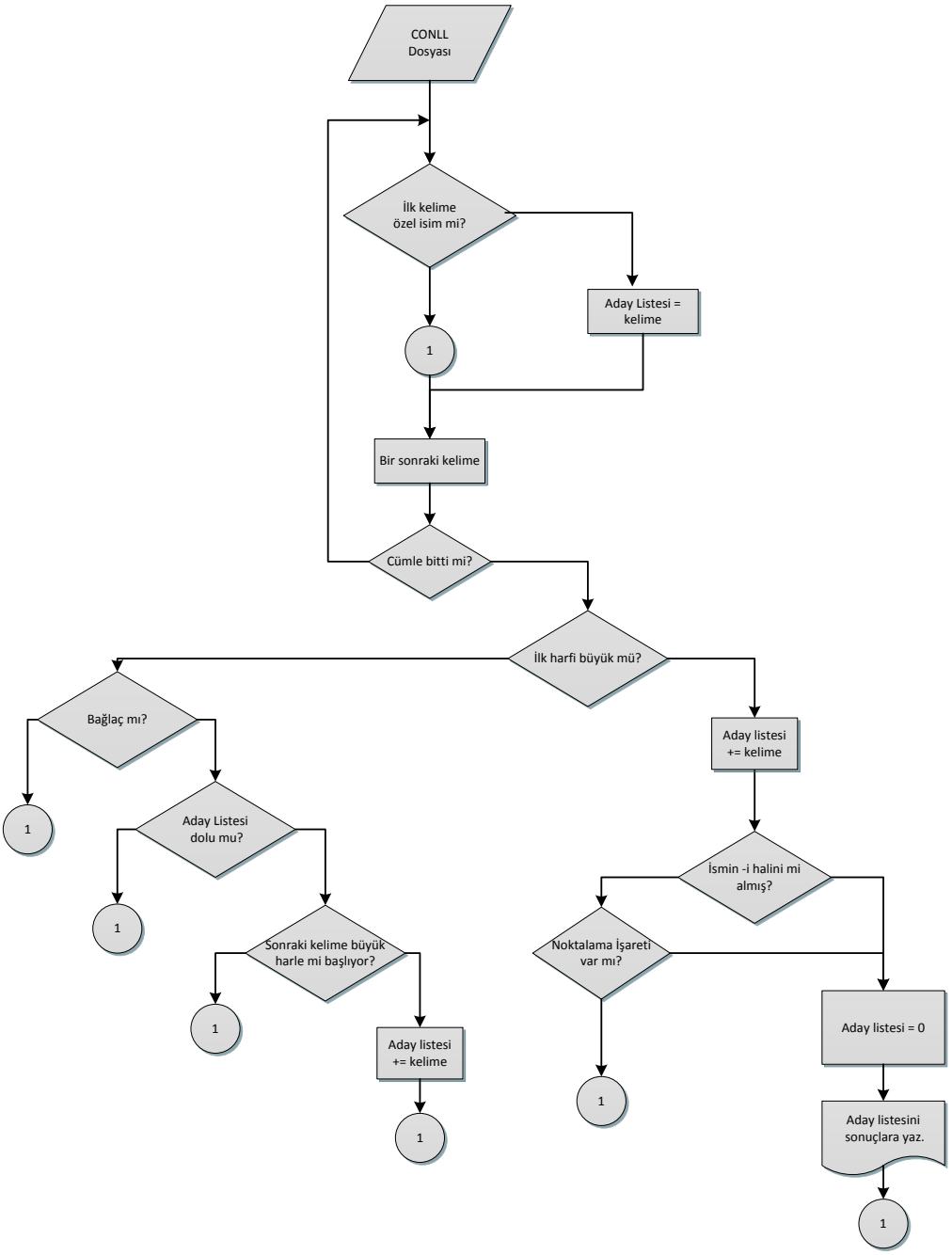
4.5 Başarımın Ölçülmesi

Çalışmamızın başarımını ölçmek için şöyle bir yol izlenmiştir:

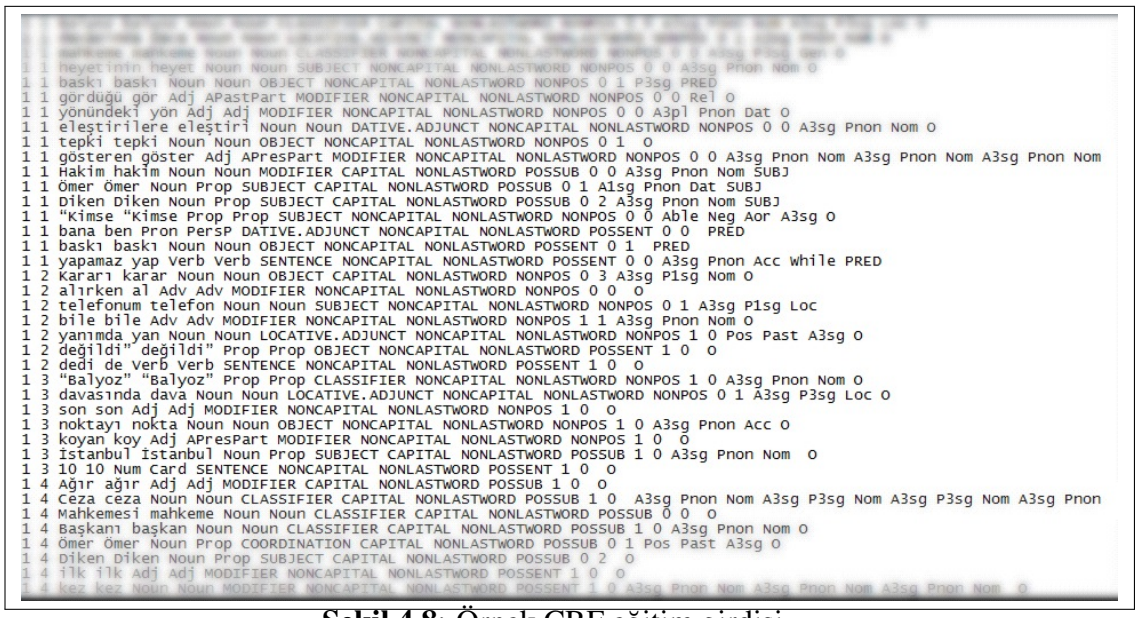
1. Tarafımızdan etiketlenmiş belgeler eğitilmiş CRF modülüne girilir.
2. Eğitilmiş CRF'in ürettiği etiketler her bir belge için elde edilir.
3. Bu etiketler tarafımızdan yazılmış etiketlerle karşılaştırılır.

4. Bunun sonucunda çalışmamızın başarımı bulma,tutturma ve F-ölçümü ile hesaplanır.

Tüm metinler için sırayla bu veriler Ölçme ve Değerlendirme kısmında kullanılmak üzere kaydedilmiştir.



Şekil 4.7: Özne ve yer niteliği bulma akış diyagramı

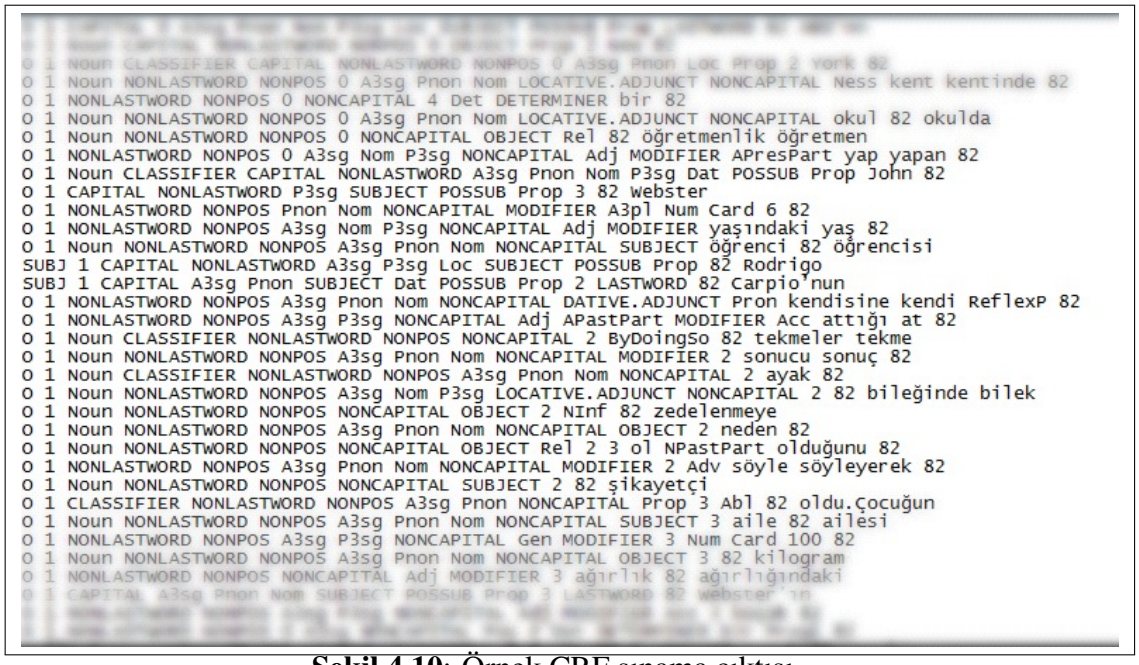


```
1 1 heyetinin heyet Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom O
1 1 baskı baskı Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 0 1 P3sg PRED
1 1 gördüğü gör Adj APastPart MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 Rel O
1 1 yönündeki yön Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3pl Pnon Dat O
1 1 eleştirilere eleştiri Noun Noun DATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom O
1 1 tepki tepki Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 0 1  O
1 1 gösteren göster Adj APresPart MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0 A3sg Pnon Nom A3sg Pnon Nom A3sg Pnon Nom
1 1 Hakim hakim Noun Noun MODIFIER CAPITAL NONLASTWORD POSSUB 0 0 A3sg Pnon Nom SUBJ
1 1 Ömer Ömer Noun Prop SUBJECT CAPITAL NONLASTWORD POSSUB 0 1 A1sg Pnon Dat SUBJ
1 1 Diken Diken Noun Prop SUBJECT CAPITAL NONLASTWORD POSSUB 0 2 A3sg Pnon Nom SUBJ
1 1 "Kimse "Kimse Prop Prop SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 0 Able Neg Aor A3sg O
1 1 bana ben Pron PersP DATIVE.ADJUNCT NONCAPITAL NONLASTWORD POSSENT 0 0  PRED
1 1 baskı baskı Noun Noun OBJECT NONCAPITAL NONLASTWORD POSSENT 0 1  PRED
1 1 yapamaz yap Verb Verb SENTENCE NONCAPITAL NONLASTWORD POSSENT 0 0 A3sg Pnon Acc While PRED
1 2 Kararı karar Noun Noun OBJECT CAPITAL NONLASTWORD NONPOS 0 3 A3sg P1sg Nom O
1 2 alırken al Adv Adv MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 0  O
1 2 telefonum telefon Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg P1sg Loc
1 2 bile bile Adv Adv MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 1 A3sg Pnon Nom O
1 2 yanımda yan Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 1 0 Pos Past A3sg O
1 2 değildi" değildi" Prop Prop OBJECT NONCAPITAL NONLASTWORD POSSENT 1 0  O
1 2 dedi de Verb Verb SENTENCE NONCAPITAL NONLASTWORD POSSENT 1 0  O
1 3 "Balyoz" "Balyoz" Prop Prop CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 1 0 A3sg Pnon Nom O
1 3 davasında dava Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg P3sg Loc O
1 3 son son Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 0  O
1 3 noktayı nokta Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 1 0 A3sg Pnon Acc O
1 3 koyan koy Adj APresPart MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 0  O
1 3 İstanbul İstanbul Noun Prop SUBJECT CAPITAL NONLASTWORD POSSUB 1 0 A3sg Pnon Nom  O
1 3 10 10 Num Card SENTENCE NONCAPITAL NONLASTWORD POSSENT 1 0  O
1 4 Ağır ağır Adj Adj MODIFIER CAPITAL NONLASTWORD POSSUB 1 0  O
1 4 Ceza ceza Noun Noun CLASSIFIER CAPITAL NONLASTWORD POSSUB 1 0  A3sg Pnon Nom A3sg P3sg Nom A3sg P3sg Nom A3sg Pnon
1 4 Mahkemesi mahkeme Noun Noun CLASSIFIER CAPITAL NONLASTWORD POSSUB 0 0  O
1 4 Başkanı başkan Noun Noun CLASSIFIER CAPITAL NONLASTWORD POSSUB 1 0 A3sg Pnon Nom O
1 4 Ömer Ömer Noun Prop COORDINATION CAPITAL NONLASTWORD POSSUB 0 1 Pos Past A3sg O
1 4 Diken Diken Noun Prop SUBJECT CAPITAL NONLASTWORD POSSUB 0 2  O
1 4 ilk ilk Adj Adj MODIFIER NONCAPITAL NONLASTWORD POSSENT 1 0  O
1 4 kez kez Noun Noun MODIFIER NONCAPITAL NONLASTWORD POSSENT 1 0 A3sg Pnon Nom A3sg Pnon Nom A3sg Pnon Nom  O
```

Şekil 4.8: Örnek CRF eğitim girdisi

```
82 1 york York Noun Prop CLASSIFIER CAPITAL NONLASTWORD NONPOS 0 2 A3sg Pnon Loc
82 1 kentinde kent Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 0 1 Ness A3sg Pnon Nom
82 1 bir bir Det Det DETERMINER NONCAPITAL NONLASTWORD NONPOS 0 4
82 1 okulda okul Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg Pnon Nom A3sg Pnon Nom
82 1 öğretmenlik öğretmen Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 0 1 Rel
82 1 yapan yap Adj APresPart MODIFIER NONCAPITAL NONLASTWORD NONPOS 0 1 A3sg P3sg Nom
82 1 John John Noun Prop CLASSIFIER CAPITAL NONLASTWORD POSSUB 1 1 A3sg Pnon Nom A3sg Pnon Nom A3sg P3sg
82 1 Webster Webster Prop Prop SUBJECT CAPITAL NONLASTWORD POSSUB 1 3 P3sg
82 1 6 6 Num Card MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 1 A3pl Pnon Nom
82 1 yaşındaki yaş Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 1 A3sg P3sg Nom
82 1 öğrencisi öğrenci Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 1 1 A3sg Pnon Nom
82 1 Rodrigo Rodrigo Prop Prop SUBJECT CAPITAL NONLASTWORD POSSUB 1 1 A3sg P3sg Loc
82 1 Carpio'nun Carpio'nun Prop Prop SUBJECT CAPITAL LASTWORD POSSUB 1 2 A3sg Pnon Dat
82 1 kendisine kendi Pron ReflexP DATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 1 1 A3sg Pnon Nom
82 1 attığı at Adj APastPart MODIFIER NONCAPITAL NONLASTWORD NONPOS 1 1 A3sg P3sg Acc
82 1 tekmeler tekme Noun Noun CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 2 1 ByDoingSo
82 1 sonucu sonuç Noun Noun MODIFIER NONCAPITAL NONLASTWORD NONPOS 2 1 A3sg Pnon Nom
82 1 ayak ayak Noun Noun CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 2 2 A3sg Pnon Nom
82 1 bileğinde bilek Noun Noun LOCATIVE.ADJUNCT NONCAPITAL NONLASTWORD NONPOS 2 2 A3sg P3sg Nom
82 1 zedelenmeye  Noun NInf OBJECT NONCAPITAL NONLASTWORD NONPOS 2 2
82 1 neden neden Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 2 1 A3sg Pnon Nom
82 1 olduğunu ol Noun NPastPart OBJECT NONCAPITAL NONLASTWORD NONPOS 2 3 Rel
82 1 söyleyerek söyle Adv Adv MODIFIER NONCAPITAL NONLASTWORD NONPOS 2 2 A3sg Pnon Nom
82 1 şikayetçi şikayetçi Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 2 1
82 1 oldu.Çocuğun oldu.Çocuğun Prop Prop CLASSIFIER NONCAPITAL NONLASTWORD NONPOS 3 1 A3sg Pnon Abl
82 1 ailesi aile Noun Noun SUBJECT NONCAPITAL NONLASTWORD NONPOS 3 1 A3sg Pnon Nom
82 1 100 100 Num Card MODIFIER NONCAPITAL NONLASTWORD NONPOS 3 1 A3sg P3sg Gen
82 1 kilogram kilogram Noun Noun OBJECT NONCAPITAL NONLASTWORD NONPOS 3 1 A3sg Pnon Nom
82 1 ağırlığındaki ağırlık Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 3 1
82 1 webster'ın Webster'ın Prop Prop SUBJECT CAPITAL LASTWORD POSSUB 3 1 A3sg Pnon Nom
82 1 küçük küçük Adj Adj MODIFIER NONCAPITAL NONLASTWORD NONPOS 3 1 A3sg P3sg Acc
82 1 bir bir Det Det DETERMINER NONCAPITAL NONLASTWORD NONPOS 0 4 Pos Prog1 A3sg
```

Şekil 4.9: Örnek CRF sınama girdisi



Şekil 4.10: Örnek CRF sınama çıktısı

Çizelge 4.2: Karşılaştırma dosyası örnek satırı

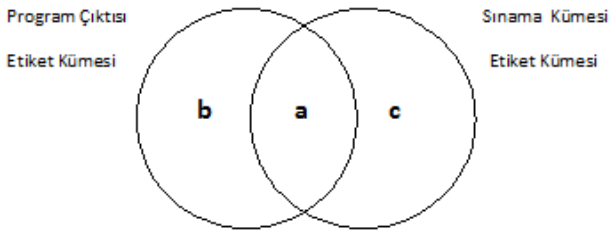
ÖZNE					
Program Tarafından Bulunan Etiketler	Elle Atanan Etiketler	Çakışan Etiketler	Program Tarafından Bulunan Etiket Sayısı	Elle Atanan Etiket Sayısı	Ortak Etiket Sayısı
[Muhammed El Baradei]	[Muhammed El Baradei] ,Atom Enerjisi Kurumu]	[Muhammed El Baradei]	1	2	1
YÜKLEM					
Program Tarafından Bulunan Etiketler	Elle Atanan Etiketler	Çakışan Etiketler	Program Tarafından Bulunan Etiket Sayısı	Elle Atanan Etiket Sayısı	Ortak Etiket Sayısı
[saldırıyı üstlendi]	[gitti]	[-]	1	1	0
YER					
Program Tarafından Bulunan Etiketler	Elle Atanan Etiketler	Çakışan Etiketler	Program Tarafından Bulunan Etiket Sayısı	Elle Atanan Etiket Sayısı	Ortak Etiket Sayısı
[Mısır]	[Mısır]	[Mısır]	1	1	1
ZAMAN					
Program Tarafından Bulunan Etiketler	Elle Atanan Etiketler	Çakışan Etiketler	Program Tarafından Bulunan Etiket Sayısı	Elle Atanan Etiket Sayısı	Ortak Etiket Sayısı
[_]	[_]	[_]	0	0	0

5. DEĞERLENDİRME

Çalışmamızın değerlendirmesi amacıyla çeşitli sınama yöntemleri kullanılmıştır. Sınama kümesi olarak elle etiketlenmiş haber metinleri kullanılmıştır. Tez çalışmasının ürünü olarak geliştirilen yazılım tarafından atanan etiketler ile sınama kümesindeki etiketler karşılaştırılarak sistemin başarımı oranı ölçülmüştür.

Bunun için bulma ve tutturma oranları ile F-ölçümleri sayesinde başarı oranı ölçülmüştür.

Bulma ve tuturma oranlarını küme diyagramları şekilde Şekil 5.1'de gösterilmiştir.



Şekil 5.1: Bulma ve tutturma kümesi

Buna göre bulma ve tuturma oranları aşağıdaki formüllerle hesaplanabilir.

$$Tutturma = \frac{a}{(a+b)} \tag{5.1}$$

$$Bulma = \frac{a}{(a+c)} \tag{5.2}$$

Yukarıdaki bağıntıda verilen bulma ve tutturma oranları çalışmamızda kullanılmıştır. Bu bağıntılarda sınama kümesiyle kastedilen taramızdan elle işaretlenmiş etiketlerdir. Bu etiketler her bir metin için ayrı olarak değerlendirilir. Her bir metinde elle etiketlenmiş sınama kümesi ile, yazılımın o metin için bulduğu etiketler karşılaştırılır.

Bulma: Sınama kümesi içinde var olan etiketlerden bulabildiklerimizin oranıdır.

$$Bulma = \frac{\text{sınama kümesinde bulunan etiketler içersinde programın bulduğu etiketler}}{\text{sınama kümesindeki etiketler}} \tag{5.3}$$

Tutturma Program ile bulabildiğimiz etiketlerin elle etiketlenmiş etiket listesinde olmasının oranıdır.

$$Tutturma = \frac{\text{sınama kümesinde bulunan etiketler içersinde programın bulduğu etiketler}}{\text{programın bulduğu etiketler}} \tag{5.4}$$

Sonuçların karşılaştırılmasında F-değerleri kullanılmıştır. Bulma ve tutturma oranları hesaplanan bir dizinin başarımını yorumlayabilmek için, bulma ve tutturma oranlarının harmonik ortalaması alınarak bir değer elde edilir.

Bu değer F-değeri olarak bilinir, karşılaştırmalar için kullanılır. Bulma ve tuturma değerlerinin F-değerinin hesaplanmasını sağlayan formül aşağıda verilmiştir.

$$F = \frac{2 * tutturma * bulma}{tutturma + bulma} \tag{5.5}$$

Bu bilgiler ışığında çakarılan sonuçlar Çizelge 5.1'de verilmiştir.

Çizelge 5.1: Herbir etiketin başarı oranları

Elle Etiketlenen Etiket Sayısı	**Programın Bulduğu etiket Sayısı**	**Çakışan Etiket Sayısı**
50	62	40
50	64	39
12	11	8
11	7	5
Bulma	**Tuturma**	**F Ölçümü**
0.8	0.645	0.72
0.78	0.61	0.68.4
0.667	0.727	0.695
0.45	0.71	0.55

Burada belirtilecebilecek diğer bir önemli husus ise, karşılatıran etiketlerin aynı varlığı işaret edip etmediğidir. Eğer etiketlerden herhangi biri diğerinin alt kümesi ise bulunan etiket doğru bulunmuş sayılır. Örneğin elle etiketlenmiş sonuca göre"Başbakan

İnönü", program çıktısına göre "Başbakan İsmet İnönü" geçiyorsa bu etiket doğru bulunmuş sayılır.

6. SONUÇ VE ÖNERİLER

Bu çalışmada metni en iyi şekilde temsil eden özne, yüklem, yer ve zaman etiketlerinin metinden çıkarımı hedeflenmiştir. Bu hedef doğrultusunda öncelikle kural tabanlı yöntemler üzerinde çalışılsa da, başarım oranındaki düşüklükten dolayı, daha sonra makine öğrenmesi yöntemlerinden CRF üzerinde çalışılmıştır. Kural tabanlı çalışma sırasında gözlemlediğimiz bazı kuralları nitelik olarak kullanarak CRF yönteminin başarımı arttırılmıştır.

Bu çalışmada sadece haber metinleri üzerinde çalışıldığı için, haber metinlerine özel bazı özellikler kullanılmıştır. İlerdeki çalışmalarda başka tür metinleri de incelemek hedeflenmiştir. Ancak bu çalışmada incelenen haber kümesi bir çok avantajın yanında dezavantaj da barındırmaktadır. Bir dezavantaj olarak, haber metinleri imla ve noktalama kurallarına uymayan bir çok hatayı da içinde barındırır. İnternet ortamında ne yazık ki haber metinleri editörler tarafından yeterince incelenmeden yayımlanmaktadır. İmla ve noktalama kuralları özellikle sözdizimsel çözümleme olmak üzere çalışmanın bütününde olumsuz etki yaratmaktadır. Başarı oranlarındaki düşüklüğün en büyük nedeni olarak bu hataları varsayabiliriz. Bu nedenle ilerideki çalışmalarımızda yazım ve noktalama hatalarını düzelten ek bir modülün geliştirilmesi hedeflenmektedir. Bu modülün sistemin başarı oranını büyük düzeyde etkileyeceği düşünülmektedir. Bununla birlikte biçimbilimsel çözümleme, belirsizlik giderici ve sözdizimsel çözümleme işlemlerini kapsayan modülün hata oranları da sistemin başarı oranını doğrudan etkiler. Bu modüllere yönelik geliştirmelerde sistemin başarısını olumlu yönde etkileyecektir.

Başarım oranını etkileyen önemli bir neden ise eğitim kümemizin niceliğinin sınırlı olmasıdır. Eğitim kümesi arttırılarak başarım oranında artış olacağı öngörülmektedir.

İlerdeki çalışmalarımızda başarım oranımızı arttırmayı ve elde edilen çözümü, anlamsal web uygulaması olarak kullanmayı hedeflemekteyiz.

KAYNAKLAR

[1] **Oflazer, K.**, 1993. Two-level description of Turkish morphology, Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics, EACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA, s.472–472, http://dx.doi.org/10.3115/976744.976810.

[2] **Sak, H.**, **Güngör, T. ve Saraçlar, M.**, 2008. Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus, Proceedings of the 6th international conference on Advances in Natural Language Processing, GoTAL '08, Springer-Verlag, Berlin, Heidelberg, s.417–427, http://dx.doi.org/10.1007/978-3-540-85287-2_40.

[3] **Eryiğit, G.**, 2007. ITU Treebank Annotation Tool, Proceedings of the ACL workshop on Linguistic Annotation (LAW 2007), Prague.

[4] **Cucerzan, S. ve Yarowsky, D.**, 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence, s.90–99.

[5] **Soderland, S.**, **Fisher, D.**, **Aseltine, J. ve Lehnert, W.**, 1995, CRYSTAL: Inducing a Conceptual Dictionary.

[6] **Bikel, D.M.**, **Miller, S.**, **Schwartz, R. ve Weischedel, R.**, 1997. Nymble: a high-performance learning name-finder, Proceedings of the fifth conference on Applied natural language processing, ANLC '97, Association for Computational Linguistics, Stroudsburg, PA, USA, s.194–201, http://dx.doi.org/10.3115/974557.974586.

[7] 1997. NetOwl Server, Proceedings of the fifth conference on Applied natural language processing, ANLC '97, Association for Computational Linguistics, Stroudsburg, PA, USA, s.15–16, http://dx.doi.org/10.3115/974281.974290.

[8] **Kucuk, D. ve Yazici, A.**, 2009. Named Entity Recognition Experiments on Turkish Texts, Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS '09, Springer-Verlag, Berlin, Heidelberg, s.524–535, http://dx.doi.org/10.1007/978-3-642-04957-6_45.

[9] **Özkan Bayraktar**, 1991. Local Grammar, Person Name Recognition in Turkish Financial Texts by Using Local Grammar Approach, METU, s.19–27.

[10] **Tür, G.**, **Hakkani-tür, D. ve Oflazer, K.**, 2003. A statistical information extraction system for Turkish, *Nat. Lang. Eng.*, **9(2)**, 181–210, http://dx.doi.org/10.1017/S135132490200284X.

[11] **Nallapati, R.**, **Allan, J. ve Mahadevan, S.**, Extraction of Key Words from News Stories.

[12] **Manning, C.D. ve Schütze, H.**, 1999. Foundations of statistical natural language processing, MIT Press, Cambridge, MA, USA.

[13] **Oflazer, K.**, **çetinoğlu, O. ve Say, B.**, 2004. Integrating morphology with multi-word expression processing in Turkish, Proceedings of the Workshop on Multiword Expressions: Integrating Processing, MWE '04, Association for Computational Linguistics, Stroudsburg, PA, USA, s.64–71, http://dl.acm.org/citation.cfm?id=1613186.1613195.

[14] **Metin, S.K. ve Karaoğlan, B.**, 2010. Collocation extraction in Turkish texts using statistical methods, Proceedings of the 7th international conference on Advances in natural language processing, IceTAL'10, Springer-Verlag, Berlin, Heidelberg, s.238–249, http://dl.acm.org/citation.cfm?id=1884371.1884400.

[15] **Bouma, G.**, 2010. Collocation extraction beyond the independence assumption, Proceedings of the ACL 2010 Conference Short Papers, ACLShort '10, Association for Computational Linguistics, Stroudsburg, PA, USA, s.109–114, http://dl.acm.org/citation.cfm?id=1858842.1858862.

[16] **Cohen, J.D.**, 1995. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting, *Journal of the American Society for Information Science*, **46(3)**, 162–174.

[17] **Matsuo, Y. ve Ishizuka, M.**, 2004. Keyword extraction from a single document using word co-occurrence statistical information, *International Journal on Artificial Intelligence Tools*, **13(1)**, 157–169, http://dx.doi.org/10.1142/S0218213004001466.

[18] **van der Plas, L.**, **Pallotta, V.**, **Rajman, M. ve Ghorbel, H.**, 2004. Automatic Keyword Extraction from Spoken Text. A Comparison of two Lexical Resources: the EDR and WordNet, *CoRR*, **cs.CL/0410062**, http://arxiv.org/abs/cs.CL/0410062.

[19] **Hulth, A.**, 2003. Improved automatic keyword extraction given more linguistic knowledge, Proceedings of the 2003 conference on Empirical methods in natural language processing, EMNLP '03, Association for Computational Linguistics, Stroudsburg, PA, USA, s.216–223, http://dx.doi.org/10.3115/1119355.1119383.

[20] **Witten, I.H.**, **Paynter, G.W.**, **Frank, E.**, **Gutwin, C. ve Nevill-Manning, C.G.**, 1999. KEA: Practical Automatic Keyphrase Extraction, *CoRR*, **cs.DL/9902007**, http://arxiv.org/abs/cs.DL/9902007.

[21] **Pala, N. ve Çiçekli, I.**, 2007. Turkish Keyphrase Extraction Using KEA, in, Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS 2007.

[22] **Wang, J.**, **Peng, H. ve Hu, J.s.**, 2006. Automatic keyphrases extraction from document using neural network, Proceedings of the 4th international conference on Advances in Machine Learning and Cybernetics, ICMLC'05, Springer-Verlag, Berlin, Heidelberg, s.633–641, http://dx.doi.org/10.1007/11739685_66.

[23] **Quinlan, J.R.**, 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA.

[24] **J.R., Q.**, 1993. C4.5:Programs for Machine Learning, California:Morgan Kaufmann.

[25] **Cicekli, I. ve Kalaycilar, F.**, 2008. TurKeyX: Turkish Keyphrase Extractor, Proceedings of the 23rd International Symposium on Computer and Information Sciences, TeX Users Group, s.84–89.

[26] **Ergin, M.**, 1976. Türk Dili, Milli Eğitim Bakanlığı, İstanbul.

[27] **Kurumu, T.D.**, 2012, Büyük Harflerin Kullanıldığı Yerler, http://www.tdk.gov.tr/index.php?option=com_content&view=article&id=183:Buyuk-Harflerin-Kullanildigi-Yerler-&catid=50:yazm-kurallar&Itemid=132.

[28] **Bilgi, T.**, 2012, Özel İsimler, http://www.turkcebilgi.org/edebiyat/dil-bilgisi/cins-isimler-cins-isimleri-233815.html.

[29] **McCallum, A.**, **Freitag, D. ve Pereira, F.C.N.**, 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation, Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, s.591–598, http://dl.acm.org/citation.cfm?id=645529.658277.

[30] **Lafferty, J.**, **McCallum, A. ve Pereira, F.**, 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, s.282–289.

[31] **Wallach, H.M.**, 2004. Conditional random fields: An introduction, Teknik Rapor.

[32] **Eryiğit, G.**, 2012, Turkish NLP Pipeline, http://web.itu.edu.tr/gulsenc/pipeline.html.

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

ÖZGEÇMİŞ

Ad Soyad: Seda Kazkılınç

Doğum Yeri ve Tarihi: Kırşehir 1985

Adres: Kozyatağı Mah. Kaya Sultan Sokak N:27/23 KADIKÖY/İSTANBUL

E-Posta: kazkilinc@itu.edu.tr

Lisans: Yıldız Teknik Üniversitesi - Elektronik ve Haberleşme Mühendisliği

Yayın ve Patent Listesi: