

T.C.
TRAKYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Türkçe Derlemlerdeki Artgönderimlerin
Tümdengelimli ve Tümevarımlı
Yöntemlerle Çözümlemesi

Savaş YILDIRIM

Doktora Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Yılmaz KILIÇASLAN

T.C.
TRAKYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Türkçe Derlemlerdeki Artgönderimlerin Tümdengelimli ve Tümevarımlı
Yöntemlerle Çözümlemesi

Savaş YILDIRIM

Doktora Tezi
Bilgisayar Mühendisliği Anabilim Dalı

Bu tez ... / ... / 2008 tarihinde aşağıdaki jüri tarafından kabul edilmiştir.

Jüri

Danışman: Yrd. Doç. Dr. Yılmaz KILIÇASLAN

Yrd. Doç. Dr. Erdem UÇAR

Yrd. Doç. Dr. Rafet AKDENİZ

Yrd. Doç. Dr. Yılmaz ÇAN

Yrd. Doç. Dr. Özlem UÇAR

Özet

Bu tez çalışmasında Türkçe Artgönderim Problemi çeşitli yaklaşımlar ışığında ele alınmıştır. Bilgi tabanlı tümdengelimli yöntemlerle, öğrenme tabanlı tümevarımlı yöntemler birbirleriyle çeşitli açılardan karşılaştırılmıştır. Belirginlik Tabanlı yaklaşım ile Merkezleme Algoritması, bilgi tabanlı yöntemleri değerlendirmek için ele alınmıştır. Makine Öğrenme algoritmalarından karar ağaçları, naïve Bayes algoritması, k-en yakın komşu algoritması, Destek Vektör makineleri ve algılayıcılar artgönderim problemine uygulanmıştır. Bazı Türkçe metinlerdeki artgönderimsel ilişkiler bilgisayarlı bu modellerle az bilgi kullanılarak çözümlenmiş ve deney sonuçlarına dayanarak birçok bulgu sunulmuştur. Bu alanda Türkçe için yapılan çalışmaların azlığı sebebiyle araştırmacılar tarafından kullanılan ortak bir derlem bulunamamıştır. Bu nedenle 10000 kelime kapasiteli 1114 adıl içeren bir derlem yaratılmıştır. Çeşitli deneylerle bu derlem üzerindeki gönderimsel ilişkilerin davranışı analiz edilmiştir. Deney sonuçlarını değerlendirebilmek için derlemdeki gönderimsel ilişkiler elle işaretlenmiştir. Kapsam olarak adısal artgönderim, dönüşlü artgönderim, boş artgönderim ve karşılıklı artgönderim tipleri ele alınmıştır. Bu tezin temel amacı Türkçe **Artgönderim Çözümlemesi** için az bilgili ve bilgisayarlı yaklaşımlar sunmak, bu yaklaşımları birbirleriyle kıyaslamak ve bu problemin bilgisayarlı çözümlenmesi için bir tartışma başlatmaktır.

Anahtar Kelimeler: Artgönderim Çözümlemesi, Bilgisayarlı Dilbilim, Makine Öğrenmesi

Abstract

In this study, the problem of anaphora resolution in Turkish is addressed in the light of various approaches. Knowledge-based deductive methods and learning-based inductive methods are compared from various points of view. The salience-based approach and the centering algorithm are chosen as examples of knowledge-based methods to evaluate. Of machines learning algorithms, decision trees, the naïve Bayes algorithm, the k-nearest neighbor classifier, support vector machines, and the voted perceptron are chosen and applied to the problem of anaphora resolution. Anaphoric relations in some Turkish texts are resolved by means of these computational techniques in a knowledge-poor framework and numerous findings are presented as results of the experiments. Since there is no publicly released corpus on which we can conduct our experiment, we needed to create a Turkish corpus- with the capacity of 10000 tokens and of 1114 pronouns- in which anaphoric relations are manually annotated. Pronominal anaphora, reflexive anaphora, null pronoun, and reciprocals are in the scope of this study. So, main goal of the study is to provide a computational resolution for the turkish language, to compare deductive and inductive approaches, and to start a discussion about the problem.

Keywords: Anaphora Resolution, Computational Linguistics, Machine Learning

TEŞEKKÜR

Bu tez çalışması süresince her türlü katkıyı sağlayan ve sonsuz enerjisiyle beni destekleyen danışmanım Yrd. Doç. Dr. Yılmaz Kılıçaslan'a, tez çalışmasında beraber çalıştığım Edip Serdar Güner'e, Bilgi Üniversitesinde beraber çalıştığım Tuğba Yıldız'a, Trakya Üniversitesi Bilgisayar Mühendisliği Bölümüne ve bu tezi tamamlamam için bana gereken fırsatı ve motivasyonu veren Bilgi Üniversitesi ve Bilgi Üniversitesi Bilgisayar Bilimleri Bölümüne sonsuz teşekkürlerimi sunarım.

İçindekiler

1	GİRİŞ	1
1.1	Problem	1
1.2	Artgönderim ve Önemi	2
1.3	Artgönderim Çalışmaları	4
1.4	Kapsam	6
2	DİLBİLİMSEL ARKAPLAN	8
2.1	Artgönderim Biçimleri	10
2.1.1	Adılsal Artgönderim	10
2.1.2	Dönüştürücü Artgönderim	11
2.1.3	Boş Artgönderim	11
2.1.4	Sözlüksel Ad Öbeği Artgönderimleri	12
2.1.5	Karşılıklı Artgönderim (Reciprocal)	13
2.1.6	Boş Adsal Artgönderim	13
2.2	Yan Tümcelerde Artgönderim	14
2.3	Diğer Dilbilimsel Artgönderim Çalışmaları	15
2.3.1	Bağlama Kuramı (Binding Theory)	15
2.3.2	Givon	18
2.3.3	Fox	18
2.3.4	Merkezleme Teorisi	19

2.4	Türkçe için Dilbilimsel Artgönderim Çalışmaları	19
3	BİLGİSAYARLI ARTGÖNDERİM ÇALIŞMALARI	26
3.1	Artgönderim Çözümleme Süreçleri	26
3.2	Bilgi Tabanlı (Knowledge-Based) Artgönderim Çözümleme Çalışmaları	29
3.2.1	Hobbs'un Algoritması(Hobbs's Naive Algorithm)	30
3.2.2	Merkezleme Teorisi ve Algoritması (Centering Teory)	34
3.2.3	BFP	35
3.2.4	Lappin ve Leass'in Sözdizim Tabanlı Yaklaşımı	37
3.2.5	Kennedy ve Boguraev'in Yöntemi	38
3.2.6	Baldwin's Cogniac	39
3.2.7	MARS	40
3.3	Öğrenme Tabanlı Artgönderim Çözümleme Çalışmaları	42
3.3.1	Artgönderim İçin Öğrenme Süreçleri	42
3.3.2	Diğer Dillerdeki Öğrenme Tabanlı Çalışmalar	44
3.4	Türkçe için Bilgisayarlı Artgönderim Çözümleme Çalışmaları	46
4	TÜRKÇE İÇİN ARTGÖNDERİM ÇÖZÜM DENEMELERİ	48
4.1	Derlem	49
4.2	Gönderimsel İfadelerin Tespiti	51
4.3	Bilgi Tabanlı Çözüm Önerileri	53
4.3.1	Merkezleme Tabanlı Çözüm Önerisi	54
4.3.2	Belirginlik Tabanlı Çözüm Önerisi	62
4.4	Öğrenme Tabanlı Çözümler	70
4.4.1	Eşleşme	70
4.4.2	Kullanılan Veri Madenciliği Teknikleri	71
4.4.3	Sonuçlar	76

5	DENEY SONUÇLARININ DEĞERLENDİRİLMESİ	92
6	SONUÇ	104
6.1	Amaç	104
6.2	Yöntem	105
6.3	Sonuçlar ve Bulgular	105

Şekil Listesi

3.1	Birinci tümcenin sözdizim ağacı	31
3.2	İkinci tümcenin sözdizim ağacı	32
3.3	Ayrıştırma Ağacı	34
4.1	Derlem Girme ve Güncelleme Arayüzü	51
4.2	Merkezeleme Yaklaşımının Akış Diagramı	55
4.3	K-en yakın komşu algoritması	72
4.4	Karar Ağacı	74
4.5	Destek Vektör Makineleri - KDM	75
4.6	Yapay Sinir Ağları	76
4.7	Aşırı Uyma	85
5.1	Belirginlik Temelli Yaklaşım Gösterimi	96

Tablo Listesi

2.1	AÖ Sınıflandırma Tablosu	16
3.1	Merkezdeki Değişiklikler	36
3.2	Belirgenlik Türleri ve Ağırlıkları	38
3.3	Öğrenme Veritabanı	43
4.1	Öncüllerdeki İsim Hallerinin Oluşma Sıklığı	57
4.2	Senaryolar	60
4.3	Etkenler ve Ağırlıkları	66
4.4	Aday Öncüllerin Ağırlıklandırılmaları	67
4.5	Bağıl Önem Tablosu	68
4.6	Karar Gücü Tablosu	69
4.7	Algoritma ve Parametreleri	77
4.8	Sınıflandırıcıların Sonuçları	77
4.9	Karışıklık Matrisi	78
4.10	Özelliklerin Katkı Değeri Ortalamaları / 10'lu Çapraz Sağlama .	87
4.11	Özelliklerin Bilgi Kazancı	87
4.12	Artgönderim Başarı Oranı	89
5.1	Artgönderim Başarı Oranı	95
5.2	Tüm Yaklaşımların Genel Görünümü	99
5.3	Artgönderim Çalışmaların Başarı Oranları	103

Bölüm 1

GİRİŞ

1.1 Problem

Doğal dildeki gönderimlerin yada dildeki bağlantıların çözümlenmesi Bilgisayarlı Dilbilim çalışmalarının temel adımlarından biridir. Örneğin Türkçe’de

Arda Alin'den onun kitabını istedi.

gibi bir tümcede **onun** ifadesi hangi varlığa, Alin’e mi yoksa Arda’ya mı gönderimde bulunur? En azından Türkçe bilen birçok kişi sezgisel olarak **onun** ifadesinin Alin’e gönderimde bulunduğunu çıkarabilir. Yapısal olarak yukarıdaki tümceye benzeyen fakat farklı bir çözümü olan bir başka örnek tümce verilebilir. Örneğin

Arda Alin'den kendi kitabını istedi.

gibi bir tümcedeki kendi -ki bu kendi ifadesi "**onun**" adılsal artgönderimi yerine konmuş dönüşlü bir artgönderimdir- ifadesi Alin yerine Arda’ya gönderimde

bulunur. Bu çıkarımı yine Türkçe'yi bilen biri rahatlıkla yapabilir. Ancak ilk örnekte olduğu gibi, kişi bu çıkarımı nasıl yaptığını bize sistematik olarak ifade edemez. Belki o kişi "**nasıl yaptın?**" sorusu sorulduğu anda düşünüp çıkarımda bulunabilir. Bu tam olarak insandaki sezgisel ve farkında olmadığı bir yetenektir diyebiliriz. Peki yukarıdaki iki örnek yapısal olarak birbirlerine benzemelerine karşın neden farklı bir çözüm oluşmuştur. Bunun için akla gelen ilk cevap birinci örnekte adılısı bir artgönderim, ikinci örnekte dönüşlü bir artgönderim kullanıldığı olacaktır. İki farklı artgönderim kullanıldığı için çözümleme sistematığı de farklı kurallar içermiştir diyebiliriz. Bu örneklemeden hareketle şöyle bir noktaya varabiliriz: Artgönderim problemi için birçok kuraldan oluşan bir sisteminin oluşturulması gerekmektedir. Daha basit bir ifadeyle dilde kuralları gizlenmiş bu ilişkilere bir çözüm modeli oluşturulması gerekmektedir. Belki bu çok sayıdaki kurallar daha basit bir genellemeden oluşmaktadır. İşte bu problem ekseninde bu tip soruların cevabı aranmaktadır.

Bu tez kapsamında oluşturulacak ve de tartışılacak model bilgisayarlı bir alan içerisinde olacaktır. Bilgisayarlı birçok yaklaşım çözüm için denenecektir ve çok deneylerle ortaya atılan yaklaşımlar kıyaslanacak ve detaylı bir şekilde tartışılacaktır.

1.2 Artgönderim ve Önemi

Yapay Zeka alanının alt başlıklarından olan Doğal Dil İşleme (Natural Language Processing) ve Bilgisayarlı Dilbilim (Computational Linguistics) alanları, doğal dili istatistiksel ve kural tabanlı modellerle ele alarak bilgisayarlı çözümler üreten, araştırmacıların son zamanlarda yoğunlaştığı ve önemi giderek artan disiplinler arası bilimsel çalışma alanlarıdır. Doğal dil işleme alanı dilbilim teori

ve yaklaşımlarından yapay zeka çalışmaları yapmak için yararlanır. Bilgisayarlı Dilbilim ise bilgisayar bilimlerinden dilbilim alanındaki teori ve yaklaşımların modellenmesi ve doğrulanması için yararlanır. Bu tezde sunulan çalışma her iki alana da katkı sağlama potansiyeli içermektedir.

Her iki alanın da teorik olarak beslendiği dilbilim alanı ise sözdizim (syntax), biçimbilim (morphology), sesbilim (phonology), anlambilim (semantics), edimbilim (pragmatics), söylem analizi (discourse analysis) gibi birçok önemli alt başlığı içerir.

Yukarıda bahsedilen alt alanlardan biri ve bu tezin kapsadığı alt alanlardan birisi olan söylem analizi (discourse analysis) birçok uygulama için önemini korumaktadır. Söylem analizi çerçevesinde metin özetleme, makine çevirisi, bilgi çıkarımı, diyalog sistemi gibi teorik ve pratik çalışmalar yapılmaktadır. Söylem düzeyindeki analiz dilbilim çalışmalarındaki en zor problemlerden biridir. Bu düzeyde yapılacak çalışma alt düzeylerdeki birçok bilgiye ihtiyaç duyduğundan bu alt düzeylere sıkı sıkıya bağlıdır. Sözdizimsel ya da biçimbilimsel analizlerde ortaya çıkan problemler söylem düzeyinde yapılacak başka bir analizi dolaysız bir şekilde etkiler.

Söylem analizi için çözülmesi gereken problemten biri ve aynı zamanda bu tez çalışmasının da konusu artgönderim (**anaphora**) problemidir. Artgönderim söylem içindeki öğelerin birbirlerine gönderimde bulunmasıdır. Bir başka deyişle bir dünya varlığına çeşitli şekillerde eşgönderimde bulunma durumudur. Örneğin, John isimli birine söylem içinden *John, adam, o, çocuk, Mr. John, John Green, Bay John* şeklinde gönderimler oluşturulabilir. Bu gönderimsel ilişkilerin söylem içindeki rolü çok güçlüdür. Her söylem kendi içinde çok sayıda gönderimsel ilişki barındırır. Bu gönderimsel ilişkilerin varlığı bilgisayarlı dilbilim çalışmaları için bilgisayarla çözümlenmeleri gerekliliği gibi zor problem-

ler oluşturur. Söylem içinde yapılan herhangi bir gönderimin çözülmesi söylem analizleri için çok gereklidir. Örnek olarak bir söylemde, **o** adılının Ahmet'e mi yoksa Ayşe'ye mi gönderimde bulunduğu söylem analizi için önemlidir. Bu bilgi olmadan söylem üzerinde herhangi bir çıkarım yapılamayacaktır. Bu gönderimsel olguya İngilizce'de **Anaphora** denirken Türkçe için sıklıkla kullanılan karşılığı **Artgönderim**dir. Bu gönderimsel ilişki için aşağıdaki iki tümcelik basit söylemi ele alalım:

Alin çantasındaki kitapları masaya koydu. ∅ Sonra masaya oturup onları okumaya başladı.

Örneğin yukarıdaki söylemde ikinci tümce iki artgönderim barındırır: ∅, *onlar*. Bu artgönderimlerin hangi öncüle gönderimde bulunduğu Türkçe bilen biri tarafından kolaylıkla çözülebilir. ∅ boş artgönderimi birinci tümcedeki **Alin** kişisine gönderimde bulunurken, **onlar** adılı **kitaplar** söylem varlığına gönderimde bulunur. Bu çözümlemenin insan tarafından yapılması normal ve olağandır. Ancak, bu çözümlemenin bilgisayar tarafından yapılmaya çalışılması birçok güçlüğü beraberinde getirecektir. Yani hangi adılın hangi öncüle gönderimde bulunduğunu tespit etmek için bilgisayarlı bir model yaratmak gerekmektedir. İşte bu problemin çözümü için uygulanacak çözümleme sistemine Bilgisayarlı Artgönderim Çözümlemesi (Computational Anaphora Resolution) denir.

1.3 Artgönderim Çalışmaları

Artgönderim probleminin çözümü için çeşitli yöntemler önerilmiştir. [Hobbs, 1978] bu problemi çözmek için tümcelerin ayrıştırma ağacından faydalanmıştır. Ayrıştırma ağaçlarını belli bir sisteme göre tarayarak adılların öncüllerine ulaşmaya

çalışmış ve bir sistem önermiştir. [Grosz ve Sidner, 1986] ve [Brennan vd., 1987] merkezleme teorisini geliştirmişlerdir. Bu teoriye göre söylemin merkezindeki değişiklikler belli bir davranış ilkesine göre oluşmaktadır. Söylemin, merkezindeki öğeyi mümkün oldukça koruma eğilimi içinde olduğu vurgulanmıştır. Bilgi tabanlı bu teori başarılı ve önemli olduğundan birçok çalışmaya ilham vermiştir. [Lappin ve Leass, 1994] ise belirginlik tabanlı bir yöntem geliştirmiştir. Bu yaklaşıma göre aday öncüller puanlanarak bir yarış içine sokulur. En yüksek puanı alan öncül hesaptaki adılın öncülü olur. Basit bir toplama işlemi çerçevesinde geliştirilmiş bir yaklaşım olmasına karşın çok başarılı sonuçlar vermiştir. Ayrıca günümüzdeki birçok artgönderim çözümleme sisteminde kullanılmaktadır. [Chomsky, 1981] tarafından sunulan bağlama teorisi birçok yöntem tarafından kullanılmıştır. Bağlama kuramında birçok kural ve kısıt vardır. Bu kurallar ve kısıtlar artgönderimin davranışını tanımlar. Bu konudaki detaylar ileriki bölümlerde yer alacaktır. Gerçek uygulama açısından ise en önemli çalışmalardan biri [Mitkov, 2002] tarafından yapılmıştır. Bu çalışma [Lappin ve Leass, 1994] tarafından uygulanan RAP algoritmasının geliştirilmiş bir türüdür. Gerçek uygulamalarda çalışması çalışmanın önemini artırmaktadır.

Yukarıdaki teorik yaklaşımların dışında artgönderim problemi için makine öğrenmesi ve istatistiksel yöntemler de kullanılmış ve bu yöntemler problemin çözümü açısından başarılı olmuştur. Makine Öğrenmesi tekniklerinin bilim alanındaki öneminin artması ve birçok alanda başarılı bir şekilde uygulanması, artgönderim probleminin çözümünde de başarıyla kullanılmaları doğrultusunda umut doğurmuştur. Makine öğrenme algoritmasını artgönderim çözümü için uygulayan ilk çalışma [Aone ve Bennet, 1995.] tarafından yapılmıştır. Bu çalışmada söylemler öğrenme algoritması için düzenlenmiş ve karar ağacı yöntemi kullanılarak bir karar verme sistemi oluşturulmuştur. Japonca için uygulanan bu yöntem birçok çalışmaya ön ayak olmuştur. Örnek olarak daha sonra İngilizce

için yapılan en önemli öğrenme temelli yaklaşım [Soon vd., 2001.] tarafından sunulmuştur. Bir diğer önemli çalışma ise [McCarthy ve Lehnert, 1995.] tarafından yapılmıştır. Bu çalışmaların ayrıntıları izleyen bölümlerde verilecektir.

Tüm bu çalışmalar artgönderim problemi için zengin bir literatür oluşturmuş ve bu problem çerçevesinde birçok araştırma ve bilimsel tartışma yapılmıştır. Türkçe için yapılan çalışmalar ise diğer dillerle kıyaslandığında oldukça azdır. İngilizce, Almanca, Japonca en çok çalışılan diller olurken Türkçe bu anlamda az çalışılan diller kategorisine girmiştir. Türkçe için bilgisayarlı artgönderim çözümlemesi ilk olarak

[Tın ve Akman, 1994] tarafından sunulmuştur. Bunun dışında Türkçe için bazı çalışmalardan bahsetmek mümkün. Bu çalışmalar ilerleyen bölümlerde detaylı olarak incelenecektir.

1.4 Kapsam

Bu çalışmanın amacı Türkçe söylemlerdeki artgönderim çözümü için bilgisayarlı sistemler tasarlamak, çeşitli yaklaşımları denemek ve bunları tündengelim ve tümevarım ikililiği üzerinden tartışmaktır. Bu vesileyle öncelikle çalışmanın kapsamından bahsedilmesi gerekiyor. Türkçe’de öne çıkan gönderim tipleri adılsal artgönderim, dönüşlü artgönderim, boş artgönderim, sözlüksel artgönderim, karşılıklı artönderim, boş adsal artgönderim şeklinde sıralanabilir. Bu gönderim biçimlerinden adılsal artgönderim (ben,sen, o vb.), dönüşlü artgönderim (kendim, kendi vb.), boş artgönderim (yüzeyde görünmeyen), karşılıklı artgönderim (birbirleri) bu çalışmanın kapsamındadır. Bu artgönderimlerin oluşturduğu ilişkilerin çözümü için çeşitli çözümleme yöntemleri tartışılacaktır.

Dünya’da kullanılan Artgönderim Çözümleme sistemlerini iki yaklaşıma göre

ayırabiliriz:

1. Bilgi Tabanlı Yaklaşımlar
2. Öğrenme Tabanlı Yaklaşımlar

Bu iki yaklaşımın Türkçe'deki artgönderim problemi için nasıl bir sonuç vereceği bu çalışmanın konusudur. Ancak Bilgi tabanlı tüm sistemleri denemek olanaksız olduğundan en önemli iki yöntem olan, belirginlik tabanlı yöntem¹ ile Merkezleme Algoritması yöntemi bu çalışma kapsamında tartışılacak yöntemlerdir.

Öte yandan öğrenme tabanlı yaklaşımlar yine aynı derlem üzerinde sınanacaktır. Makine öğrenme algoritmalarından Karar Ağacı, Destek Vektör Makineleri gibi önemli teknikler artgönderim çözümü için kullanılacaktır. Ortaya çıkan başarı sonuçları çerçevesinde tüm yöntemler birbirleriyle kıyaslanacak ve belirli sonuçlar çerçevesinde tartışmalar açılacaktır. Bilgi tabanlı yöntemlerle, öğrenme tabanlı yöntemler birbirleriyle çeşitli kıyaslamalar göz önünde bulundurularak tartışılacaktır. Bu yöntemlerin Türkçe söylem analizi için ne kadar başarılı ve uygulanabilir olduğu araştırılacaktır.

¹Bu yöntem [Lappin ve Leass, 1994] ve [Mitkov, 2002] tarafından kullanılan sistemlerdir.

Bölüm 2

DİLBİLİMSEL ARKAPLAN

Son otuz yıldır giderek artan söylem yapısı ile ilgili çalışmaların merkezinde hep artgönderim konusu yer almıştır. Söylem gerçek dünyadaki nesnelere gönderimde bulunurken bir bağlaşıklık düzeneği olarak artgönderimlere başvurur. Söylemde odakta bulunan anlatımsal ifadeler artgönderimsel olarak adıllaşabilir. Söylem içinde gerçek dünyadaki bir nesneye çeşitli şekillerde gönderimlerde bulunabiliriz: *Ali, bardak, adam, o, bu, onlardan biri, ora, masa vb.* Söylem içinde önce öncüller sisteme dahil olurlar. Örnek olarak tümce 2.1’de Ali, Hasan ve Kitap sisteme dahil olurlar.

Ali Hasan’a kitabı uzattı. (2.1)

Bu öğeler herhangi bir anda söylem içinde bir adıl ile eşgönderim oluşturabilirler. Bu tümcenin hemen ardından geldiğini varsayarsak, tümce 2.2 buna bir örnek oluşturur.

Ama o almak istemedi. (2.2)

Bu tmcedeki *o* aık adılı Hasan'a gnderimde bulunur. 2.1 nolu tmceden hemen sonra 2.3'deki gibi bir tmce gelseydi bu durumda *onu* aık adılı kitaba gnderimde bulunacaktı.

Ancak Hasan onu almak istemedi. (2.3)

Eęer tmce 2.4'teki gibi bir tmce gelseydi, ncelikle gzle grnr aık bir adılın olmadıęı farkedilirdi. Ancak zgn grnen birinin olduęu da kesindir.

ok zgn grnyordu. (2.4)

Bu durumda bu tmcede gremedięimiz ancak dilbilgisel olarak gizli bir znenin varlıęında bahsedebiliriz. Bu tmcede Ali'ye gizli bir artgnderim sz konusu olduęundan buna boę artgnderim diyeceęiz ve $\oslash$ ięareti ile ifade edeceęiz.

Bu rneklerden anlaşıldıęı gibi ama geriye ięaret etmek ve bunun yapısını araştırmaktır. Geriye ięaret eden veya gnderimde bulunan yapılar iin İngilizce'de **Anaphora** kelimesi kullanılmaktadır. Trke iin kullanılan ve kabul gren karşıllıęı ise **artgnderim**dir. Bu artgnderimin (Anaphora) ięaret ettięi, gnderimde bulunduęu ęeye ise İngilizce'de **antecedent** ya da **referent** denir. Trke dilbilimsel alıřmalarda oęunlukla **ncl** kelimesi kullanılır. Tmce 2.1'de Ali, Hasan ve kitap aday ncllerdir. Tmce 2.2'deki *o* ęesi ise Ali n-

cülünü işaret eden bir açık adıdır ve bir artgönderim örneğidir.

2.1 Artgönderim Biçimleri

2.1.1 Adısal Artgönderim

Ana ve yan tümcelerde özne pozisyonunda açık adıl ve boş artgönderimleri tümcenin yüklemi ile sayı/kişi uyumu içindedir. Üçüncü tekil şahıs uyum takısı boştur. Üçüncü çoğul kişide eğer açık özne kullanılmışsa fiildeki biçimsel takı bağlama bağlı olarak kullanılabilir ya da kullanılmayabilir. Aksi halde uyum gösterir ve açıktır. Artgönderim, söylem içinde çeşitli biçimlerde var olabilir. Bu biçimin zorunluluk mu yoksa konuşucunun ya da yazarın seçimine mi bağlı olduğu değişkenlik gösterir. En yaygın kullanılan artgönderim tipi adısal artgönderimdir (pronominal anaphora). **Ben, Sen, O, Biz, Siz ve Onlar** adısal artgönderim örneklerinin yalın halleridir (**Ben** yarın okulda olacağım). Bununla birlikte ismin hallerini de alırlar:

1. Yalın Hali (Nominative): *Ben, Sen, O ...*
2. Belirtme Hali (Accusative): *Beni, Seni, Onu ...*
3. Yönelme Hali (Dative): *Bana, Sana, Ona ...*
4. Ayrılma / Çıkma Durumu (Dative): *Benden, Senden, Ondan ...*
5. İlgi / Tamlayan Hali (Genitive): *Benim, Senin, Onun ...*
6. Bulunma hali (Locative): *Bende, Senden, Onda ...*

2.1.2 Dönüslü Artgönderim

Bir başka artgönderim tipi ise dönüslü (reflexive) adıllardır. "Jonh admires himself" tümcesindeki himself aynı tümcedeki John varlığına gönderimde bulunur. Türkçede bu tümceyi "John kendini beğenir" şeklinde söyleriz. Bu tümcedeki **kendi** terimi Türkçe dönüslü adılar için bir örnektir. Dönüslü artgönderim de adılsal artgönderim gibi durum hallerinden etkilenebilir: Kendi, kendini, kendisine, kendinde...

2.1.3 Boş Artgönderim

Söylemde yüzeyde görünmeyen fakat dilbilgisi kurallarınca açığa çıkarılan artgönderim tipidir. Aşağıdaki söylemi ele alalım.

Alin_i baleye gitmek için hazırlık yapıyordu.

[$\emptyset_i$ [$\emptyset_i$ Çantasına gerekli eşyaları koyarak] evden çıktı.] (2.5)

Bu tümcedeki özne pozisyonundaki ilk öge boş bir artgönderimdir ve Alin ile eşgönderime sahiptir. Benzer bir şekilde boş artgönderim pozisyonunda bulunan ikinci $\emptyset$ gönderimi de boş artgönderim'e örnektir. Bu boş artgönderimler adıl mıdır yoksa değişken midir sorusu halen tartışılmaktadır. Bu tartışma çalışmamızın dışında kaldığından biz yüzeyde görünmeyen bu ögelere boş artgönderim diyeceğiz.

2.1.4 Sözlüksel Ad Öbeği Artgönderimleri

Diğer bir artgönderim tipi ise **Sözlüksel Ad Öbeği Artgönderimleri**dir. Bu yapı sözdizimsel olarak tümcelerde belirtili isim öbekleri şeklinde bulunurlar: Özel isimler, belirtili tanımlar vb. *Noam Chomsky, bir sepet elma, masadaki vazo, adam* örnek olarak verilebilir.

Artgönderim örneklerimize aşağıdaki alıntıyı inceleyerek devam edelim.

(1) Evvel zaman içinde bir kentte küçük ve fakir bir kız yaşarmış.(2) Bu kız kibrit satarak geçinmeye çalışırmış. (3) Kızın hiçkimsesi yokmuş.(4) Yıkık dökük bir evde yaşarmış. (5) Bir sabah yine kibrit satmak için yollara düşmüş.(6) Kar yağdığı için her taraf bembeyazmış.(7) Kibritçi Kızın ayağına giyecek ayakkabısı bile yokmuş.(8) Yalın ayak karlara basa basa kibrit satmaya çalışıyormuş. (9) Üzerindeki elbiseler o kadar eskimiş ki soğuktan korumuyormuş bile. (10) Kibritçi kız soğuktan donmamak için bir köşeye büzülmüş. (11) Artık akşam olmak üzereymiş. (12) Kızcağız artık kibrit satamayacağını anlamış.

Bu metinde ilk tümceyle birlikte *kent* ile *küçük ve fakir bir kız* ad öbekleri söylem varlıkları olarak sisteme girerler. Bir sonraki tümcede (*Bu kız kibrit satarak geçinmeye çalışırmış*), *küçük ve fakir bir kız* söylem varlığına *bu kız* ad öbeğiyle gönderimde bulunulur. Aslında bu iki ad öbeği gerçek dünyadaki bir varlığa eşgönderimde bulunurlar. Bu yapıdaki iki ad öbeği de gönderimde bulunduklarından bunlar **Sözcüksel Ad Öbeği Artgönderimi**'nin birer örneğidir. Benzer bir şekilde ikinci tümceyle birlikte **kibrit** varlığı da sisteme dahil edilmiştir. Bu ad öbeğine de izleyen bir tümceden gönderim yapılabilir. Bu tip ad

öbekleri genel anlamda göderimsel ifadeler olarak değerlendirilir. Şimdi yukarıdaki alıntıda küçük kız için nasıl bir gönderim yapısı kurulduğuna bakalım. Kullanılan gönderimleri tümce sırasıyla listeleyelim: Küçük ve fakir bir kız (1), Bu kız (2), Kızın (3), $\oslash$ (4), $\oslash$ (5), Kibritçi Kızın (7), $\oslash$ (8), $\oslash + \text{GEN}$ ve $\oslash + \text{ACC}$ (9), Kibritçi kız (10), Kızcağız(12). Ortaya çıkan eşgönderimsel bu zincirde açık adıl, boş artgönderim, ve ad öbeği artgönderimleri kullanılmıştır.

2.1.5 Karşılıklı Artgönderim (Reciprocal)

Bir diğer artgönderim tipi ise karşılıklı artgönderimdir. Söylemde karşılıklı bulunan öğeler arasındaki karşılıklılığı ifade etmek amacıyla kullanılır.

They hit each other (2.6)

Yukarıdaki tümcede geçen **each other** ifadesi İngilizce'deki karşılıklı artgönderim için bir örnektir. Türkçe için kullanılan kelime ise **birbirleridir**.

Birbirlerine vurdular. (2.7)

Yukarıdaki tümcedeki birbirlerine ifadesi karşılıklı artgönderime bir örnektir.

2.1.6 Boş Adsal Artgönderim

Bir ad öbeğinin olması gereken yeri boş bıraktığı durumlarda ortaya çıkar. Buna aşağıdaki gibi bir örnek verelim.

- Ben [senin gibi bir ayyaş] yol vermem
- Ben $\emptyset$ veririm

Yukarıdaki söylemde birinci konuşucu *senin gibi bir ayyaş* ifadesi ile ikinci konuşucuyu tanımlıyor. İkinci konuşucu aynı ifadeyi boş kullanarak birinci konuşucuyu tanımlıyor. İkinci konuşucunun tümcesinde adısal boş bir artgönderim olduğunu söyleyemeyiz; bu durumda bu ifade ikinci konuşucuyu yani kendisini tanımlıyor olacaktı. Oysa ikinci konuşucunun birinci konuşucuyu belirttiği açıktır.

2.2 Yan Tümcelerde Artgönderim

Türkçede yan tümce üretmek için eyleme eklenen çekim ekleri şu şekildedir:

1. -mek: Öğrenciler *derse girmek* istemiyorlar.
2. -me: *Öğrencilerin derse geç gelmesi* en ciddi problemimiz.
3. -iş: *Hocanın dersi anlatışı* bence iyi.
4. -ecek: *Dersi geçecek öğrenciler* belli.
5. -dik: Öğrenci *aldığı notu* beğenmedi.

Yan tümcenin öznesi tamlayan eki alırken, eylemi iyelik eki alır. Ve bu iki birim kişi sayı uyumu gösterir. Ana tümcede olduğu gibi yan tümcede de üçüncü kişi çoğul uyumu eyleme seçime bağlı olarak eklenir. Artgönderim ana cümledeki işlevi yan tümcelerde de aynen devam eder.

$$\text{Onun dersi anlatışı bence iyi.} \quad (2.8)$$

Bu cümledeki **Onun** artgönderimi açık adıl olarak gerçekleşmiştir. Bununla beraber boş artgönderim içeren durumlar da olur. Boş artgönderimler benzer bir şekilde yan tümcelerde de görülür. Aşağıdaki tümce buna örnektir.

$$Adam_i [\oslash_i \text{ okula gitmek}] \text{ için otobüse bindi.} \quad (2.9)$$

2.3 Diğer Dilbilimsel Artgönderim Çalışmaları

2.3.1 Bağlama Kuramı (Binding Theory)

Artgönderim konusundaki en önemli çalışmalardan biri de [Chomsky, 1981] tarafından sunulan ilkeler ve değiştirgenler (principles and parameters) kuramıdır. Bu çalışmanın içerdiği Bağlama Kuramı artgönderimlerin nasıl gönderimde bulunacağına ilişkin önemli sözdizimsel kısıtları vurgular. Bu kuram dönüşlü adılları, kişi adıllarını ve sözlüksel ad öbeklerini kapsar. Bağlama bağlı öğenin/öncülün gönderimi ile göndergenin gönderimi aynıdır. Bağlama kuralında gönderimi tümce içinde olan adılar **gönderge** olarak adlandırılır. Gönderimi tümce dışında olan adılar adılsıl (pronominal) ve sıradan adlar ise gönderimsel ifade (referential expression) olarak tanımlanır.

1. He likes himself
2. John likes herself

He adılının gönderimi tümce dışında olduğundan adılsıl, **herself** kurucusu gönderimi tümce içinde olduğu için gönderge ve John ise gönderimsel ifadedir.

Kuram dört ifade tipinden bahseder:

Tablo 2.1: AÖ Sınıflandırma Tablosu

	Açık	Boş (Örtük)
[+ gönderge, - adılsıl]	Sözlüksel Gönderge (kendine, birbirleri)	Ad Öbeği ızı
[- gönderge, + adılsıl]	Adıl (ben, onlar,...)	adıl (pro) ($\emptyset$)
[+ gönderge, + adılsıl]	-	ADIL (PRO)
[- gönderge, - adılsıl]	Ad (Ali, Veli)	değişkenler

Sözlüksel göndergeler [+ gönderge, - adılsıl], gönderimleri sadece tümce içinde olmak durumundadır. Buna en iyi örnek dönüşlü artgönderimdir. Adıl [- gönderge, + adılsıl], tümce içine gönderimde bulunamaz. Tümce dışına gönderimde bulunurlar. Adıllar buna örnektir. Adlar ise [- gönderge, - adılsıl] hiç bir gönderimsel kısıta uğramazlar. Üçüncü kategorideki öge ise denetim ADILı (kontrol PRO) olarak bilinir. Bu yapı ya kişi-dışı bir yorum ile yapılır ya da özne ile nesne tarafından kontrol edilebilen bir yorum yapılabilir. Örnek olarak

$$[\text{ADIL}(\text{PRO}) \text{ Okumak}] \text{ gereklidir.} \quad (2.10)$$

cümlesinde okumak eyleminden önce gizli bir kişi-dışı değişken vardır. PRO özne tarafından kontrol edilebilir.

$$\text{Arda } [\text{ADIL}(\text{PRO}) \text{ Okumak}] \text{ istiyor.} \quad (2.11)$$

Yukarıdaki cümlede, okumak eyleminin önünde PRO vardır. Aşağıdaki tümcede ise nesne tarafından kontrol edilen bir PRO söz konusudur.

$$\text{Herkes}[\text{PRO}(\text{onun})\text{gitmesini}]\text{istiyor.} \quad (2.12)$$

Tüm bu ifadeler Chomsky'nin (Binding Theory) Bağlama İlkelerine uyar.

1. Bir gönderge yerel bir alana bağlıdır. (An anaphor is bound in its Governing Category)
2. Bir adılsıl yerel alana bağlı değildir (A pronominal is free in its GC)
3. Gönderimsel ifadeler özgürdür (R-expression is free)

Bu ilkeler için aşağıdaki örnekleri inceleyelim.

$$John_i \text{ likes } himself_i \quad (2.13)$$

John sevmek kendi(gönderge)

John kendini sever

Bu tümcedeki kendi göndergesi ile John adı Bağlama Kuram'ını birinci ilkesine uygun bir şekilde eşgöndergelidir.

$$He_i \text{ likes } him_{*i/j} \quad (2.14)$$

O(adılsı) sevmek o(adılsı)

O onu sever

Kuramımızın ikinci ilkesine göre adılsılar yerel alanda özgürdürler. Dolayısıyla bu cümledeki *he* ve *him* adılsıları yerel alanda özgürdürler ve eşgöndergeli değildirler. Bu nedenle *he* i ile dizinlenmişken *him* j dizinine bağlıdır.

Bağlama Kuramının üçüncü ilkesi gereği John gönderimsel ifadesi tamamen özgürdür ve O(he) adılı ile eşdizinlenemez.

$$He_i \text{ likes } John_{*i/j} \quad (2.15)$$

O(adılsı) sevmek John(Gönderimsel ifade)

O onu sever

2.3.2 Givon

[Givon, 1983] konusallık ile artgönderim biçimi arasında karşılıklı bir ilişkinin olduğuna değinir. Konusallık değeri yüksek olan göndergeler ya adıyla ya da boş artgönderimle kodlanır. Konusallık değeri düşük olanlar ise tam ad öbekleriyle (Full NP) ile kodlanır.

2.3.3 Fox

[Fox, 1987] ise aynı söylem alanı içinde sisteme dahil edilmiş öncüllerin adıyla ya da boş artgönderimlerle kodlandığını vurgular. Öte yandan başka bir söylem alanındaki öncüller ise Tam Ad Öbeği ile kodlanırlar. [Fox, 1987] söylemin aşamalı olarak sıralanmış önermeler dizisi olduğunu belirtir. Bu aşamalı yapının artgönderim seçimlerinde önemli bir rol oynadığını vurgular. Aşamalı modeldeki temel yaklaşım şu şekildedir: Söylem sıralı bir şekilde okunuyor olsa da anlamlandırılmaları aşamalı olarak gerçekleşir. Kuşku yok ki bunun artgönderim davranışı üzerinde çok büyük bir etkisi vardır.

Metinler aşamalı olarak düzenlenmiş önerme kümeleridir. Önermelerden oluşan kümeler retorik yapılarla gösterilmektedir. Bu Retorik yapılar çekirdek ve niteleyici uydu olmak üzere iki bölümden oluşur. Konuşucunun ana fikri ya da amacı çekirdekte yer alır. Çekirdekteki bilgileri nitelemek için kullanılan diğer

ek bilgiler ise uydu dediğimiz niteliyici yapıda yer alırlar.

2.3.4 Merkezleme Teorisi

[Grosz ve Sidner, 1986] ve [Grosz, 1995] söylem yapısı teorisini geliştirdiler ve odak tabanlı artgönderim çözümlemesi ile ilgili birçok çalışma sundular. Merkezleme Teorisi olarak bilinen bu çalışma yerel bağdaşıklık ele alır ve merkezin hareketine ilişkin bazı kuralları içerir. İşte bu kuralları içeren ve bu kurallar üzerinden oluşturulan artgönderim algoritmalarına merkezleme algoritmaları (Centering Algorithms) denir. Merkezleme Kuramı yerel bağdaşıklık ve belirginlik modeli olarak geliştirilmiştir. Asıl amacı bağdaşıklık örüntülerini oluşturan gönderimsel ifadeleri tanımlamaktır. Bu kurama göre her söylem bir dizi bölüitten oluşur ve her bölüm bir tümceden(sözce) oluşur. [Brennan vd., 1987] bu kuramı bilgisayarlı bir model olarak sunmuştur.

2.4 Türkçe için Dilbilimsel Artgönderim Çalışmaları

[Oktar ve Yağcıoğlu, 1997] yaptığı çalışmada [Fox, 1987]'un çalışmasını örnek alarak üçüncü-tekil-kişi insan göndermelerinde oluşan artgönderim örüntülerini saptamaya çalışmışlardır. Veri tabanındaki metinleri inceleyerek elde ettikleri bulgularda, artgönderim örüntülerinin tümüyle retorik düzenlemeyle yönetilmediğini, artgönderim seçimlerinde yapısal olmayan etkenlerin de rol oynadığını gözlemlerler. Ortaya çıkan bulgular şu şekilde özetlenebilir:

1. Eğer aktif göndermede söz konusu kişi ya da varlık hakkında daha önce bir açıklama varsa, bu öncüle yapılacak bir gönderme adıl ile gerçekleşir.

Aksi halde ad öbeği kullanılır.

2. Bu çalışmada verilen bir örnekte öncül ile adıl arasında dört önermeli bir boşluk vardır. Böylelikle tam ad öbeği kullanma zorunluluğunun olmadığı tespit edilmiştir.
3. Adıl kullanımının uygun olabileceği retorik birimlerde genellikle tam AÖ'lerinin kullanıldığına ilişkin bulgular bu tür metinlerde tam AÖ'lere geleneksel gönderge bağlantısının dışında çok daha başka işlevler yüklendiğini göstermektedir.

[Çeltek ve Oktar, 2004] ise yine [Fox, 1987]'un çalışmasındaki Retorik Yapı Çözümlemesi yöntemini temel alarak metinlerin sadece bir tümceler dizisi olmadığı, aşamalı biçimde düzenlenmiş önermeler dizisi olduğuyla ilgilenmiştir. Bu çalışmada 5000 önermenin adıl ve boş artgönderim örüntüleri çözümlenmiştir. Bu analizde sadece 3. kişi göndergeleri için 317 adet boş artgönderim ve 144 adet açık adıl olduğu tespit edilmiştir. 3. kişi göndergelerde boş art gönderimlerin daha ağırlıklı olduğu kesindir. Öte yandan çalışmadaki ilginç bir bulgu ise 1. ve 2. kişi ile cansız göndergelerine yöneliktir. Bu göndergeler için tercih edilen artgönderimsel biçimin açık adıl olduğu belirtilmektedir. Öte yandan araya giren önermelerde söz konusu kişiye ilişkin bir açıklama yapılmış ise gönderge geriye-atlama yapısında adıl ya da boş artgönderim ile kodlanmaktadır.

Peki adıl ya da boş artgönderim seçime bağlı olarak mı yapılmaktadır ? [Enç, 1986] açık adıların söylem içinde konuşanlar ve dinleyenler tarafından öğrenilmesi gereken edimbilimsel işlevleri olduğundan, boş ulamın ve açık adılın seçime bağlı olmadığını söyler.

[Erguvanlı-Taylan, 1986] söylemde öznenin konu değiştirme işlevi gördüğü durumlarda, açık adıların kullanılmasının zorunlu olduğunu ve boş ula-

mın dilbilgisi dışı yapılara neden olduğu görüşünü savunur. Ve boş adılın söylem özellikleri açısından açık adılarla örtüşmediğini gösterir. Çalışmada gönderimin olduğu üç durumun altı çizilir:

1. Boş artgönderimlerin zorunlu kullanımı
2. Açık adıların zorunlu kullanımı
3. Boş artgönderimin ya da Açık adılın seçimli kullanımı

Yan tümcelerdeki tamlayan pozisyonundaki ifadelerin düşmesi tamlayan durumlu boş bir artgönderim olarak yorumlanır.

$$Erol_i \oslash_i \text{ karısı için herşeyi yapar.} \quad (2.16)$$

2.16 nolu tümcede karısı ifadesinden hemen önce boş bir artgönderim vardır. Bu art gönderimin boş kullanılması zorunludur ki Erol ile eşgönderime sahip olsun. Benzer bir durum aşağıdaki 2.17 nolu tümce için de geçerlidir.

$$\oslash_i \text{ karısı için } Erol_i \text{ herşeyi yapar.} \quad (2.17)$$

Boş art gönderim öncülde önce olsa da farketmez. Bu durumda yine iki öge aynı eşgönderime sahiptir. Öte yandan 2.18 ve 2.19 nolu tümcelerdeki karısı ögesi açık adıl ile kullanılmıştır. Bu durumda söz konusu açık adıl (Onun) Erol ile aynı eşgönderime sahip olmaz.

$$Erol_i \text{ onun}_{*i/j} \text{ karısı için herşeyi yapar.} \quad (2.18)$$

2.19 nolu tümcede açık adıl öncülünden önce yer almasına rağmen 2.18 nolu tümceyle aynı davranışa sahiptir.

$$Onun_{*i/j} \text{ karısı için } Erol_i \text{ herşeyi yapar.} \quad (2.19)$$

Ancak öncüller birden fazla olduklarında, tümce içindeki pozisyonları önemli olur. Örnek olarak aşağıdaki 2.20 nolu tümcede boş artgönderimin öncülü belirsizdir.

$$Ahmet_i Erol_j \text{'a } \mathcal{O}_{i/j} \text{ karısını sordu.} \quad (2.20)$$

Ancak karısını ögesi değişik pozisyonlarında yer aldığında bu belirsizlik ortadan kalkar. Karısını ifadesi 2.21 nolu tümcedeki gibi yerleştiğinde, boş artgönderim Ahmet öncülüne gönderimde bulunur.

$$Ahmet_i \mathcal{O}_{i/*j} \text{ karısını } Erol_j \text{'a sordu.} \quad (2.21)$$

2.22 nolu tümcedeki gibi bir diziliş söz konusu olduğunda boş artgönderim Erol'a gönderimde bulunur.

$$Erol_j \text{'a } \mathcal{O}_{*i/j} \text{ karısını } Ahmet_i \text{ sordu.} \quad (2.22)$$

Çalışmadaki bir başka bulgu ise şudur: Ana tümce öznesi ile aynı eş gönderime sahip yan tümce öznesi boş adıl olmak zorundadır.

$$Erol_i \text{ bana } [\mathcal{O}_i / onun_{*i/j} \text{ toplantıya gelmeyeceğini}] \text{ söyledi} \quad (2.23)$$

2.23 nolu tümcede görüldüğü gibi eğer yan tümcenin öznesi açık adıl olursa gönderimi Erol, yani ana tümcenin öznesi, olmaz. Eğer bu özne boş artgönderim olursa Erol'a gönderimde bulunabilir. Dolayısı ile bu tip örneklerde adılların açık ve boş kullanılması seçime bağlı değildir.

Çalışmada altı çizilen bir diğer önemli başlık ise söylemin gönderim çözümüne etkisidir.

Erol_i yemeğe gelecek miydi ?

$$\text{Nazan bana } [onun_i / \mathcal{O}_i \text{ yedide geleceğini}] \text{ söyledi} \quad (2.24)$$

2.24 nolu söylemde $onun_i / \mathcal{O}_i$ gönderimleri Erol'a gönderimde bulunur. Oysa birinci tümce dikkate alınmaz ise bu durumda söz konusu gönderimler tümce içindeki özne pozisyonunda yeralan Nazan'a gönderimde bulunurlar.

[**Turan, 1996**] Türkçe söylemdeki artgönderimsel ilişkilerin karakteristiğini çıkarmaya çalışmıştır. Bu çalışmasını Merkezleme Teorisi üzerinden yapmıştır. Buna yönelik olarak üç temel olguyu tartışmıştır.

- Hangi ifadeler gönderimler için öncül olma potansiyeline sahiptir. Bazı ifadeler gönderimseldir, bazıları değildir.

- Ayrıca potansiyel öncüller belli bir belirginlik değerine sahiptir. Bu belirginlik değeri nasıl hesaplandığı çalışmanın ikinci olgusudur.
- Üçüncü amaç, boşartgönderim, açık adıl ve özne pozisyonundaki tam ad öbeğinin işlevini ortaya çıkarmaktır.

Bu çalışmadaki bulgular çalışmamızın ileriki aşamalarında kullanılacağından detaylarını daha sonra tartışacağız.

[Enç, 1986] Türkçe'deki boş artgönderim ve açık adılların serbest olmadığını söyler.

Ben çarşıya gidiyorum

⊗ çarşıya gidiyorum (2.25)

2.25 nolu örnekteki ilk tümcede özne açık adıyla ifade edilmiş.İkinci örnekte ise boş artgönderim kullanılmıştır. [Enç, 1986] ilk cümledeki açık adılın söylemin konusunu belirlediğini söyler. İkinci tümce "Neden ceketini giydin" gibi bir soruya cevap olarak verilmiştir. Bu durumda ceketin neden giyildiği söylemin konusudur. Bu konuya yorum yapmak için ikinci tümce kullanılır. Bu tümcede boşartgönderim kullanılmasının sebebi, söylemin konusunun özne olarak belirlenmesinden kaçınmaktır. Ayrıca [Enç, 1986] özne pozisyonundaki açık adılın konu değiştirmek için kullanıldığını söyler. Ayrıca iki ad öbeği arasında bir aykırılık (contrast) yaratmak için açık adıl kullanılır.

Arabayı Ahmet yıkamadı. Ben yıkadım

Bu havada kimse top oynamaz. Ben Oynarım (2.26)

Açık adıl kullanmanın bu iki işlevi - konu değıştirme ve aykırılık yaratma- birbirlerine bağı kavramlardır.

Bölüm 3

BİLGİSAYARLI ARTGÖNDERİM ÇALIŞMALARI

Bu bölümde artgönderim çözümlemesi için ne tip bilgisayarlı tekniklerin kullanıldığına bakacağız. Bu nedenle öncelikle bilgisayarlı artgönderim çözümleme süreçlerinin nelerden oluştuğunu inceleyip Türkçe ve diğer Dünya dilleri için hangi tekniklerin ve çözümlerin önerildiğine bakacağız. Basitleştirilmiş çözümleme süreçleriyle başlayalım.

3.1 Artgönderim Çözümleme Süreçleri

Bu kısımda artgönderim çözümlemesi için yapılması gereken temel adımlardan bahsedeceğiz. Bu adımlar bugüne kadar artgönderimle ilgili yapılan çalışmalarda uygulanan temel adımlardır. Bir çok çalışma aşağı yukarı benzer aşamalardan geçerek artgönderim için bir çözüm oluşturmaya çalışmıştır. Bu aşamaları şu şekilde sıralayalım:

1. Söylem Varlıklarının Bulunması: Bu aşamada söylemde mevcut olan öge-

lerden hangilerinin söylem için bir varlık olduğu saptanır. Bu varlıklar çözüm sistemine söylem varlığı olarak dahil edilirler. Varlıklar, gönderimde bulunabilenler ve bulunamayanlar olarak ikiye ayrılır. Ve yine gönderimde bulunamayanlar ise öncül potansiyeliğine sahip ve sahip olmayanlar olarak ikiye ayrılır. Ya da bir başka ifade ile gönderimsel ifade (referring expression) sınıfında olanlar ya da olmayanlar şeklinde ikiye ayrılabilirler. Dolayısıyla söylemdeki öğeler üç şekilde sınıflandırılır.

- (a) Gönderimsel Varlıklar: Açık Adıllar(Adısal, Dönüşlü, Karşılıklı), Boş Artgönderimler, Sözlüksel Ad Öbekleri.
- (b) Öncül Potansiyeline Sahip Varlıklar (Referring Expressions): Gönderimsel varlıklar da yine bu sınıfa dahil olurlar. Yani bir adıl bir diğer adıla gönderimde bulunabilir. Hatta bu Türkçe metinlerde çok sık olan bir durumdur. Bunun dışındaki ad öbekleri ya da tümceler bu sınıfa dahil olabilir.
- (c) Öncül Potansiyeline Sahip Olmayan Öğeler: Eylemler (çoğu zaman), Sayılar, Sıfatlar vb. Bu öğeler bazı koşullarda gönderimsel ilişki içinde olabilirler. Ama çoğunlukla bu ilişkilerin dışındadırlar.

Bu sınıflandırma, oluşturulacak bir modelin hızını ve performansını artıracaktır. Çünkü bu sınıflandırma ve de eleme sayesinde **Artgönderim Çözümleme Algoritması** gereksiz öğeler ile uğraşmayacağından modelin başarısını olumlu yönde etkileyecektir. Bir artgönderimin olası öncül adaylarının azalması, gönderim potansiyeli taşımayan söylem varlıkları için aday öncül aranmaması gibi durumlar sistemin ve de algoritmasının başarısını ve hızını artıracaktır.

2. Olası Aday Öncüllerin Oluşturulması:

Bu aşama bi önceki aşamanın devamıdır. Bu ve sonraki aşamaların or-

tak özelliği herbir gönderge için tekrarlanmasıdır. Yani tüm göndergeler çözümlenene ya da işlenene kadar bu aşamalar sürekli tekrarlanır. Sözkonusu gönderge/adıl için olası öncül adaylarının oluşturulması için dikkate alınan temel noktalar şunlardır:

(a) Aday öncüllerin, ele alınan gönderge ya da adıl ile aynı söylem sınırları içinde olması

(b) Aday öncüllerin ele alınan adıldan önce gelmesi ¹

3. Öncüllerin Süzülmesi:

Süzme aşaması söz konusu adıl için mantıksız öncüllerin elenmesidir. Bir adıl bir önceki aşamadan geçmiş olası öncüllerden sadece belli kurallara uyanlar ile gönderimsel bir ilişki içinde olabilir. Örnek olarak sayı, kişi, cinsiyet uyumu bu kurallar arasındadır. Ama sözkonusu dilin bu bilgileri kodlaması gerekmektedir. Türkçe bunlardan sadece sayı ve kişi bilgisini kodlamaktadır. Dolayısıyla bu aşamada bizim çalışmamızda kullanacağımız süzme tekniği sadece sayı ve kişi bilgilerini içerecektir. Bunun dışında başka süzme yöntemleri de vardır. Anlambilimsel süzme bunlardan en önemlisidir. Bir varlığın ve eylemin anlamından yola çıkarak belli kurallar sisteme dahil edilebilir. Örneğin sistem şu tip anlamsal kurallar içerebilir; Ağaçlar yürüyemez, Kediler konuşamaz, İnsan nefessiz yaşayamaz. Ama bunlar gerçek dünyada, belli bir bağlamda kabul görecektir anlamsal olgulardır. Başka bir evrende bu kurallar değişebilir.

4. Kalan Öncüllerin Sıralanması Bu aşamada herbir gönderge için bir aday öncül listesi vardır. Böylelikle aday öncüller arasından en uygunu seçi-

¹Adıl her zaman gönderimde bulunduğu söylem varlığından önce gelir. Ama bu kurala istisna oluşturan durumlar da vardır. Mesele İngilizce'de "He's the biggest slob I know. He's really stupid. He's so cruel. He's my boyfriend, Steve." cümlesi dilbilgisel olarak doğrudur ve kullanılan bir yapıdır. Burada **He** adılı gönderimde bulunduğu **Steve** söylem varlığından önce gelir. Bu yapılara İngilizce'de **Cataphora** denir. Türkçe için sık kullanılan karşılığı ise **Öngönderim**dir.

lebilecektir. Eğer bu liste boş gelirse, ya diğer aşamalardan bazıları iptal edilir ya da bu gönderge çözümsüz bırakılır. Bazı sistemlerde bu çözümsüz durum için özgün yöntemler kullanılır. Örneğin adıla en yakın öncül bu adılın öncülü olur. Bu yöntemi deneyimizin bir kısmında kullanacağız. Bu tamamen sistemin gereksinimlerine bağlıdır. Eğer liste boş değilse aday öncüllerin sıralanması ya da belli puanlarla ağırlıklandırılması gerekir. Buna yönelik kullanılan iki önemli bilgi tabanlı kuramdan bahsedebiliriz: Merkezleme Teorisi ve Belirginlik Tabanlı(Salient-based: RAP, MARS, etc...) algoritmalar. Bilgi tabanlı yöntemlerin dışında makine öğrenmesi algoritmaları da kullanılmaktadır. Bu iki yaklaşım - yani bilgi tabanlı yaklaşımlar ve makine öğrenmesi yaklaşımları- sonraki bölümlerde ayrıntılı olarak incelenecektir.

3.2 Bilgi Tabanlı (Knowledge-Based) Artgönderim Çözümleme Çalışmaları

Bir önceki bölümde de bahsettiğimiz gibi artgönderim çözümlemesinin bazı aşamaları vardır. Bu aşamalardan sonuncusu hariç diğerleri hemen hemen tüm yaklaşımlar için aynıdır ve çok büyük bir değişiklik göstermezler. Son aşama hangi öncülün hangi sıralamadan geçeceğine ilişkin bir yöntemini içerir. Son aşama için birçok yaklaşımdan bahsedilebilir. Bu bölümde bilgi tabanlı (knowledge-based) yaklaşımla ilgili ne tür çalışmalar yapıldığı incelenecektir. Bu yaklaşımlar daha çok dilbilimsel teorilerden elde edilir ve öne sürülen model bu teorik arkaplan üzerine oturur. Bu yaklaşımlardan öne çıkanlar tek tek inceleyelim.

3.2.1 Hobbs'un Algoritması(Hobbs's Naive Algorithm)

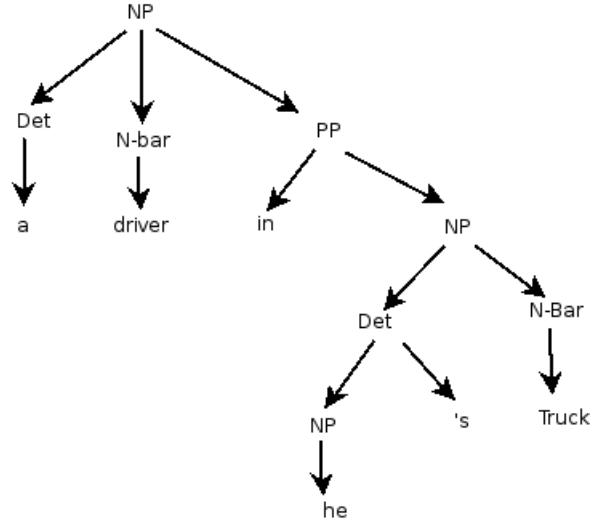
En eski bilgi tabanlı artgönderim çalışması [Hobbs, 1978] tarafından sunulan sözdizim tabanlı bir çözümdür. Her bir gönderge için aday öncüllerin bulunması ve tespiti sözdizim ağacının taranması ile yapılır. Öncül arama işi metnin yönünün tersinedir. Öncül arama sözkonusu göndergenin ya da adılın içinde bulunduğu tümceden başlar ve geriye doğru gider. Bu aramada ağaç soldan-sağa ve yayılma öncelikli arama (breadth-first) yöntemiyle taranır. Tarama sonucunda uyum ve bağlama koşullarına uyan ilk aday gönderge ya da adıl için önerilir. Bu kadar basit olmasına karşın tam ayrıştırılmış bir sözdizim ağacına bağımlı olması bir dezavantaj olarak görülür.

Bu algoritma sözkonusu adılın bulunduğu tümce ile önceki tümcelerin sözdizim ağaçlarını girdi olarak alıp bu ağaçlar üzerinde uygun bir öncül arar. Ağaç tarama algoritmasında belirlenmiş bir Söylem Modeli vardır. Bundan dolayı ağacı tarayacak olan algoritma bir dilbilgisi belirlemek zorundadır. Çünkü sözdizimsel olarak ağacın yapısına ilişkin kabuller sonuçları etkileyecektir. Bu yaklaşımda Ad Öbeğinin (NP) her zaman için bir A-Bar (N-Bar) içerdiği kabul edilir. Aşağıdaki iki tümceyi inceleyelim:

1. Mr. Smith saw a driver in his truck.
2. Mr. Smith saw a driver of his truck. (3.1)

İlk tümcede **his** adılı **a driver** aday öncüle gönderimde bulunabilir. Ama ikinci tümcede bu gönderim mümkün değildir. Bu iki tümcenin sözdizim ağaçlarına bir gözatalım.

Birinci tümcenin sözdizim ağacı Şekil 3.1 'deki gibi olur. Benzer bir şekilde



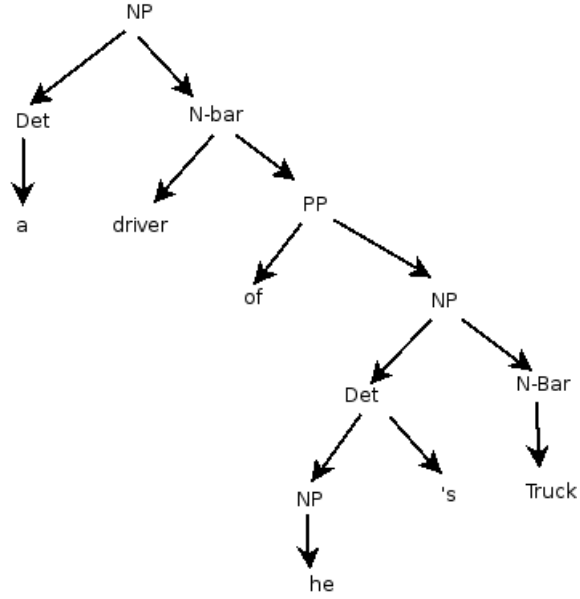
Şekil 3.1: Birinci tümcenin sözdizim ağacı

ikinci tümcenin sözdizim ağacı Şekil 3.2'deki gibi olur.

Hobbs algoritması bu sözdizim ağaçlarını belli bir yöntem ile tarayarak hesaptaki adıl ile sayı-kişi uyumu gösteren öncülleri arar. Bu arama işlemi belli kurallardan oluşur. Bu kuralları şu şekilde sıralayalım:

1. Sözkonusu gönderimi içeren Ad Öbeği ile başla.
2. Üst seviyede ilk karşılaşılan AÖ veya S ² düğümüne çık. Bu düğümü **X**, bu düğüme ulaşmak için izlediğin yolu ise **p** (PATH) olarak kaydet.
3. Bu **X** düğümünün altında ve bu **p** yolunun solunda kalan ağacı soldan sağa (left-to-right) ve yayılma öncelikli arama (breadth-first) biçimiyle tara. Karşılaşılan herhangi bir NP'yi öncül olarak öner.
4. Bu **X** düğümü tümcedeki en üst S düğümü ise, bir önceki tümceye geç ve bu ağacı sağdan sola (left-to-right) biçimiyle tara. Karşılaşılan Ad

²Buradaki S düğümü ana tümceyi ifade eder. Örneğin **S->NP VP** gibi bir tekrar yazma kuralları içeren bir sistem düşünülebilir. Burada NP(Noun Phrase) Ad Öbeğini, VP (Verb Phrase) Eylem Öbeğini ifade eder. Burada S düğümünü tümce ayrıştırma ağacının en tepesindeki düğüm olarak kabul ediyoruz. Çünkü başka sistemlerde bu tepe düğüm daha farklı olabilir.



Şekil 3.2: İkinci tümcenin sözdizim ağacı

Öbeklerini öncül olarak öner. Eğer bu X düğümü en yüksek S düğümü değil ise 5. adımdan devam et.

5. X düğümünden üst seviyede ilk karşılaşılan Ad Öbeği veya S düğümüne çık. Bu düğümü X de ve bu düğüme giden yolu p olarak kaydet.
6. Eğer X düğümü AÖ ise ve p yolu X'in yönettiği ³ herhangi bir N-bar'dan geçmiyor ise, bu düğümü öncül olarak öner.
7. Bu X düğümünün altında ve bu p yolunun solunda kalan ağacı sağdan sola, yayılma öncelikli biçimle tara. Karşılaşılan herhangi bir AÖ'yü öncül olarak öner.
8. Eğer X bir S düğümü ise, p yolunun sağında, X'in altında kalan tüm dalları tara. Ancak Karşılaşılan bir AÖ (NP) veya S düğümünün altına inme. Karşılaşılan AÖ'leri öner.

³Bu yönetme olgusu (c-command) sözdizim ağacındaki düğümlerin birbiriyle olan ilişkisidir. Örneğin A düğümü B düğümünden yuksekte ise, aynı zamanda A düğümü B'den köke giden yol üzerinde değilse ve A düğümü B düğümünün solunda ise A düğümü B düğümünü yönetir (c-command) eder

9. 4. Adıma geri dön.

Bu algoritmadaki 2. ve 3. adımlar dönüşlü adılların seviyeleriyle ilgilenirler. Bununla birlikte 5 ile 9 arasındaki döngü S ve NP düğümleri ile ilgilenir. 4 adım ise daha önceki tümcelere geçişi sağlar.

Görüldüğü üzere Bu algoritma bir dilbilgisine ihtiyaç duyar. Bu algoritma için ilk kullanılan dilbilgisi aşağıdaki gibidir.

$$S \rightarrow NP VP \quad (3.2)$$

$$NP \rightarrow (Det) N\text{-bar} (PP)^*$$

$$NP \rightarrow \text{pronoun}$$

$$Det \rightarrow \text{article}$$

$$Det \rightarrow NP's$$

$$N\text{-bar} \rightarrow \text{noun} (PP)^*$$

$$PP \rightarrow \text{preposition NP}$$

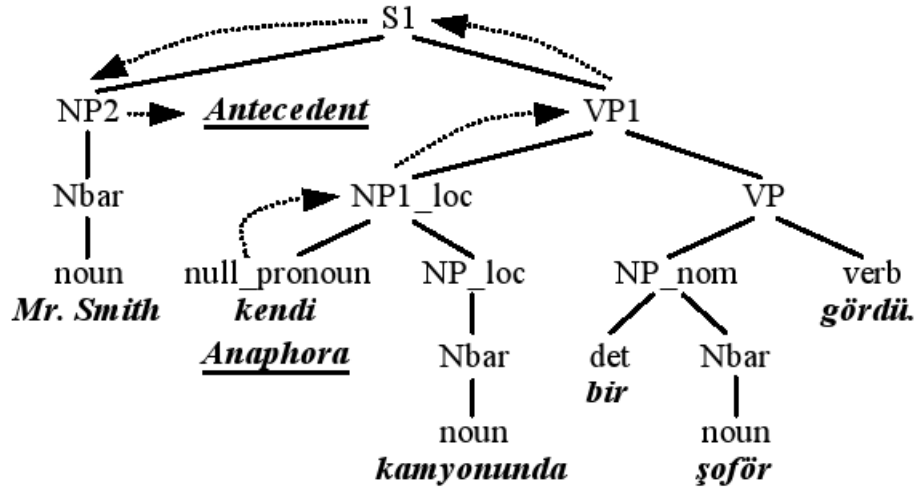
$$VP \rightarrow \text{verb NP} (PP)^*$$

Buradaki * karakterleri sıfır (0) ve sonsuz tekrar anlamına gelir.

[Tüfekci ve Kılıçaslan, 2007] Türkçe bir tümce üzerinde bu algoritmaya açıklık getirmiş. Aşağıdaki tümceyi bu algoritma üzerinden açıklamış.

$$\text{Mr. Smith kendi kamyonunda bir adam gördü} \quad (3.3)$$

Bu tümcenin ayrıştırma ağacı Şekil 3.3 'deki gibidir.



Şekil 3.3: Ayrıştırma Ağacı

Bu örnekte görüldüğü gibi **kendi** dönüşlü adlından, onu yöneten $NP1_loc$ düğümüne geçiliyor. Ardından VP1 ve S1 düğümleri takip edilerek Mr. Smith öncülüne ulaşıyor. Bu aşamada program doğru çalışıyor. Ancak bu tümcedeki **kendi** dönüşlü adlı yerine, onun açık adlı olsaydı bu algoritmanın çözümü yanlış olacaktı. Bunu için başka bir kural oluşturmak gerekiyor. Bu konu bahsettiğimiz çalışmalarda ayrıntılı bir şekilde tartışılmıştır.

3.2.2 Merkezleme Teorisi ve Algoritması (Centering Theory)

Bu kuramın arkaplanı bölüm 2.3.4'de verilmiştir. Bu bölümde de belirtildiği gibi her söylem bir dizi tümceden oluşur. Her tümce daha sonraki adılar için öncül olabilecek bir dizi söylem varlığına sahiptir. Bu söylem varlıkları belli bir sıralama yöntemine göre sıralanır ve **ileriye dönük merkez** (M_i) listesini oluşturur. Bu ileriye dönük merkez listesi söylem boyunca sürekli güncellenir ve sıralama değişir. Bu liste yeni bir listeye değişecektir. Dolayısıyla M_i başka bir liste olacaktır. Eski M_i listesi ise bu aşamadan sonra geriye dönük liste M_g

olacaktır. M_i listesindeki en belirgin öge ya da ilk adılın olası öncülü olabilecek söylem varlığı o listenin odağı olacak ve M_o olarak ifade edilecektir.

3.2.3 BFP

Merkezeleme Algoritmalarından en önemlisi [Brennan vd., 1987] tarafından tasarlanan BFP algoritmasıdır. Algoritma, verili bir adılın çözümlenmesi için, süzme aşamasında aday öncülleri biçimbilgisel kısıtlarla (cinsiyet, sayı, kişi uyumu gibi) sözdizimsel kısıtları(bağlama kısıtları gibi) kullanarak eler. Ardından aday öncüllerin sıralanması merkezeleme teorisi ilkeleriyle yapılır.

Ancak daha sonra yapılan çalışmalar Merkezeleme Algoritmasının açıklarını ya da eksiklerini tamamlamaya çalışmıştır. Merkezeleme Teorisi sadece yerel söylem alanında bir çözüm arar ve verilen bir gönderge ya da adıl için yerel söylemin dışına çıkmaz. Ama gerçek uygulamalarda yerel söylem dışına gönderimde bulunan göndergeler mevcut olduğundan bu açığı kapamaya yönelik başka çalışmalar da olmuştur. Bir diğer problem ise her bir göndergenin gönderimi için bir önceki tümce analiz edilir. Bu da teorik olarak sorun olmasa da pratik uygulamalarda bir sorun olarak görülmektedir. Çünkü bir adıl iki tümce öncesine gönderimde bulunabilir. Aynı zamanda sıralama yöntemi de yeterince belirgin olmadığından bazı varlıklar aynı sıraya sahip olabilmektedir.

Merkezeleme Kuramı dikkat odağındaki değişimleri ve dönüşümleri de incelemektedir. Söylemdeki merkezde meydana gelen değişiklikler **devam**, **içerme**, **yumuşak-dönüş** ve **sert-dönüş** olmak üzere dört şekilde ifade edilmektedir.

[Brennan vd., 1987] bu geçişleri 3.1 nolu tablodaki gibi açıklamıştır.

Bu merkezeleme geçişleri söylem varlıklarının belirginlik derecelerine göre sıralanır. Merkezdeki her türlü değişiklik bu geçişlerle ifade edilmiştir. Merkezeleme

Tablo 3.1: Merkezdeki Değişiklikler

	$Mg(T_n)=Mg(T_{n-1})$	$Mg(T_n)!\!=Mg(T_{n-1})$
$Mg(T_n)=Mo(T_n)$	Devam	Yumuşak Dönüş
$Mg(T_n)!\!=Mo(T_n)$	İçerme	Sert Dönüş

Teorisine göre söylem bu geçişlerden bazılarını daha çok tercih eder. Mesele Devam geçişi ilk tercihtir. Devam geçişi iki önceki ve bir önceki tümcede merkezde yer alan öğrenin yine merkezde olması durumudur. Yani merkezin değişmemesi söylem yapısının ilk tercihidir. Bu anlamda teoriye göre söylemler merkezlerinin fazla değişmemesini tercih eder. Dolayısıyla bu geçişleri belli bir öncelikle sıralamak gerekiyor.

$$\text{Devam} > \text{İçerme} > \text{Yumuşak-dönüş} > \text{Sert-Dönüş} \quad (3.4)$$

Söylemdeki adılların gönderimsel ilişkileri çözümlenirken bu sıralamadan yararlanılır. Tümcedeki adıllar aday öncüllerle birçok biçimde eşleşebilir. Bu eşleşmeler farklı senaryolar oluşturur. Bu farklı senaryolar ise farklı geçiş durumlarına denk gelir. Bu aşamada hangi geçişin seçileceği işte yukarıdaki sıralamaya göre seçilir. Yani **Devam** geçişi **İçerme** geçişine tercih edilir. Benzer bir şekilde **İçerme Yumuşak Dönüşe**, **Yumuşak Dönüş Sert-dönüşe** tercih edilir. Bu tercih yapıldıktan sonra hangi adılın hangi öncüle gönderimde bulunduğu ortaya çıkar. Bu da Merkezleme Algoritmasının çözümü olmuş olur. Bu algoritmanın adım adım bir örneği izleyen bölümlerde verilecektir.

3.2.4 Lappin ve Leass'in Sözdizim Tabanlı Yaklaşımı

[Lappin ve Leass, 1994] belirginlik temelli ve sözdizimsel kısıtlar ve kurallar içeren bir yöntem sunar. Bu yöntemde herhangi anlambilimsel ve edimsel kısıt bulunmaz. Sunulan yaklaşım sadece biçimbilimsel ve sözdizimsel kısıtlar uygular. Yöntemin temel yaklaşımı aday öncüllere biçimbilimsel ve sözdizimsel özelliklerine bakarak belli puanlar(scoring) vermesidir. Her bir öncül topladığı puana göre sıralanır ve en çok puanı toplayan öncül verilen adılın ya da göndergenin öncülü olarak sisteme önerilir. Toplanan bu puan o öncülünün söylem içindeki belirginliğini (salient) ifade eder. Dolayısıyla bu teorinin sadece doğrusal bir yaklaşımı vardır.

Tüm söylem varlıklarına ya da gönderimsel ifadeler bu belirginlik ağırlıklandırılması uygulanır. Ve her söylem varlığının puanı hesaplanır. Ancak bu puan ele alınan adıla göre belirlenir. Her söylem varlığının adıldan bağımsız puanı olmaz. Ele alınan adıl ile olan uyumu en büyük etkendir. Örnek olarak adıl ile öncülün tümce yakınlığı önemli bir etkendir. Aynı zamanda adıl ile sayı/kişi ve cinsiyet uyumu da bir diğer önemli etkendir. Ancak yine de her ögenin ele alınan adıldan bağımsız özellikleri bulunur. Örneğin ele alınan öncülün dilbilgisel görevi (özne, nesne vb.) ağırlıklandırmayı etkileyen güçlü bir bilgidir ve bu bilgi adıldan bağımsızdır.

Şimdi hangi etken için ne gibi bir ağırlıklandırma yapılacağına bakılacak.

[Lappin ve Leass, 1994] Tablo 3.2 içindeki ağırlık değerlerini kendi modellerinde uygulamıştır.

Gönderimi çözümlenecek adıl ile aday öncül arasındaki uzaklık en çok bir tümce kadarsa, öncülün belirginlik değerine 100 puan eklenir. Eğer öncül özne pozisyonundaysa 80 puan eklenir. Öncül varoluşsal pozisyonunda ⁴ ise 70, be-

⁴Buradaki Existential ifadesi İngilizce'de *there* kalıbıyla ilgilidir. Bu yapı içinde bulunan ad

Tablo 3.2: Belirgenlik Türleri ve Ağırlıkları

Etkenin Türü	Ağırlığı
Tümce Yakınlığı	100
Özne Olma Durumu	80
Varoluşsal Vurgusu (Existential Nominal)	70
Belirtme Durumu (Accusative Case)	50
Dolaylı Nesneler (Indirect Object)	40
Baş Ad Öbeği Vurgusu (Head Noun Emphasis)	80
Belirteç Öbeğinde olmama vurgusu (Non-adverbial Emphasis)	50

lirtme durumuna sahipse 50, dolaylı nesne ise 40, baş ad öbeği ⁵ ise 80 ve Belirteç öbeğinde ise 50 puan eklenir. Bu ağırlıklandırma tablosu kullanılarak rastgele seçilen yaklaşık 360 adıl üzerinde algoritma sınıandı. Bu sınama sonucunda % 86 oranında bir başarı kaydedildi.

3.2.5 Kennedy ve Boguraev'in Yöntemi

[Kennedy ve Boguraev, 1996] RAP alogritmasının değiştirilmiş bir versiyonunu sundular. Bu çalışmada kullandıkları sistem, ayrıştırma ağacına (parse tree) ihtiyaç duymuyor. Bu yöntemde dilbilgisel görevler (Grammatical Role) yerine sadece sözdizimsel bilgiye göre, verilen bir adılın aday öncülleri puanlanır ve sıralanır. Bu özelliği ile RAP algoritmasından daha avantajlı bir noktada duruyor.

Çünkü gerçek sistemlerde mümkün olduğu kadar az bilgi kullanarak tasarlanan

öbekleri için ayrıca bir ağırlık değeri eklenir. Çünkü bu yapıda bulunan ad öbekleri söylemde vurgulanırlar. Örnek olarak "There are some books on the table" ifadesindeki **some books** ad öbeği varoluşsal vurgusu içerir.

⁵Öncül söyleme ilk ad öbeği olarak dahil olmuşsa o öncülü ödüllendiriyoruz. Örneğin "Ahmet birden uyandı..." ile başlayan bir söylemde daha sonra her Ahmet öncülü için baş ad öbeği vurgusu yapılacaktır.

modeller her zaman çok daha etkili ve pratik oluyorlar. Bilgiye bağımlılığın en aza indirilmesi son zamanlardaki çalışmalar için önem verilen bir noktadır. Bim-birimsel analiz yapan LINGSOFT isimli modelden gelen bilgileri kullanan bu yaklaşım, RAP algoritmasında olduğu gibi belirginlik ağırlıklandırması içerir. Magazin dergilerinden, ürün tanıtımlarından ve gazete haberlerinden oluşturulan derlem üzerinde test edilen algoritma 306 adıl için uygulandı ve % 75 gibi bir başarı elde etti.

3.2.6 Baldwin's Cogniac

[Baldwin, 1997] az bilgili ve bir takım kurallardan oluşan bir model sundu. Bu kurallar sırayla adılar üzerinde uygulanır. Bu kurallar önemleri ve artgönderim çözümüyle olan ilişkilerine bakılarak sıralanır. Adıla ilişkin uygun bir kural sağlandığında süreç tamamlanır ve sonuca varılır. Dolayısıyla kuralların sırası çok önemlidir.

Bu yöntemde uygulanan kuralları ve bu kuralların ne gibi bir başarı sağladığını aşağıdaki şekilde sıralayalım:

1. Söylemde Teklik (Unique in Discourse): Eğer sadece bir aday öncül varsa onu adılın öncülü olarak seç. *Bu kural 8 kere doğru sonuç verdi.*
2. Dönüslü (Reflexive): Eğer adıl dönüslü ise tümcedeki en yakın adayı öncül olarak kabul et. *Bu kural 16 kere doğru, bir kere yanlış çalıştı.*
3. Önceki ve mevcut tümcedeki tekillik (Unique in current and prior): Eğer bir önceki tümce ya da mevcut tümcede tek bir adıl varsa onu öncül olarak seç. *Bu kural 114 kere doğru, 2 kere yanlış cevap verdi.*

4. İyelik durumundaki adılar (Possessive Pronoun): Eğer adıl iyelik halinde ise (my, your, his, her vb.) ve bir önceki cümlede iyelik halinde olan bir öncül var ise onu adılın öncülü olarak kaydet. *Bu kural 4 kere doğru, 1 kere yanlış çalışmıştır.*
5. Mevcut Tümcede Teklik (Unique current sentence): Eğer mevcut tümcede sadece bir aday öncül varsa, onu adılın öncülü olarak kaydet. *Bu kural 21 kere doğru ve 1 kere yanlış sonuç vermiştir.*
6. Tekil Özne / Adılsı Özne (Unique Subject / Subject Pronoun): Eğer bir önceki tümcenin özne Ad Öbeği sadece bir aday öncül içeriyorsa ve mevcut tümcenin öznesi adıl ise, bir önceki tümcenin ad öbeğini öncül olarak kaydet. *Bu kurallar 11 kere doğru çalışmış ve hiç yanlış sonuç vermemiştir.*

Sırasıyla uygulanan bu kurallar dışına çıkan durumlar için daha sonra iki ek kural daha eklenmiştir. Bu çalışma The Wall Street Dergisi'nden elde edilen derlem üzerinde 200 adıl için uygulandı ve % 92 kesinlik (precision), % 64 anma (recall) oranında bir başarı elde etti.

3.2.7 MARS

[Mitkov, 2002] tarafından geliştirilen MARS (Mitkov's Anaphora Resolution System) isimli sistem tam otomatik çalışabilme ⁶ yeteneğine sahiptir. Yani hiçbir işaret koyulmamış serbest metinler sistem tarafından analiz edilir. Ve içerdği artgönderimsel ilişkiler otomatik olarak belirlenir ve belli bir başarı oranıyla çözümlür. MARS02 kodlu ilk versiyonda toplam beş aşamada işlemler tamamlanır.

⁶Bu sistemin ilk versiyonu yarı-otomatik olarak çalışıyordu. Daha sonraki denemelerde tam otomatik bir sistem kuruldu. Ancak doğal olarak tam otomatik sistem yarı-otomatik sisteme göre daha az başarılı olmuştur.

Birinci aşamada değeriendirilecek işlenmemiş metin sözdizimsel olarak ayrıştırılır. Bu aşamada ayrıştırılmış her bir parçanın biçimbilimsel, sözdizimsel, dilbilgisel analizi yapılır. Ad Öbekleri bu süreçte belirlenmiş olur. İkinci aşamada artgönderimsel adılar, gönderimsel olmayan öğeler makine öğrenmesi algoritmaları aracılığı ile ayıklanır. Üçüncü aşamada her bir adıl için, adılın bulunduğu tümce ile iki tümce öncesine kadar olan metin parçasındaki olası adaylar belirlenir. Bu aşamada mevcut adaylar belli kısıtlar çerçevesinde yarışır. Yarışta sözdizimsel kısıtlar uygulanır. Öncelikle her bir aday öncül adıl ile sayı, kişi ve cinsiyet uyumu içinde olmalıdır. Dördüncü aşamada RAP algoritmasında olduğu gibi her bir aday öncülün puanı hesaplanır. Bunun için etkenlere bağlı olarak adaylar değeriendirilir ve puanlanır. Beşinci aşamada ise en yüksek puana sahip aday adılın öncülü olur. MARS'ın en son versiyonunda bazı güncellemeler yapılmıştır. MARS06 olarak nitelendirilen yeni versiyonda daha katı ve kesin uyum yöntemleri uygulanır. MARS sistemin en önemli özelliği dünyadaki en iyi sistem olması ve bu sistemin metin özetleme (text summarization), bilgi çıkarma (information extraction) ve metin sınıflandırılmasında (text categorisation) kullanılmasıdır.

MARS bilgisayar kullanım klavuzlarından oluşan 8 farklı döküman üzerinde çalıştırılmıştır. Bu derlem 247.000 kelimedenden ve 2263 adıldan oluşur. MARS'ın tam otomatik versiyonundaki deneyler sonunda ortaya çıkan başarı oranı %61.55 olarak ölçülmüştür. Bu oran doğru çözümlenmiş adıl sayısının sistemde bulunan adıl sayısına oranıdır. Öte yandan yarı otomatik versiyonundaki sonuçlar %85-90 arasında çıkmıştır. Bu yöntemin Bulgarca dili adaptasyonu % 72.6, Japonca dili adaptasyonu % 75.8 başarı elde etmiştir.

3.3 Öğrenme Tabanlı Artgönderim Çözümleme Çalışmaları

Makine öğrenmesi yapay zeka alanının en önemli alt alanlarından biridir. Bu disiplin içinde üretilen teoriler diğer disiplinlerdeki hemen hemen tüm alanlara, özellikle de yapay zeka alt alanlarına uygulanabilmektedir. Bu anlamda bilgi tabanlı yöntemlere bir alternatif oluşturur. Öğrenme tabanlı algoritmalar daha önceki deneyim durumlarından bir model oluşturur. Oluşturulan modeller, dayandığı veriyi iyi bir şekilde ifade etmeye çalışır ve bu verinin içindeki özellikle gizli olan kuralları çıkarmaya çalışır. Kısacası, veri madenciliğinin bir parçası olan makine öğrenme algoritmaları insanın bulamayacağı kuralları çıkarmak için kullanılır. En önemlisi ele aldığı veriyi analiz ederek gelecekte oluşacak durumlara tahminde bulunmasıdır. Bu gücü sebebiyle öğrenme algoritmaları bilgisayarlı dilbilim çalışmalarının hemen hemen tüm alt başlıkları için uygulanmaktadır. Bu alt görevlerden biri olan Artgönderim Çözümlemesi için de uygulanmaktadır. Bu uygulamaların ne tür süreçlerden geçtiğine bakalım.

3.3.1 Artgönderim İçin Öğrenme Süreçleri

Öğrenme algoritmaları bir veritabanındaki veriye dayanarak bir model ortaya çıkarır. Bu veritabanı birçok özellik ve tek bir hedef özellik (ya da sonuç) içeriyor ise bu öğrenme yöntemine kontrollü (supervised), eğer sonuç alanı içermiyorsa kontrolsüz (unsupervised) öğrenme denir. Kontrollü öğrenmede amaç, bağımsız özelliklerin hedef özellik ile olan ilişkisini bulmak. Buradaki Hedef ya da sonuç diğer özellik öğelerinden hangilerine daha çok bağlıysa o özellik ön plana çıkar. Bu özellikler birbirinden ne kadar bağımsız olurlarsa ve hedef sınıfa ne kadar bağımlı olurlarsa model o kadar başarılı olur.

Peki Artgönderim çalışmaları bu sürece nasıl bağlanacak ? Söylemdeki her bir artgönderimsel adıl ya da öge kendisinden önce gelen tüm söylem varlıklarıyla eşleşir. Bu eşleşmelerin her biri bir kayıda denk düşer. Bu kayıtların hedef sınıfı, sözkonusu artgönderim ile eşleştiği varlığın ya da gönderimsel ifadenin gönderimsel bir ilişki içinde olup olmadığına bakılarak belirlenir. Eğer iki öge arasında gönderimsel bir ilişki var ise hedef sınıf (sonuç) **EVET** olarak imlenir, aksi durumda ise **HAYIR** olarak imlenir. Bu duruma basit bir örnekle açıklık getirelim. Diyelim ki

$$e1...e2...a1(-> e1)...e3...e4...e5...a2(->e4) \quad (3.5)$$

söylem 3.5' deki gibi bir söylem üzerinde çalışıyoruz. Bu söylemde $e1, e2, e3, e4$ ve $e5$ ögelerinin gönderimsel ifade ve $p1$ ve $p2$ ögelerinin ise adıl olduğunu kabul edelim. Adıl $a1$ 'in $e1$ 'e ve adıl $a2$ 'nin $e4$ 'e gönderimde bulunduğunu varsayalım. Bu durumda oluşturulacak veri kayıt listesi Tablo 3.3'deki gibi olacaktır.

Tablo 3.3: Öğrenme Veritabanı

Adıl	Gönderimsel İfade	Sonuç
a1	e1	EVET
a1	e2	HAYIR
a2	e1	HAYIR
a2	e2	HAYIR
a2	e3	HAYIR
a2	e4	EVET
a2	e5	HAYIR

Bu tablodan da görüldüğü gibi her bir adıl ile ondan önce gelen her bir gönderimsel ifade eşleşiyor. Bu eşleşmeler iki şekilde yapılmaktadır. Birincisi olumsuz eşleşme, ikincisi ise olumlu eşleşme. Bu tablo makine öğrenme algoritması için bir eğitim kümesini temsil eder. Ancak bu aşamada belirtilmesi gereken bir

nokta şudur: **a** ve **e** ile ifade ettiğimiz adıl ve gönderimsel ifadeler yerine bu öğelerin ya da varlıkların özellikleri koyulmalıdır. Yani adıla ait özellikleri ve aday öncüle ait özellikler bu satırlarda bulunmalıdır. Ayrıca adıl ile aday öncül arasındaki yakınlık, uyum gibi ilişkişel bilgiler de bu özelliklere eklenmelidir. Böylelikle ortaya çıkan model, çözümü belli olmayan bir söylem üzerine uygulanır ve bir çözüm sunar. Makine öğrenmesi tekniğini Türkçe artgönderim çözümlemesi için uygulayan henüz yoktur. Bu nedenle bir sonraki bölümde öğrenme yaklaşımı içeren bu tekniği diğler diller için kullanan çalışmalara bakacağız.

3.3.2 Diğler Dillerdeki Öğrenme Tabanlı Çalışmalar

[Aone ve Bennet, 1995.] Japonca metinlerde geçen farklı türdeki artgönderimler için makine öğrenme tabanlı bir sistem geliştirdiler. Bu çalışma artgönderim problemine makine öğrenmesi yaklaşımını sunan ilk çalışmadır. Bu çalışmadan sonra birçok makale bu çalışmaya atıfta bulunmuştur. Öğrenme algoritması için gereken özellik vektörü için toplam 66 özellik kullanmışlardır. Bu özellikler sözlüksel bilgiler, sözdizim bilgileri, anlamsal bilgiler ve adıl ile öncül arasındaki mesafe bilgisinden oluşur. Çalışmada iki olumlu örnek yaratma metodu kullanılmıştır: Geçişli ve geçişsiz. Geçişli olumlu örnekler, herbir artgönderim ile dahil olduğu artgönderimsel zincir üzerindeki onlardan önce gelen tüm ad öbekleriyle oluşturulan örneklerden oluşur. Geçişsiz örnekler ise, herbir artgönderim ile dahil oldukları artgönderimsel zincir üzerindeki onlardan önce gelen ilk (en yakın) ad öbekleriyle oluşturulan örneklerden oluşur. Olumsuz örnekler ise benzer bir şekilde her bir artgönderimsel ifade ile ondan önce gelen ve artgönderimsel zincir üzerinde olmayan gönderimsel ifade eşleşmelerinden yaratılır. Bu şekildeki eşleşmeler eğitim ve sağlama kümesini oluşturur. Bu deney [Quinlan, 1993.] - C4.5 karar ağacı algoritmasını kullanmış ve çalışmanın sonunda öğrenme ta-

banlı sistemin bilgi tabanlı sisteme oranla daha başarılı olduğu vurgulanmıştır. Bu çalışmadaki model 1971 adlı kapasiteli eğitim kümesi ile eğitilmiş 1359 adlı kapasiteli sağlama kümesi ile sınanmıştır. Bu test sonucunda alınan en iyi anma sonucu (recall) % 70.20, kesinlik sonucu (precision) % 88.55 ve f-ölçütü ise % 77.27 olarak ölçülmüştür.

C4.5 karar ağacını kullanan bir başka çalışma [McCarthy ve Lehnert, 1995.] tarafından uygulanmıştır. Ad öbekleri arasındaki gönderimsel ilişkiler (coreference resolution) gibi tüm ilişkileri kapsayan bir sistem kurulmuştur. Kullandıkları 1230 örnek durumun % 26'sı olumlu ve %74'ü olumsuzdur. RESOLVE ismini verdikleri sistem MUC-5 isimli İngilizce metinlerden oluşan bir derlem üzerinde denenmiş ve başarılı sonuç alınmıştır. Bu çalışmada da yine öğrenme tabanlı yöntemlerin bilgi tabanlı yöntemlere göre daha başarılı olduğu vurgulanmıştır. Çalışmada kullanılan ağaç budamasız olduğunda anma, kesinlik ve f-ölçütü sırasıyla % 85.4, % 87.6 ve % 86.5 çıkmıştır. Budamalı bir ağaç kullanıldığında ise sırasıyla %80.1, % 92.4 ve % 85.8 sonucu elde edilmiştir.

[Soon vd., 2001.] yine karar ağacı öğrenme algoritması kullanarak az bilgili ⁷ bir yaklaşım sunmuştur. Toplam 12 adet özellik kullanılmıştır. Sistem bir sözlüksel özellik (eşsesli), sekiz dilbilgisel özelliksel (cinsiyet, sayı, ek, ad öbeği türü vb.), iki anlamsal özellik ve bir pozisyon özelliği (ad öbekleri ve adılar arasındaki uzaklık bilgisi) içerir. Bu çalışmada olumlu örnekler ile olumsuz örnekler arasındaki denegersizliği ⁸ gidermek için olumsuz örneklerin sadece bazıları alınmıştır. Olumsuz örnekler şu şekilde oluşturulmuştur: Her bir adlı ya da gönderimsel ad

⁷Az bilgili yaklaşımlar bilgiye bağlılığı azaltmak ve böylelikle daha pratik, daha hızlı, daha az bağımlı bir sistem oluşturmak isterler. Özellikle gerçek uygulamalarda az bilgiyle çalışmak çok önemlidir

⁸Olumlu örnekler olumsuz örnekler göre daha az olduğundan bu oran bazı uygulamalarda yüzde 10 seviyesinin altına düşer. Bu seviye makine öğrenmesi algoritmaları için oldukça zararlıdır. Çünkü öğrenme algoritmaları her iki sınıf değeri -yani evet ve hayır ya da 1 ve 0- için mümkün olduğunca dengeli bir veritabanı olmasını ister. Bu dengesizlik ve bunun giderilmesi öğrenme tabanlı artgönderim çözümlemelerinin en ciddi problemlerindendir.

öbeği ile onun en yakın öncülü arasında kalan ve gönderimsel bir ilişki içinde olmayan ad öbekleri ya da gönderimsel ifadeler ile oluşan küme olumsuz küme olarak belirlenir. Oluşturulan sistem, gönderimsel ilişkileri önceden belirlenmiş iki veri kümesi (MUC-6, MUC-7) üzerinde denenmiş ve başarılı bir sonuç elde etmiştir. 13.000 kapasiteli bir derlem üzerinde gerçekleştirilen deneylerden çıkan sonuçlar % 52 /anma, % 68 /kesinlik ve % 58.9 /f-ölçütü şeklindedir.

3.4 Türkçe için Bilgisayarlı Artgönderim Çözümleme Çalışmaları

Türkçe'ye yönelik dilbilimsel çalışmaların bolluğuna karşın bilgisayarlı dilbilimsel çalışmaları çok fazla değildir. Artgönderim çözümlemesi için de bu durum geçerlidir. Türkçe için yapılan ilk bilgisayarlı çalışmalardan biri

[Tın ve Akman, 1994] tarafından yapılmıştır. Durum Anlambilimini (Situation Semantic) temel alan bir çerçeve sunulmuştur. Bilgisayarlı bir çerçeve içinde çözüm sunan bu yapı BABY-SIT olarak adlandırılmıştır. Bu durum teorisine göre bir söylem yapısını bireyler, onların özellikleri, birbirleri arasındaki ilişkileri ve durumları belirler. Ancak bu teorik çerçevede sunulan bilgisayarlı model deneysel bir sonuç sunmamıştır. Dolayısıyla bir başarı oranı belirtilmemiştir.

[Turhan-Yöndem ve Şehitoğlu, 1997] tümce içi boş artgönderimler için öbek-yapı dilbilgisi çerçevesinde bir çalışma sundular. Bu çalışmada boş artgönderimlerin çözümünün tümce bileşenlerinin sırasına bağlı olduğunu belirtilmiştir. HPSG (Head-Driven Phrase Structure Grammar) dilbilgisi tabanlı bir ayrıştırıcı kullanılmıştır. Çalışmada vurgulanan bir başka nokta ise çözümlemenin anlambilimsel bağlantılara bağlı olduğudur. Konuşucular ve dinleyicilerin

artgönderimlerin çözümü için anlambilimsel bağlantıları kullandıkları belirtilmiştir.

[Yüksel ve Bozşahin, 2002] önerdikleri sistemde bağlama teorisi ve merkezleme teorisi için bazı kurallar kullanarak yerel ve yerel olmayan gönderimleri modellemişlerdir. Yerel gönderimler bağlama kuramının temel kurallarına göre yapılandırılmıştır. Yerel olmayanlar tümce dışı bir gönderim çerçevesinde ele alınmıştır. Çalışmanın içerdiği bir diğer konu ise Türkçe'nin adıl düşürme yapısı ile ilgilidir. Bu başlık altında adılların nasıl düştüğü ya da nasıl boş artgönderime dönüştüğü tartışılmıştır. Bu bilgisayarlı yaklaşım deneyi makine çevirisi sisteminde test edilmiştir ve başarı oranı %70 olarak kaydedilmiştir.

[Tüfekci ve Kılıçaslan, 2005] Hobbs Algoritmasını Türkçe metinler üzerinde uygulamıştır. Bu uygulamada kendi oluşturduğu derlem üzerinde deneylerini gerçekleştirmiştir. Kullanılan model tümcelerin ayrıştırma ağaçları üzerinden bir analiz yapar. Dolayısıyla her bir tümcenin ayrıştırma ağacı da derleme girilmiştir. Bu çalışmada elde edilen başarı oranı % 84 olarak belirtilmiştir.

[Küçük, 2005] ve **[Küçük ve Turhan, 2007]** az bilgili adıl çözümleme modeli sunmuştur. Sistem sınırlı dilbilimsel ve anlamsal bilgi kullandığı için az bilgili bir model olarak ifade edilmiştir. Kural tabanlı bu çalışmada ise % 85.3 anma ve % 88 kesinlik sonuçlarına ulaşılmıştır.

Bölüm 4

TÜRKÇE İÇİN ARTGÖNDERİM ÇÖZÜM DENEMELERİ

Bu bölümde Türkçe için artgönderim çözümünün nasıl uygulandığı anlatılacaktır. Bir çok farklı algoritma sınanarak, ortaya çıkan sonuçlar üzerinden bir tartışma oluşturulacaktır. Asıl amacımız farklı yaklaşımları Türkçe metinler üzerinde ne gibi tepkiler verdiğini incelemektir. Bu deneyler daha sonraki deneylerin için yol gösterici olduğuna inanmaktayız.

Diğer diller için, özellikle İngilizce için son yirmi yılda birçok hesaplamalı model ve yaklaşım sunulmuştur. Oldukça fazla sayıda olan bu çalışmaların ön plana çıkanları önceki bölümlerde tartışıldı. Bu aşamadaki amaç bu yaklaşımlardan bazılarını ele alarak Türkçe için uygulanabilirliğini, nasıl uygulanacağını ve sonuçlarını tartışmaktır. Tartışmadaki en önemli vurgu **Bilgi Tabanlı** (Knowledge-Based) yaklaşımlarla Öğrenme Tabanlı yaklaşımların karşılaştırılmasıdır. İki ayrı yaklaşımın Türkçe derlem üzerindeki etkisi sonuçları ekseninde değerlendirilecektir.

4.1 Derlem

Bu çalışma kapsamındaki en önemli unsurlardan biri, üzerinde deneylerimizi gerçekleştireceğimiz Türkçe derlemidir. Türkçe için bir kaç derlem çalışması yapılmıştır. Ancak bu derlemlerden araştırmacılara açık olanlar eksik bilgiler içermektedir. Özellikle söylem analizleri için çok uygun değildir. Çünkü derlemin içerdiği söylemler sıralı ve tutarlı tümcelerden oluşmamaktadır. Tümce düzeyindeki çalışmalar için uygun olmasına karşın, söylem düzeyi için çok uygun değildir. Bir başka olumsuzluk ise artgönderimsel ilişkilerin bu derlemlerde belirtilmemiş olmasıdır. Bu çalışmada ihtiyaç duyulan ve üzerinde deneyler yapılacak olan derlemin artgönderimsel ilişkilerinin belli olması gerekir. Bu nedenden dolayı 20 farklı çocuk hikayesinden oluşan bir örnek derlem oluşturulmuştur. Bu derlemin toplam 10165 kelimeden oluşur ve içindeki adıl sayısı ise 1114'tür. Bu adılar tiplerine ve derlemdeki oranlarına göre şu şekilde sıralanır;

1. %82.3 kişisel adıl (personal pronoun)
2. %6.6 yerel adıl (locative pronoun)
3. %10.7 dönüşlü adıl (reflexive pronoun)
4. %0.4 karşılıklı adıl (reciprocal pronoun).

Derlem özellik-değer şeklinde bir veritabanı tablosu olarak kaydedilmiştir. Bu derlemin özelliklerini şu şekilde sıralayalım.

- **Durum Eki:** Adılın ya da aday öncülün dilbilgisel durum eki; Yalın hali, Belirtme Hali, Yönelme Hali, Ayrılma Hali, İlgi Hali, Bulunma Hali
- **Dilbilgisel İşlev:** Bir ögenin Özne ya da Konu olma durumu

- **Açıklık :** Bir adılın açık ya da boş olup olmadığı
- **Adıl Tipi :** Olası adıl tipleri kişisel, yerel, dönüşlü ve karşılıklı.
- **Anlamsal Tip :** Bir ögenin olası anlamsal tipleri insan, hayvan, bitki, yer, cansız varlık ve soyut varlık olabilir.
- **Kişi Sayı:** Kişi sayı bilgisi bir ögenin tekil-çoğul olma durumu ile 1., 2. ve 3. kişi olma durumu.
- **Pozisyon:** Ögenin hangi söylemde, hangi tümcede ve hangi sırada olduğu
- **Öncül Pozisyonu** Bir adılın işaret ettiği öncülün pozisyonu
- **Gönderimsel İfade:** Bir söylem varlığının gönderimsel ifade olup olmadığı

Yapılan tüm artgönderim çalışmalarında kullanılan özellik sayısı ([McCarthy ve Lehnert, 1995.]) 8’den ([Aone ve Bennet, 1995.])66’a kadar değişebilmektedir. Sistemimiz yarı otomatik olduğu için bu özelliklerin bir kısmı elle bir kısmı yardımcı kütüphaneler ile otomatik olarak gerçekleştirilmiştir. Özelliklerin elle girilmesi için java ile yazılmış bir arayüz tasarladık. Şekil 4.1’de de görülen arayüz ile derleme yeni söylemler girebiliyoruz ve onları güncelleyebiliyoruz. Kullanılan bu arayüz aynı zamanda web tabanlı bir versiyona da sahiptir. Bu web arayüzü ile birçok kişi aynı anda sisteme bağlanıp söylem dahil edebiliyor.

Derlemdeki bu özelliklerin bazıları yardımcı kütüphaneler kullanılarak otomatik girilmiştir. Bu açıdan oluşturulan sisteme yarı-otomatik artgönderim çö-

Word	ID	Case	isSbj	isRE	numPer	Overt	refer	wType
Annem	108:0:0	nom	☒	☒	a3sg	☐		noun
bana	108:0:1	dat	☐	☒	a1sg	☐		noun
bu	108:0:2	empty	☐	☐	empty	☐		empty
öyküyü	108:0:3	acc	☐	☒	a3sg	☐		noun
sık sık	108:0:4	empty	☐	☐	empty	☐		empty
anlatırdı.	108:0:5	empty	☐	☐	empty	☐		empty

Şekil 4.1: Derlem Girme ve Güncelleme Arayüzü

zümleme sistemi denebilir. Bu yardımcı kütüphanelerden biri biçimbilimsel analiz yapan Zemberek Kütüphanesidir ¹

Girilen söylemlerin genişliği içerdiği adılların gönderimde bulunduğu öncülleri kapsayacak şekilde belirlenmiştir. Yani gereksiz tümceler söylemden çıkarılmıştır. Bu nedenden adılların oranı çokmuş gibi durmaktadır. Her bir adılın öncülü mutlaka söylem içerisinde bulunmaktadır. Derlemdeki en büyük söylem 26 tümceden oluşur.

4.2 Gönderimsel İfadelerin Tespiti

En önemli aşamalardan biri söylemdeki öğelerin birer söylem varlığı olup olmadığına karar vermektir. Eğer bir öğe söylemde gönderimsel varlık olamıyorsa, daha sonraki tümcelerden ona bir gönderimde bulunulamaz. Bu anlamda, bir söylem ögesi söylem varlığı olamıyorsa bir adılın olası öncülü de olamaz ve öncül aday listesinden çıkarılır. [Turan, 1996] bu olgu için aşağıdaki kuralları tanımlamıştır. Bu kuralların bazıları çalışmada oluşturulan derlemin işaretlenmesi için kullanılmıştır. Kullanılan kurallar şu şekilde sıralanabilir.

¹Zemberek biçimbilimsel analiz yapmak için kullanılan açık kaynak bir yardımcı Java API'sidir. Kelimelerin kökünü bulma, eklerini belirleme gibi sözlük düzeyi işlemleri yapabilmektedir.

1. Belirtisiz yüklemsel nominaller (indefinite predicative nominal) söylem varlığı oluşturamazlar.

Her şey sonsuz bir yarış.

Ben sevmem yarışları/*onları (4.1)

Yukarıdaki söylemde birinci tümcedeki **yarış** (indefinite predicative nominal) belirtisiz yüklemsel olduğundan **onları** adlı ile gönderimsel bir ilişki içinde olamaz.

2. Bir işaret ile kullanılan ad bir söylem varlığı oluşturur.

Şu elbise çok güzel görünüyor. Onu denemek istiyorum. (4.2)

Yukarıdaki söylemde elbise bir isimdir ve şu işareti ile kullanılmıştır. Bu kullanım bir söylem varlığı oluşturduğundan **elbise** söylem varlığına **Onu** adlı ile ulaşılabilir.

3. (Accusative case) Belirtme durum eki genel anlamda kullanılan isimleri söylem varlığı yapabilir.

Ben hayvanları severim. (4.3)

Onların acı çekmelerini istemem.

Bu söylem örneğinde, hayvanlar bir söylem varlığıdır. Çünkü belirtme durum eki ile kullanılmıştır. Dolayısı ile bir adlı ile ulaşılabilirler.

4. Belirtme takısı içindeki generic kullanılan isimler söylem varlığı oluştururlar. Fakat belirtme eki yoksa bir söylem varlığı oluşturamazlar.

Ben eti severim. (4.4)

Onu her zaman yerim.

Yukarıdaki söylemde **et** adı belirtme durum eki ile kullanılmıştır. Dolayısı ile ulaşılabilir. Ancak aşağıdaki söylemde **et** adı hiçbir durum eki takısı almamıştır. Sonuç olarak bir söylem varlığı oluşturamaz ve bir sonraki tümceden yapılacak adılsal gönderme hatalı olur.

Kasap iyi et satar. (4.5)

Ben hep eti *ondan alırım.

4.3 Bilgi Tabanlı Çözüm Önerileri

Bu bölümde artgönderim çözümü için uygulanacak bilgi tabanlı yöntemlerden bahsedilecektir. Bu yöntemler dilbilim çalışmalarından oluşturulan teoriler üzerine kurulur ve bazı kurallardan oluşur. Bu yöntemlerden bazıları daha önce tartışıldı. Bu bölümde bu yaklaşımlardan ikisi ele alınarak Türkçe derlem için bazı denemeler yapılacak ve sonuçları tartışılacaktır.

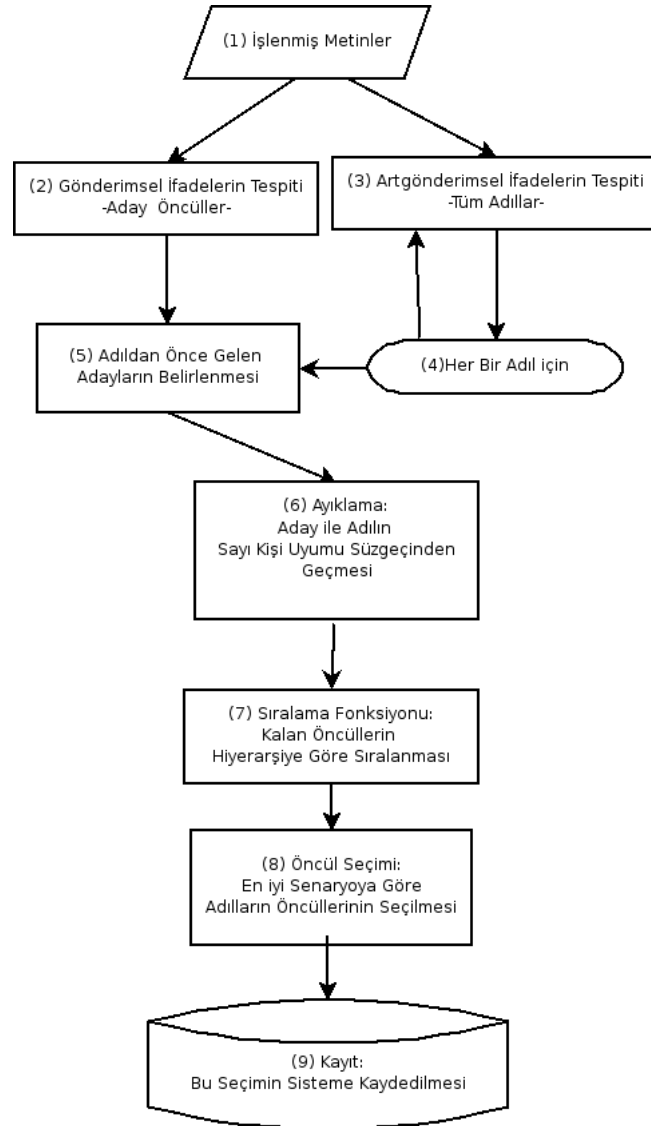
4.3.1 Merkezleme Tabanlı Çözüm Önerisi

Merkezleme Tabanlı yaklaşımlardaki süreç daha önceki bölümlerde bahsedilmiştir. Bu çalışmada uygulanan Merkezleme Teorisine dayanan yöntem ayrıntılı bir şekilde ele alınacaktır. Sistemin bir akışı Şekil 4.2’de gösterilmiştir. Bu şekilde sistem dokuz aşama ile ifade edilse de her aşama kendi içersinde başka aşamalar içerebilmektedir.

(1) nolu öge, işlenmiş metin (Annotated Text), daha önceki süreçlerden (sözdizimsel, anlambilimsel, biçimbilimsel analizlerden) geçtiğinden, bu analizlere ilişkin işaretleri ve bilgileri içermektedir. (2) nolu aşamada gönderimsel ifadeler, (3) nolu aşamada adısal ögeler belirlenir. (4) nolu aşama her bir adıl için bir iterasyon başlangıcını ifade eder. Sistem her seferinde bu noktaya geri döner. (5) nolu aşama ise adıldan önce gelen tüm öncüllerin listesini içerir. Merkezleme Kuramı gereğince her bir adıl hep bir önceki tümcedeki öncüllerle bağlantı oluşturur. Dolayısı ile bu aşamada bir önceki tümcede yer alan öncüller listelenir.

Aday Öncüllerin Ayıklanması

(6) nolu aşama ayıklama aşamasıdır. Her bir gönderge ya da adıl için onu önceleyen bir aday öncül listesi oluşturduktan sonra bu listenin ayıklanması ve gereksiz öncüllerin listeden çıkarılması gerekir. Bunun için önce gereksiz öncüllerin belirlenmesi gerekiyor. Uygulanan yöntemler sözdizimsel ve anlamsal olabilir. Sözdizimsel ayıklamada adılın sayı kişi bilgisiyle ona aday olacak öncülün sayı kişi bilgisinin aynı ya da tutarlı olması gerekiyor. İngilizce için geçerli olan cinsiyet (gender) uyumu Türkçe için geçerli değildir. Öte yandan adıl, ana tümcenin ya da yan tümcenin öznesi pozisyonunda ise o tümcenin yüklemiyle kişi-sayı uyumu içinde olmalıdır. Bu koşullara uymayan aday öncüller listeden



Şekil 4.2: Merkezeleme Yaklaşımının Akış Diagramı

çıkarılır. Benzer bir şekilde anlamsal kısıtlamalar ya da ayıklamalar da yapılabilmektedir. Örneğin "bir ev bağıramaz", "bir köpek kitap okuyamaz" gibi anlamsal kısıtlar uygulanabilir. Ama edebiyatta bir evin bağırması ya da bir çocuk hikayesinde bir köpeğin kitap okuması olağan bir durum olabilir. Dolayısıyla, çalışma alanına göre bir anlamsal kurallar oluşturulmalıdır.

Bu çalışma için uygulanacak ayıklama yöntemi sadece sayı-kişi uyumu olacaktır. Çünkü diğer ayıklama teknikleri bilgiye çok fazla ihtiyaç duyarlar. Az bilgili bir yaklaşım için sadece sözdizimsel bilgi ile bu ayıklamanın iyi sonuçlar verdiği gözlemlenmiştir.

Sıralama

Bir sonraki aşama (7) numaralı sıralama aşamasıdır. Merkezleme Teorisi yaklaşımındaki en önemli aşama gönderimsel ifadelerin ya da aday öncüllerin nasıl *sıralanacağı* aşamasıdır. Daha önce yapılan çalışmada, ([Yıldırım ve Kılıçaslan, 2005]) bu sıralama hiyerarşisini aşağıdaki şekilde tanımlamıştır.

$$\begin{aligned} & Agent > Time > Duration > Location > \\ & Instrument > Manner > Benefactive > \\ & Theme(orPatient) > Source - Goal \end{aligned} \quad (4.6)$$

Konu ilişkili (Thematic relation) bu sıralamanın bilgiye bağımlılığı fazla olduğundan ve aynı zamanda bu tez çalışmasında az bilgili bir yaklaşım denendiğinden başka bir sıralama önerilmesi gerekmiştir. Bu sıralamayı, tespit edilmesi basit olan ad öbeklerinin isim halleri (case) üzerinden yapılması uygun bulunmuştur. 4.1 nolu bölümde bahsedilen derlem üzerinden elde edilen istatistiksel sonuç Tablo 4.1’de gösterilmiştir. Bu tablo gönderimde bulunulan söylem varlık-

larının almış olduğu isim hallerinin sıklığını (case frequency) ifade etmektedir. Bu istatistiksel tablo isim hallerinin söylem varlıklarının belirginlikleri üzerindeki etkisini gösterir. Daha çok hangi isim haline sahip söylem varlıkları sonradan adıllaştırılarak ifade edilmiştir? Bu sorunun tek bir cevabının olmadığını bu tablo açıklıyor.

Tablo 4.1: Öncüllerdeki İsim Hallerinin Oluşma Sıklığı

Öncüllerin Durum Eki	Sıklığı
Yalın Hal (NOM)	0.69
Belirtme Hali	0.12
Yönelme Hali	0.06
Ayrılma Hali	0.02
Bulunma Hali	0.01
İlgi Hali	0.10

En çok vurgu alan isim hali **Yalın Hal**dir (Nominal). Bu durum beklenen bir sonuçtur. Çünkü Yalın Hal tümcenin öznesinin sahip olduğu haldir. Diğer isim halleri ise nesne öğelerinin sahip olduğu hallerdir. Dilbilgisel rolün belirginlik üzerindeki etkisi daha önce ifade edilmişti. Özne nesneden belirgin olduğundan bu tablodan bir hiyerarşi çıkarmak çok mantıksız olmayacaktır. Bu istatistiksel tabloyu temel alarak aşağıdaki sıralama oluşturulabilir.

$$\begin{aligned} &\text{Yalın Hal} > \text{Belirtme Hali} > \text{İlgi Hali} > \\ &\text{Yönelme Hali} > \text{Ayrılma Hali} > \text{Bulunma Hali} \end{aligned} \quad (4.7)$$

Görüldüğü gibi oluşturulan sistem az bilgili bir çerçeve içindedir. Sadece ismin hallerine bakılmıştır. Bu anlamda Merkezleme Teorisi temelinde yapılan çalışma diğer çalışmalara göre en az bilgi kullanan sistemdir. Bu sıralamada yalın hale sahip söylem varlıkları daha belirgin ya da daha önemli oluyor. Sıralamada sağa doğru gidildikçe önem değerleri azalmaktadır.

(8) nolu aşamada, (7) nolu sıralamaya göre sıralanan öncüllerden ilki ele alınan adılın seçili öncülü haline gelir. Ancak bir tümcede birden fazla adıl varsa hangi adıl en belirgin öncül ile eşleşecektir ? Bu sorunun çözümü yine Merkezleme algoritması yanıtlamaktadır. Birden fazla adıl olduğu durumlarda bu adıllar yukarıdaki sıralamaya göre sıralanır. Bu sıra göz önüne alınarak adaylar eşleştirilir. Yani aslında burada en iyi birleşimleri seçmemiz gerekiyor. Şimdi ilerleyen bölümlerde bu ifadelerle biraz daha açıklık getirelim.

En İyi Birleşimin Seçimi

Verili bir tümcedeki tüm adıllar için bir önceki tümcedeki tüm gönderimsel ifadelerden oluşan adıl-aday çiftleri üzerinden birden fazla ilişki senaryosu oluşturulur. Bu birleşimlerden hangisinin seçileceği Merkezleme Teorisi yardımıyla belirlenir. Bu senaryoların veya birleşimlerin yol açtığı geçiş dikkate alınır. Her bir birleşimin geçişi 3.4 nolu sıralamaya göre belirlenir. En iyi geçişe sahip birleşim bu aşamanın çıktısı olur. Daha öncede belirtildiği gibi Merkezleme Teorisine göre söylemler merkez var olan varlıkların yine merkezde kalması eğilimindedirler. Yani bir öge bir önceki tümcede merkezde ise, bir sonraki tümcede de merkezde kalması tercih edilir. Bunun için Merkezleme Teorisi Bölümüne bakılabilir.

Son aşamada, yani (9) nolu aşamada ise tüm karar verilen bu eşleşmeler sisteme kaydedilir. Ve bir sonraki tümceye geçilir. Bu geçişten sonra tüm listeler tekrar güncellenir.

Basit Bir Örnek

Basit bir örnek ile bu uygulamaya açıklık getirilebilir. Aşağıdaki üç tümceli söylem ele alındığında yöntem şu şekilde uygulanabilir.

T_1 :Ayşe Ali'yi yolda gördü

T_2 : Ona elindeki kitapları gösterdi

T_3 :Ve onları eve götürmesini rica etti (4.8)

T_1 için oluşturulacak ileriye dönük liste, $M_i(T_1)$, **Ayşe**, **Ali** ve **yol** söylem varlıklarını içerir.

1. Ayşe (Yalın Hal)
2. Ali (Belirtme Hali)
3. Yol (Bulunma Hali)

Bu söylem varlıklarından Yol ögesinin gönderimsel ifade elemesinden geçemediği varsayılın. Bu liste 4.7 nolu hiyerarşiye göre sıralandığında

$$\text{Ayşe} > \text{Ali} \quad (4.9)$$

gibi bir sıralama oluşur.

T_2 için oluşturulacak ileriye dönük liste, $M_i(T_2)$ ise, $\emptyset$ (**boş artgönderim**), **Ona** ve **Kitap** söylem varlıklarını içerir. Bu söylem varlıklarının elemekten başarılı bir şekilde geçtiği varsayılın.

1. $\emptyset$ (boş artgönderim), Üçüncü Tekil Kişi
2. Ona, Açık adıl, Yönelme Hali (Dative)

3. Kitap, Çoğul, Belirtme Hali

Bu liste sıralandığında ise

$$\emptyset > \text{Ona} > \text{Kitapları} \quad (4.10)$$

gibi bir sıralama oluşur. Bu listedeki adılların her biri için aday öncül belirlenmesi gerekiyor. Her iki adıl için {Ayşe, Ali} listesi olası öncül listesidir.

Öncül Listesi($\emptyset$) : (Ali, Ayşe)

Öncül Listesi (Ona) : (Ali, Ayşe) (4.11)

Bu aşamada 4.2 nolu tabloda gösterildiği gibi iki seçenekli senaryo oluşur.

Tablo 4.2: Senaryolar

	$\emptyset$	Ona
Senaryo 1	Ali	Ayşe
Senaryo 2	Ayşe	Ali

Bu aşamada hangi senaryonun Merkezleme Kuramı açısından en iyi geçişi verdiğine bakılmalıdır. Birinci senaryo seçildiğinde 3.1 nolu tabloda belirtilen geçişlerden DEVAM geçişi seçilmiş olacaktır. Eğer ikinci senaryo seçilirse bu durumda geçişin ismi YUMUŞAK-GEÇİŞ olacaktır. DEVAM geçişi YUMUŞAK-GEÇİŞ'e tercih edildiğinden bu aşamada birinci senaryo tercih edilecektir. Sonuç olarak $\emptyset$ boş artgönderimi **Ayşe**'ye, **ona** açık adlı **Ali**'ye gönderimde bulunmaktadır.

T_3 için bu algoritma tekrar çalıştırıldında bu tümcedeki **onları** artgönderimi Ali ve Ayşe öğeleriyle sayı/kişi uyumu sebebiyle eşleşemez. Ardından bir tek **kitaplar** söylem varlığı kalır ve bu öge **onlar** adılının öncülü olarak seçilir.

Merkezeleme Yaklaşımın Sonuçları

Yukarıda tanımlanan algortimayı tüm aşamalarıyla Java diliyle yazılmış ve derlem bölümünde bahsedilen 10.000 kapasiteli derlem üzerinde denenmiştir. Uygulama sonucunda derlemdeki 1114 adılın 792'si doğru bir şekilde çözümlendi. [Mitkov, 2002]'un tanımladığı 4.12 nolu formülü kullanırsak

$$\text{Başarı Oranı} = \frac{\text{Doğru Çözülmüş Adıl Sayısı}}{\text{Toplam Adıl Sayısı}} \quad (4.12)$$

$$\text{Başarı Oranı} = \frac{792}{1114} = .711 \quad (4.13)$$

gibi bir başarı oranı elde edilir. Bilgisayarlı Artgönderim için yapılan çalışmaların sayısı çok az olduğundan ve ortak bir derlem olmadığından ortaya çıkan sonuçların doğru bir şekilde karşılaştırılmasının imkanı yoktur. Ama yine de Türkçe için yapılan diğer bilgisayarlı çalışmalarla basit anlamda bir karşılaştırma yapılabilir. Bununla ilgili son çalışmalardan iki tanesi ile bir karşılaştırma yapılabilir. [Tüfekci ve Kılıçaslan, 2007] Hobbs'un algoritmasını kullanarak %84 gibi bir başarı yakaladığını belirtmiştir. Öte yandan [Küçük ve Turhan, 2007] az bilgili sistemleri üzerinde iki farklı deney yapmıştır. Ve bu deneylerin sonucunda %85.3 ve %73.7 gibi iki farklı sonuç elde etmişlerdir. Bu sonuçlara bakmaktaki

temel amaç sayısal bir karşılaştırma değildir. Türkçe için genel resmi görmek içindir. Zaten bu tez çalışmasının sonunda tüm diller için yapılan çalışmaların sonuçlarına bakılacaktır.

Merkezeleme Algoritması çerçevesinde yapılan bir diğer değişiklik ise 4.7 nolu formüldeki değişikliktir. Çok iyi bilinen bir gerçek vardır ki adılsı öğeler söylem içinde diğer söylem varlıklarına göre daha belirgindir. Bu nedenden 4.7 nolu sıralamayı sadece bu bilgiyi ekleyerek 4.14 nolu formül halinde yeniden ifade edildi ve uygulandı.

$$\begin{aligned} \text{Yalin[adıl]} &> \text{Belirtme[+adıl]} > \text{İlgi[+adıl]} > \\ \text{Yönelme[+adıl]} &> \text{Çıkma [+adıl]} > \text{Bulunma[+adıl]} > \\ \text{Belirtme[-adıl]} &> \text{İlgi[-adıl]} > \text{Yönelme[-adıl]} > \\ \text{Çıkma[-adıl]} &> \text{Yönelme[-adıl]} \end{aligned} \quad (4.14)$$

Bunun ardından başarı oranında yüzde 1 oranında küçük bir artışı oldu. Bu sonuçtan sonra adılsı söylem varlıklarının çok az da olsa daha belirgin olduğu Türkçe içinde sözkonusu olduğu söylenebilir. Bu bilgi daha sonra yapılacak çalışmalar açısından önemlidir.

4.3.2 Belirginlik Tabanlı Çözüm Önerisi

Artgönderim çalışmalarının en önemli yaklaşımlarından biri [Lappin ve Leass, 1994] tarafından önerilmiş RAP algoritmasıdır. Bu yaklaşıma benzer bir sistem ise [Mitkov, 2002]'un sunduğu MARS sistemidir. Bu yöntemler sadece İngilizce için değil Lehçe, Arapça, ve Norveççe için de başarılı bir şekilde uygulanmıştır. İn-

gilizce dışındaki diğer diller için, bu yöntemler basit bir değişiklik ile uygulanabilmiştir. Türkçe için de basit bir değişiklikle bu yöntem bu tez kapsamında uygulanmıştır. Bu yöntem belirginlik temmellidir ve sözdizimsel bilgi kullanması açısından az bilgili sistemdir. Yöntem aday öncüllere, özelliklerine bakarak belli puanlar verir. En çok puanı alan öncül asıl öncül olarak seçilir.

Dolayısıyla bu noktadaki en temel soru puanlamanın nasıl belirleneceğidir. Örneğin İngilizce için düşünüldüğünde dilbilgisel işlev özelliği, en önemli özelliktir ve bu yüzden en çok puanı bu özellik alır. Ancak her dilde bu böyle olmayabilir. Örnek olarak [Holen, 2007] Norveççe için yaptığı çalışmada özne pozisyonunun belirginlik için önemli olmadığını vurgulamıştır. Türkçe için baktığımızda bu özelliğin önemli olduğu söylenebilir. Dolayısıyla bu tez için oluşturulan sistemde **Özne** dilbilgisel görevine sahip öğeler nesne konumunda olan öğelere göre daha çok puan alacaklardır. Sayı, kişi ve cinsiyet ise diğer sözdizimsel özelliklerdir. Söz konusu adıl ile o adıyla sayı ve kişi uyumu gösteren öncüller diğer öncüllere göre daha çok puan alır. Öte yandan İngilizce için kullanılan cinsiyet kodlaması Türkçe’de olmadığından bu bilgi kullanılmayacaktır. Tüm bu belirginlik etkenleri aşağıdaki şekilde sıralanabilir:

- **Özne Etkeni:** Bu etken daha önce de bahsedildiği gibi en önemli unsurlardan biridir ve Türkçe için bu durum geçerlidir. Merkezleme Teorisi tabanlı yaklaşımda da görüldüğü gibi, özne pozisyonunda bulunan merkezin değişmeme eğilimi vardır. Özne pozisyonundaki öğeler çoğu zaman boşartgönderim halinde sürekli kendilerini var ederler.

Ali sabah okula gitmek için evden ayrıldı.

İlk olarak kalabalık bir yoldan karşıya geçti.

Okula vardığında arkadaşlarının onu beklediğini gördü...

(4.15)

Yukarıdaki söylemde görüldüğü gibi özne pozisyonundaki söylem varlığı **Ali**, özne pozisyonundaki boşartgönderimlerle sürekli tekrarlanır. Bazen uzun hikayelerde bu durum çok daha uzun olur ve artgönderim çözümü de o derece basit olabilir. Bu anlamda özne pozisyonundaki öğelerin boşartgönderimlerle sürekli tekrarlanması ya da gönderimsel ilişki içinde olmaları Türkçe için oldukça sık görülen bir durumdur.

Türkçe metinlerde özne değişikliği başka bir söylem oluşturmak için kullanılır. Her zaman bu kural geçerli olmasa da çoğunlukla yeni bir söylemin sınırlarının işareti olduğu söylenebilir.

- **Nesne Etkeni:** Eğer öncül dolaysız nesne pozisyonunda ise belli bir puan alabilir. Ancak Özne Etkenine oranla daha az puanlanır.
- **Kişi/Sayı Uyumu:** Bu özellik en önemli özelliklerden biridir. Çünkü bu uyum içinde olmayanlar genellikle gönderimsel bir ilişki içinde bulunmazlar. Genellikle demek zorundayız çünkü bunun kuraldışı durumları da vardır. Örnek olarak **aile, takım, gurup** gibi çoğul eki almayan sözcükler **biz, siz, onlar** çoğul artgönderimlerle ifade edilebilirler. Ya da iki söylem varlığı sisteme dahil olmuşsa bunlar daha sonra çoğul bir artgönderimle ifade edilebilirler.

Ali ve Ayşe sabah erken kalkarlar.

Çünkü onlar günlerini daha iyi geçirmek isterler.

(4.16)

Bu örnek görüldüğü gibi Ali ve Ayşe söylem varlıkları iki ayrı ve tekil ifadelerdir. Ardından ikisine ortak bir çoğul artgönderim olan **onlar** gönderimde bulunur.

- **Adılsı Vurgusu:** Eğer öncülün kelime tipi adıl ise bu öncül adıl olmayanlara göre daha çok vurgu alır. Bu durum birçok çalışmada vurgulanmıştır.

Ali gazete okumayı sever.

$\odot_1$ Sabah erkenden g azete almak için kalkar.

$\odot_2$ Kahvaltıda gazete okumak onun için zevktir.

(4.17)

Üçüncü tümcedeki $\odot_2$ boşartgönderimi için bir önceki tümcede yine bir ($\odot_1$) boşartgönderim ile gazete aday öncüdür. ($\odot_1$) boşartgönderimi hem özne hem de adılsı olduğundan gazeteye göre daha belirgindir.

- **İlk Ad Öbeği Vurgusu:** Eğer bir öge söyleme ilk ad öbeği olarak girmiş ise sonraki her tekrarlanışında ilk ad öbeği vurgusu alır ve puanlanır. Örnek olarak 4.17 nolu söylemde Ali söylem varlığı sonra tekrarlanırsa ya da bir artgönderimle gönderimsel ilişki içinde olursa o artgönderim **ilk ad öbeği** vurgusu alır. Bu etken [Mitkov, 2002], [Lappin ve Leass, 1994] ve [Holen, 2007] tarafından kullanılmış ve önerilmiştir. Söylemin davranışına ilişkin bir ipucundan hareket ile bu İlk Ad Öbeği etkisi vurgulanmıştır. Söylemler ilk ad öbeğini sürekli vurgulamak isterler. Bu özellikten dolayı bu ad öbekleri sürekli adılsallaştırılırlar.
- **Bahsedilme Sayısı:** Bir öge söylem içinde ne kadar tekrarlanırsa o ögenin belirginliği o kadar artar. Bu özellik birçok çalışma tarafından da kullanılmıştır. Bir önceki **İlk Ad Öbeği** etkeni gibi bu etken de benzer bir söylem özelliğinden gelmektedir. Sürekli tekrarlanan ögeler daha fazla vurgu alırlar.
- **Uzaklık:** Bu etken sözkonusu adıl ile söz konusu aday öncül arasındaki uzaklığın tümce bazındaki karşılığıdır. Uzaklık arttıkça ögeler arasında

gönderimsel ilişki olma olasılığı azalır. Bu çalışmadaki derlem analiz edildiğinde bu özelliğin çok önemli olduğu tesbit edildi. Çünkü adılların öncülleri daha çok bir önceki ve iki önceki tümce içerisinde yer almıştır. Dolayısı ile uzaklığın bu çalışma için önemli olduğu açıktır.

Tüm bu etkenler belli bir ağırlık ile ağırlıklandırıldı. Bu aşamada deneme yanılma yöntemiyle birçok değer tahmini olarak atandı.

Tablo 4.3: Etkenler ve Ağırlıkları

Etken	Ağırlığı
Özne Vurgusu	100
Kişi Sayı Uyumu	150
Adıl Olma Vurgusu	50
Dolaysız Nesne Vurgusu	30
İlk Ad Öbeği Olma Vurgusu	20
Bahsedilme Sayısı Vurgusu	2 kere için 10 daha fazlası için 20

Yapılan bu denemelerde en iyi ağırlıklandırma değerleri tablo 4.3.2'deki gibi olmuştur. Söylemdeki herbir adıl ondan önce gelen herbir öncül ile eşleşir ve yukarıdaki ağırlıklar tüm bu öncüllere uygulanır. Uygulama sonunda ağırlıklar toplanır ve adıla aday her bir öncül için bir puan oluşur. En iyi ağırlığa sahip öncül başarılı aday olarak belirlenir ve söz konusu adılın gönderimde bulunduğu öncül olur.

Basit Bir Örnek

Bu algoritmayı basit bir örnekle açıklayalım. Kabul edelimki aşağıdaki gibi iki tümcelik basit bir söylem ile çalışıyoruz.

Ali çantasındaki kalemli arıyordu.

$$\oslash \text{ Ama kalemeleri bir türlü bulamıyordu.} \quad (4.18)$$

İkinci tümcedeki $\oslash$ boş artgönderimi için olası öncüller kümesi { Ali, çantasındaki , kalemeleri} gibidir. Tüm bu aday öncüller yukarıdaki tablo kullanılarak ağırlıklandırılır. Ağırlıklandırma sonuçları 4.3.2 nolu tablodaki gibidir.

Tablo 4.4: Aday Öncüllerin Ağırlıklandırılmaları

Etken	Ağırlığı	Ali	Çantasındaki	Kalemeleri
Özne Vurgusu	100	100	0	0
Kişi Sayı Uyumu	150	150	150	0
Adıl Olma Vurgusu	50	0	0	0
Dolaysız Nesne Vurgusu	30	0	30	30
İlk Ad Öbeği Olma Vurgusu	20	20	0	0
Bahsedilme Sayısı Vurgusu	10 / 20	0	0	0
Toplam		270	180	30

Ali öncülü 270 puanla en çok oy alarak ikinci tümcedeki $\oslash$ boş artgönderimin öncülü olarak kabul edilir. Bu örnek iki tümceden oluştuğu için birçok farklı durumu içinde barındıramıyor. Örnek olarak bashedilme sayısı özelliğini örnekleyemedik.

Bu algoritmayı derlem üzerinde uyguladığımızda, 1114 adılın 598'i doğru bir şekilde çözümlendi. [Mitkov, 2002] artgönderim sisteminin başarı oranını 4.12 nolu formüldeki gibi tanımlamıştır.

Bu formüle göre bizim modelin başarı oranı $\frac{598}{1114} = 0.53$ 'e eşittir. Bu orana baktığında ve de diğer artgönderim çalışmalarıyla karşılaştırıldığında normal bir çözüm yönteminden daha düşük performansa sahip olduğu söylenebilir. Bu yaklaşım Merkezleme Teorisine göre daha fazla bilgiye ihtiyaç duyar ve bu bakımdan yöntem bilgiye fazla bağımlıdır. Yöntemde kullanılan özelliklerin sisteme ne kadar katkısı olduğu her zaman tartışma konusudur. Seçilen özelliklerin bir çoğu gereksiz ve yararsız olabilir. Bu bakımdan sistemin bu gereksiz özellikleri çözüm-

leme öncesi (pre-processing) süreçlerde boşu boşuna aramaması için bir eleme işlemine tabi tutulabilir. Bu noktadaki sorulacak soru şu: Çalışma alanı içinde bu özelliklerin yararı nasıl sayısallaştırılabilir ? Ya da sayısallaştırılabilir mi ? Bunun için [Mitkov, 2002] (Relative Importance) Bağlı Önem isimli bir formül önermiştir. Bağlı Önem **K** etkeni için 4.19 nolu formüldeki gibi hesaplanır.

$$RI_K = \frac{SR - SR_K}{SR} \quad (4.19)$$

Bu formüldeki SR kısaltması tüm sistemin başarı oranıdır. Öte yandan SR_K kısaltması ise K etkeni olmadığında sistemin göstermiş olduğu başarı oranını ifade eder. Görüldüğü gibi RI_K formülü K etkeni olmadığı durumda sistemin bundan nasıl etkileneceğini bağlı bir şekilde hesap eder. Bu formül buradaki etkenler için uygulandığında 4.5 nolu tablodaki sonuçlar elde edilir.

Tablo 4.5: Bağlı Önem Tablosu

Etken	RI
Sayı / Kişi	0.07
Adılsı Vurgusu	0.01
Özne Vurgusu	0.28
Nesne Vurgusu	0.01
İlk Ad Öbeği Vurgusu	0.01
Tekrar Vurgusu	0.01

Görüldüğü gibi sayı/kişi uyumu etkeni ile özne vurgusu Türkçe’de önemli bir yere sahip. Bu etkenleri sistemden çıkardığımızda neredeyse artgönderim çözümlemesi imkansız hale gelecektir. Bu **Bağlı Önem** ölçütüne benzer başka ölçüm yöntemleri de mevcuttur. Çalışmanın sonuç kısmında tüm bu ölçüm sonuçlarını karşılaştıracamız.

Bir başka ölçüm tekniği yine [Mitkov, 2002] tarafından önerilen Karar Gücüdür (Decision Power). Bir etkenin karar gücü hesabı (K), onu sağlayan öncül adayların asıl öncül olarak seçilip seçilmemesine bakılarak yapılır. O etkenin asıl öncül olmasındaki rolüne bakılmaksızın yapılan bu değerlendirmede sadece o aday öncüllerin asıl aday olup olmadıklarına bakılır. K etkeninin Karar gücü, bu etkenin desteklediği ve doğru şekilde asıl öncül olmuş durumların sayısının (S_K) K etkeni tarafından desteklenmiş tüm öncüllerin sayısına (A_K) oranına eşittir.

$$DP_K = \frac{S_K}{A_K} \quad (4.20)$$

Sistemdeki etkenlerin Karar Gücüne bakıldığında 4.6 nolu tablodaki gibi bir sonuç elde edilir.

Tablo 4.6: Karar Gücü Tablosu

Etken	DP
Sayı / Kişi	0.107
Adılsı Vurgusu	0.09
Özne Vurgusu	0.16
Nesne Vurgusu	0.03
İlk Ad Öbeği Vurgusu	0.17
Tekrar Vurgusu	0.06

Yukarıdaki tabloda en önemli üç etken; **sayı/kişi uyumu**, **özne vurgusu** ve **ilk ad öbeği vurgusu**. Bu özellikler diğer algoritmalar içinde önemli olduğunu izleyen bölümlerde göreceğiz.

4.4 Öğrenme Tabanlı Çözümler

Bu bölümde Türkçe için öğrenme tabanlı bir çözüm tartışılacaktır. Artgönderim sorununu makine öğrenmesi teknikleriyle nasıl çözüleceği bölüm 3.3.1’te kısaca bahsedilmiştir. Türkçe için benzer bir yöntem uygulanacaktır.

4.4.1 Eşleşme

Makine öğrenmesindeki en önemli unsur kullanılan verinin niteliğidir. Artgönderim çözümlemesi için oluşturulacak olan çiftlerin seçimi bu anlamda önemlidir. Tablo 3.3 basit anlamda bu çiftlerin nasıl oluşturulduğunu göstermektedir. Her bir gönderge ile o göndergenin öncülü olabilecek, ondan önce gelen tüm gönderimsel ifadeler bir çift-eşleşme kümesi oluşturur. Bu eşleşme kümesindeki öğeler olumlu(EVET) ve olumsuz(HAYIR) olarak işaretlenir. Ve böylelikle makine öğrenme algoritması eşleşmenin kurallarını çıkarmaya çalışır. Bizim veritabanımızdaki en geniş söylem 26 cümleden oluşur. Fakat söylem listemizdeki adılar en çok 10 tümce öncesine gönderimde bulunduğundan sadece söz konusu adıl ile ondan önceki 10 tümce arasındaki öğelerle bir eşleşme oluşturulmaktadır.

Bu aşamadaki en büyük sorun olumsuz örnek sayısının olumlu örnek sayısından oldukça fazla olmasıdır. Çünkü hedef sınıftaki bu değerler birbirlerine sayıca ne kadar yakın olursa, makine öğrenme algoritmaları o kadar başarılı olur. Bu anlamda olumsuz örneklerin bir şekilde azaltılması ve olumlu örneklerle dengelenmesi gerekmektedir. Bunun için çeşitli eleme yöntemleri kullanılabilir. Bu çalışmada kişi ve sayı uyumu süzgeçi kullanılarak bir eleme gerçekleştirildi. Bu ayıklama aşamasından hemen sonra makine öğrenme algoritmaları uygulanabilir. Bu deneyde kullanılan algoritmalar bi sonraki bölümde anlatılacaktır.

4.4.2 Kullanılan Veri Madenciliği Teknikleri

Makine öğrenmesi ile bu çalışmada yapılan tüm deneyler Weka Yazılımı ² ile gerçekleştirilmiştir. Bu deneyde değerlendirme yöntemi için 10'lu çapraz geçirme (10-fold cross validation) yöntemi kullanılarak 5 farklı makine öğrenme algoritması derlem üzerinde sınılandı. Bu algoritmalar şunlardır: Naïve Bayes Algoritması, Örnek tabanlı yöntem ya da bellek tabanlı yöntem (K-en yakın komşu yöntemi), Karar Ağaçları, Destek Vektör Sistemleri ve Yapay Sinir Ağları (Algılayıcılar).

Naïve Bayes Algoritması

Her özellik birbirinden bağımsızdır kabulüne dayanan bir yöntemdir. $(x_1, ..., x_n)$ özelliklerine sahip bir X örneği verildiğinde bu algoritma, bu örnek için $P(X | C)$ olasılığını en yüksek yapan C sınıfını seçer.

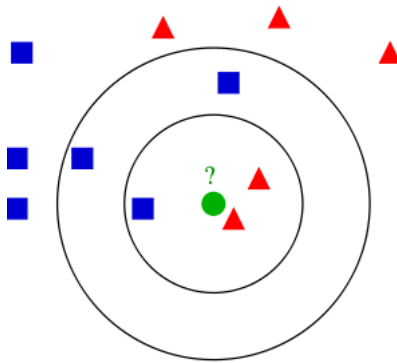
$$P(X|C) = P(x_1, ..., x_n|C) \quad (4.21)$$

Bu algoritma hesaplama maliyetini azaltıyor. Çünkü sadece sınıf dağılımlarına bakıyor. Gerçek uygulamalarda özelliklerin birbirlerine bağlı olması bu algoritmanın bir eksiği olarak görülmektedir.

²Weka (Waikato Environment for Knowledge Analysis) Java programlama dili ile yazılmış en yaygın veri madenciliği aracıdır. Waikato Üniversitesi tarafından oluşturulan WEKA açık kaynak bir yazılımdır ve GPL (GNU General Public License) lisansına sahiptir.

K-En Yakın Komşu Sınıflandırıcısı

Bu algoritma örnekleme tabanlı yöntemlerden biridir. Bu tip bir algoritmanın en belirgin özelliği kullandığı veriden bir model oluşturmaması ve bu veriyi saklamasıdır. Her bir yeni örnek saklanan verideki örneklerle kıyaslanır. Ve yeni örneğe benzeyen k adet eski örneklerin sınıflarına bakılarak bir karara varılır. Bu nedenle algoritmanın öğrenme sürecinden bahsedilemez. Karar verme süreci ise diğer algoritmalara göre daha uzundur. Temel yaklaşımda sınıflandırılacak örneğe en yakın k adet örnek bulunur. Bu k adet örnekteki baskın sınıf o örneğin sınıfı olarak atanır. Bütün örnekler n-boyutlu bir uzayda bir noktaya karşı düşürülür. Örnek olarak 4.3 nolu şekilde ortadaki daire yeni bir örneği temsil eder. Bu örnek ya kare ya da üçgen olarak sınıflandırılacaktır. Peki bu dairenin aslında kare mi yoksa üçgen mi olduğu nasıl tespit edilecek ? Eğer knn algoritmasında k değeri 3 seçilirse şekilde de görüldüğü gibi içteki çember içinde kalan örneklere bakılır. Çünkü içteki çember 3 eski örneği çevreliyor. Bu örneklerden ikisi üçgen biri kare olduğundan knn algoritmasının kararı **üçgen** olur. Eğer k değeri 5 seçilirse 5 eski örneği çevreleyen dıştaki çembere bakılır. Bu çember içinde kalan örneklerin üç tanesi kare iki tanesi üçgen olduğundan knn algoritmasının kararı baskın sınıf olan **kare** olur.



Şekil 4.3: K-en yakın komşu algoritması

Nesneler arasındaki uzaklık belli bir uzaklık formülüne (Euclidean, Manhat-

tan, ...) göre hesaplanır. Örnek olarak $X(x_1, ..., x_n)$ ve $Y(y_1, ..., y_n)$ iki noktayı ele alalım. Bu iki nokta arasındaki uzaklık 4.22 nolu formül ile hesaplanır.

$$D = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2} \quad (4.22)$$

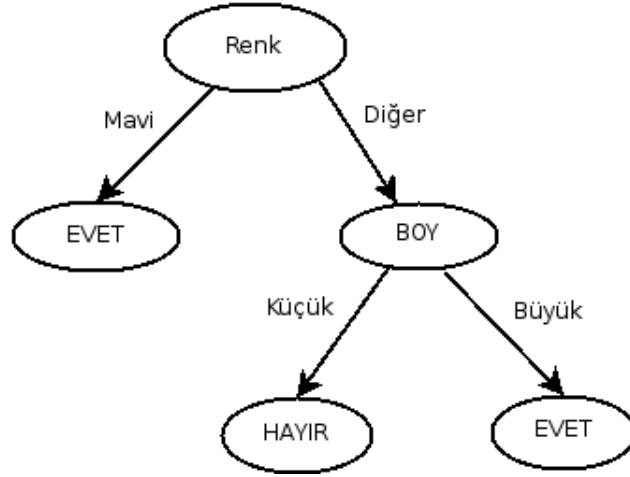
Daha öncede bahsesildiği gibi makine öğrenmesi deneylerinin öğrenme aşamalarında WEKA yardımcı kütüphanesini kullandı. Knn algoritması için WEKA'daki IBk isimli algoritma kullanılmıştır.

Karar Ağaçları

Karar ağacı algoritmaları veriden bir ağaç modeli ortaya çıkar. Bu ağaç modeli veriyi doğru bir şekilde ifade etmeye çalışır. Karar ağaçlarının en önemli özelliği kurallarının insanlar tarafından rahatlıkla okunabilmesi ve yorumlanabilmesidir. Peki bu ağaçtaki düğümler, dallar ve yapraklar ne anlama gelir ? Karar Ağaçlarında düğümler özelliklere ilişkin soruları ifade eder. 4.4 nolu şekilde görüldüğü dallar ise o özelliğin değerlerine karşılık gelir. Bir anlamda düğüm özelliğe ilişkin soru ve dal ise bu sorunun cevabıdır. Soru-cevap şeklinde ağaç üzerinde gezindikten sonra bir yaprağa ulaşılır. Bu yaprak ise modelin kararıdır. Ya da hangi sınıfın seçildiğini ifade eder.

Şekildeki ağac bir maddenin iyi olup olmadığına karar veren bir karar ağacı olarak düşünülebilir. Bu ağaç eğer şekil maviyse başka bir soru sormadan olumlu cevap veriyor. Eğer şekil mavi değilse bu durumda başka bir soru sorma ihtiyacı duyuyor. Eğer maddenin boyu büyük ise olumlu, küçük ise olumsuz karar üretiyor.

Karar ağacı algoritmalarında en önemli ilerleme [Quinlan, 1993.] tarafından



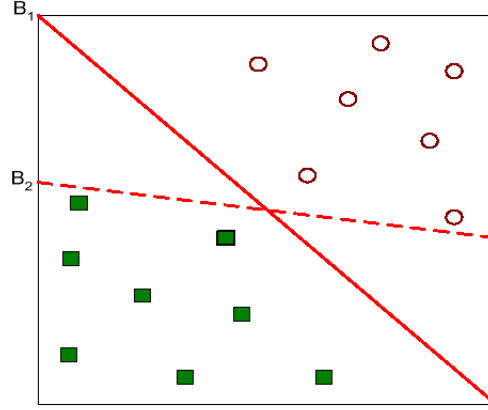
Şekil 4.4: Karar Ağacı

kaydedilmiştir. Entropi formülü sayesinde ağaç oluşturulurken özellikler önem sırasına göre dizilir. En önemli yada bilgi kazancı (information-gain) en yüksek olan özellik ilk düğüme yerleştirilir. Böylelikle ağaç hem daha küçük hem de daha başarılı bir model oluşturur. İlk olarak ayrık özellikler için geliştirilen ID3 isimli bu algoritma ardından sürekli (reel) özellikler içinde uygulanabilmiş ve C4.5 ismini almıştır. WEKA yazılımı içerisinde bu algoritmaya J4.8 ismi verilmiştir. Bu çalışmada karar ağacı için ayrıca toplu öğrenme (boosting) kullanılmıştır. Bu yöntemde tek bir model oluşturmak yerine çeşitli ayar değişikliği ve yöntem yardımıyla birden fazla ağaç modeli oluşturulur. Ve ardından yeni bir örnek sisteme geldiğinde, bu örnek tek bir sınıflandırıcı yerine birden fazla sınıflandırıcı tarafından işleme tabi tutulur. Modellerden gelen baskın sınıf yeni örneğin sınıfı olur. Buradaki amaç toplu öğrenmedeki başarı oranını artırmaktır. Ancak bu yapının her zaman için başarılı olduğu söz konusu değildir.

Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machines) veriyi doğrusal olarak ayırarak bir karar vermeye çalışır. Bir çok hiper düzlem kullanarak olumlu ve

olumsuz örnekleri birbirlerinden ayırmaya çalışır. Olumlu, olumsuz örnekleri birbirinden en iyi ayıran hiper düzlem sonraki kararlar için kullanılır.



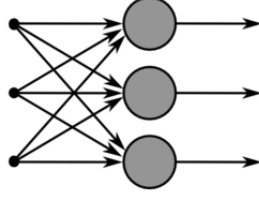
Şekil 4.5: Destek Vektör Makineleri - KDM

Örnek olarak 4.5 nolu şekilde B_1 ve B_2 isimli iki doğru, sınıfları birbirinden ayırmaya çalışıyor. En uygun hiper-düzlem algoritması [Vapnik ve Lerner, 1963] tarafından önerilmiş olan doğrusal sınıflandırıcıdır. Öte yandan KDM sadece doğrusal değil doğrusal olmayan bir sınıflandırma da uygulayabilmektedir. [Boser vd., 1992] tarafından sunulan bu yöntem doğrusal olmayan bir çekirdek fonksiyonu $K(x_i, y_i)$ içermektedir.

Bu tez çalışmasında doğrusal ve radyal temelli çekirdek fonksiyonu kullanıldı. Bu algoritmalar WEKA yazılımında LIBSVM kütüphanesiyle gelen C-SVC olarak yerleştirilir.

Algılayıcılar (Voted Perceptron)

Algılayıcılar (Voted Perceptron) bir yapay sinir ağı uygulamasıdır. Bu yöntem orjinal yöntemin ağırlıklandırma sürecinde bir değişiklik yapar. Orjinal yöntem $w \cdot x + b > 0$ formülünü kullanır. Bu formülde w ağırlıklandırma vektörü ve x ise giriş vektörüdür ya da bir başka ifade ile özellik vektörüdür. b ise sabit



Şekil 4.6: Yapay Sinir Ağları

bir katsayıdır. Bu formül doğru olursa çıktı 1, yanlış ise çıktı 0 olur. Orjinal algılayıcı eğitim sürecinde, bir sınıflandırma eğitimi için, özellik vektörü peş peşe işleme sokulur. Eğer belli bir eşik değerinin üzerinde bir hata oluşursa, ağırlıklar güncellenir. Yapay Sinir Ağları 4.6 nolu şekilde görüldüğü gibi birden çok algılayıcıları iç içe kullanır.

Ancak orjinal algoritma eğer veri doğrusal bir şekilde ayrılmaya uygunsa başarılı bir sonuca ulaşır. Eğer verinin doğrusal olmayan bir karakteri var ise, diğer bir algılayıcı olan **voted perceptron** kullanılabilir. Bu algoritma orjinal süreçte oluşan tüm ağırlıklandırmaları kullanarak tahminde bulunmaya çalışır. Bunu her bir ağırlıklandırma vektörünün kaç eğitim örneğini doğru bir şekilde sınıflandırdığını bellekte tutarak başarabilir.

4.4.3 Sonuçlar

Ele alınan bu beş algoritma, önceki bölümlerde bahsedildiği gibi derlem üzerine uygulanmıştır. Her bir algoritmanın farklı ifade gücüne sahip iki farklı sınıflandırıcısı kullanıldı. Bu deneyde kullanılan parametreler 4.7 nolu tablodaki gibidir. Bu tablodaki ayar isimleri WEKA programında kullanıldığı gibi verilmiştir.

Bu beş algoritmanın iki farklı sınıflandırıcısının uygulanması sonucu ortaya çıkan sonuçlar 4.8 nolu tablodaki gibidir.

Şimdi bu tablolarda kullanılan performance ölçüm teknik ifadelerden kısaca

Tablo 4.7: Algoritma ve Parametreleri

Algoritma	Parametreler
IBk	No Distance Weighting Use Euclidean Distance
Naïve Bayes	No Supervised Discretization
J48	Confidence Factor: 0.25 No Reduced Error Prunning Use Subtree Raising On Pruning No Laplace-based Smoothing Do not Force Binary Split on Nominal Attributes
Voted Perceptron	Maximum Number of Alternation:10000
C-SVC	Cost: 1.0 Gamma: $1/\max_i \text{index}$ No Normalization

Tablo 4.8: Sınıflandırıcıların Sonuçları

No	Sınıflandırıcı	Parametre Değişikliği	Sınıflandırma Sonuçları				
			Doğruluk	Precision	Recall	F-Ölçütü	Kappa
	Baseline / Dayanak-1	Çoğunluk tabanlı	67.3	-	-	-	0
	Baseline / Dayanak-2	Sınıf Dağılım tabanlı	55.2	31.5	31.5	31.5	0
	Baseline / Dayanak-3	En Yakın Aday	-	-	46.7	-	-
1	IBk	k=1	0.75	0.60	0.76	0.67	0.4792
2	C-SVC	Doğrusal Çekirdek	0.77	0.66	0.62	0.64	0.4736
3	Naïve Bayes	Normal Dağılım	0.78	0.65	0.71	0.68	0.5202
4	J48	boosted ve budamasız	0.77	0.64	0.70	0.67	0.4968
5	Voted Perceptron	üs= 1	0.79	0.71	0.65	0.68	0.5155
6	Naïve Bayes	kernel estimator	0.80	0.69	0.71	0.70	0.5402
7	IBk	k=11	0.80	0.67	0.76	0.72	0.5534
8	J48	Budamalı	0.81	0.71	0.71	0.71	0.5712
9	Voted Perceptron	Üst = 2	0.81	0.72	0.70	0.71	0.5739
10	C-SVC	Radyal Tabanlı Çekirdek	0.82	0.73	0.74	0.74	0.6017

bahsedilebilir. Tablodaki Doğruluk oranı (Accuracy)

$$\frac{(TP + TN)}{(TP + TN + FN + FP)} \quad (4.23)$$

formülüyle, kesinlik (precision)

$$\frac{TP}{(TP + FP)} \quad (4.24)$$

formülüyle, anma (recall)

$$\frac{TP}{(TP + FN)} \quad (4.25)$$

formülüyle ve F-ölçütü

$$\frac{(2 * kesinlik * anma)}{kesinlik + anma} \quad (4.26)$$

formülüyle hesaplanır. Bu formüllerde geçen TP, TN, FP, ve FN ifadeleri 4.9 nolu tabloya göre açıklanabilir.

Tablo 4.9: Karışıklık Matrisi

		Tahmin	
		Pozitif	Negatif
Gerçek	pozitif	TP	FN
	Negatif	FP	TN

Çalışma alanında varolan bir öğeye var dendiğinde ya da gerçekte doğru olan bir şeyi tasarlanan sistem de doğru olarak tahmin etmiş ise bu durum TP (True Pozitif) olarak ifade edilir. Gerçekte doğru olan bir şey yanlış olarak tahmin edilirse bu FN (False Negative) olarak ifade edilir. Gerçekte yanlış olan bir şeyin doğru olduğu söylenirse, bu durum FP (False Pozitive) olarak ifade

edilir. Son olarak gerçekte yanlış olan bir duruma sistem yanlış derse, bu durum TN (True Negative) olarak isimlendirilir.

Şimdi bu performans değerlendirme ölçümlerinin artgönderim çözümlemesi için ne ifade ettiğine bakılacak.

Doğruluk (Accuracy),

$$\frac{\text{Sistem tarafından doğru tahmin edilmiş tüm durumlar}}{\text{Sistemdeki tüm durumlar}} \quad (4.27)$$

formülüyle, kesinlik (Precision)

$$\frac{\text{Sistem tarafından doğru şekilde çözülmüş artgönderim ilişkileri}}{\text{Sistem tarafından ifade dılmış toplam artgönderim ilişkiler}} \quad (4.28)$$

formülüyle ve anma (Recall) ise

$$\frac{\text{Sistem tarafından doğru şekilde çözülmüş artgönderim ilişkileri}}{\text{Derlemde bulunan toplan artgönderimsel ilişkiler}} \quad (4.29)$$

formülüyle hesaplanır edilebilir.

4.8 nolu tablodaki değerlere bakıldığında, tüm sonuçlar belli bir aralıkta ve birbirine yakındır. Çok net bir şekilde belli algoritmaların diğerlerine göre daha başarılı olduğu söylenemez. Ancak yine de sınıflandırıcıları, genel performansları açısından değerlendirildiğinde olumlu bir sonuç olduğu söylenebilir. En azından bu alanda yapılmış diğer çalışmalarla kıyaslandığında bu daha net söylenebilir. Bunun söylenmesindeki en önemli dayanak ise, tabloda da gösterilen ilk üç dayanak algoritmalarıdır. Dayanak (baseline) algoritmaları elde edilen sonuç için bir temel oluşturur. Bu algoritmalar ile önerilen algoritma kıyaslanır ve ortaya bir başarı oranı çıkar. Bu çalışmadaki deneyde üç dayanak algoritması kullanıldı: Çoğunluk tabanlı, Sınıf dağılım tabalı ve en yakın aday. Birinci dayanak algorit-

ması, çoğunluk tabanlı algoritma, tüm yeni örneklerle eğitim kümesinde baskın olan sınıfı atar. Bu tez çalışması alanında baskın sınıf değeri **hayır** etiketlidir. Yani iki öge arasında bir ilişki var mı sorusunun cevabı hep **hayır**dır. İkinci dayanak algoritması Sınıf Dağılımı üzerine kuruludur. Test kümesindeki sınıflandırılacak yeni örneklerle yapılacak rastgele sınıf etiketlerinin dağılımı eğitim kümesindeki dağılımla aynı tutulur. Üçüncü dayanak algoritması söz konusu adıldan önceki en yakın adayı öncül olarak kabul eder. Buradaki yöntem basit bir ilkeyi temel alır. Artgönderimler onlardan önce gelen ve en yakın ögeye gönderimde bulunmaya eğilimlidirler. Yapılan çalışmaların bu dayanak algoritmalarından ne kadar başarılı olduğu önemlidir. Bunlardan birine yakın bir sonuç veren sistemin bir yatarıcılığı yoktur. Çünkü basit bir bilgi ile kurulmuş dayanak algoritmalarıyla aynı sonucu vermesi istenilen bir durum değildir.

Bu noktada rahatlıkla belirtilmesi gereken bir nokta şudur: Bu çalışmada sunulan modeller ve kullanılan makine öğrenme algoritmaları, bu dayanak algoritmalarından çok daha iyi bir performans göstermiştir. Bu performansı daha iyi ifade etmek için makine öğrenmesi modeliyle dayanak algoritması arasındaki farka bakılabilir.

Ele alınan ve uygulanan diğer bir yöntem **Kappa İstatistiği**dir. Tablo 4.8’de sunulan sonuçlarda bu istatistiksel bilgi mevcuttur. Kappa istatistiği, sınıflandırıcıların performansını bir temel çerçevesinde ölçmek için kullanılan alternatif bir yöntemdir. Ölçüm formülü

$$K = \frac{P_0 - P_c}{(1 - P_c)} \quad (4.30)$$

şeklinde ifade edilir. Buradaki P_0 ifadesi sınıflandırıcının doğruluk oranı, P_c

ifadesi kararın rastgele verilmesiyle ortaya çıkan başarı oranıdır. Bu formül-
den de anlaşıldığı gibi bu istatistiksel ölçüm değeri -1 ile 1 arasında yer alır.
-1 tümüyle kötü bir sistemi, +1 ise tamamen başarılı bir sistemi ifade eder.
[Landis ve Koch, 1977]'e göre Kappa değerinin 0.4'ten büyük olması o sistemin
şansın ötesinde makul bir başarı gösterdiğini anlamına gelir. 4.8 nolu tablodaki
Kappa sonuçları çalışmada denenen yaklaşımların rastgele sınıflandırıcılardan
çok daha iyi olduğunu ifade ediyor.

Her ne kadar öğrenme algoritmalarının sonuçları rastgele sınıflandırıcıların-
dan başarı olsa da, problemin çözümüne ilişkin değerlendirmede tam anlamıyla
başarılı sayılmaz. [Ho vd., 2006] bunun sebebine ilişkin üç olası etkenden bah-
seder:

1. Algoritmadan kaynaklı eksiklik
2. Verideki içkin zorluklar
3. Metod ile problemin uyuşmaması

Şimdi bu üç başlıklı çerçevede problem tartışılacak. Tablo 4.8'de sunulan
sonuçlara bakıldığında belli bir algoritmanın daha başarılı olduğu söylenemez.
Aynı algoritma değişik parametrelerle değişik sonuçlar üretebiliyor. İlk beş sıra-
lamada doğrusal çekirdekli C-SVC neredeyse en kötü başarıya sahip olmasına
karşın, ikinci beşlide radyal çekirdekli C-SVC en iyi başarıya sahiptir. Sonuç
olarak artgönderim problemi için başarılı olabilecek evrensel bir algoritmadan
bahsedilmesi mümkün değil. Bu değerlendirme kural tabanlı yöntemler içinde
söylenebilir.

Bir diğer zorluk ise verideki ya da derlemdeki zorluklardır. Artgönderim
problemi için baktığımızda bu durumu aşağıdaki söylemle örneklemek müm-
kündür.

Ayşe Hasan’a yardım etti. Çünkü Onun yardıma ihtiyacı vardı. (4.31)

Yukarıdaki örnekte **Onun** ifadesi Ayşe’ye mi yoksa Hasan öncülüne mi gönderimde bulunacak ? Merkezleme kuramının yaklaşımı ele alındığında Özne nesneye tercih edilecektir. Dolayısı ile Ayşe öncülü seçilecektir. Ancak normal bir okumada Hasan öncülü daha doğru gözükmemektedir. Bu çıkarıma ulaşmak için belki de anlambilimsel yöntemler kullanılması gerekiyor. İşte Artgönderim problemine dair buna benzer birçok örnek vardır. Bu tip örnekler muğlaklık (ambiguity) olarak niteleniyor Tüm bu örnekler burada sıralanmayacak. Öte yandan Türkçe için cinsiyet kodu bir dezavantaj oluşturmaktadır. Yukarıdaki örnek İngilizce için muğlaklık yaratmazdı. Çünkü Ayşe ve Hasan farklı şekilde cinsiyet kodlamalı adıyla ifade edileceklerdi. Örneğin **his** ya da **her** adıyla kullanılırdı ve bir muğlaklıktan bahsedilmezdi.

Artgönderim problemindeki bir diğer önemli husus adılar için elemeler sonunda kalan öncül sayısıdır. Eğer eleme sonucu sadece bir aday kalmışsa bu problemin çözümü kolaydır. Dolayısıyla veritabanında ya da derlemde kaç adılın elemeler sonunda tek bir adayının kaldığı çok önemlidir. Bu oran yüksekse söz konusu verinin karmaşıklığı oldukça düşüktür. Bu tez çalışmasındaki derlemde bu oran % 3.1’dir ve bu anlamda çalışılan derlem karmaşıktır.

Sınıflandırıcıların performansını etkileyecek bir diğer öge ise verinin dağınıklığı veya yetersizliğidir (sparsity). Veri eğer yeterli miktarda örneğe sahip değilse, sınıflandırıcılar genel bir kural oluşturma yeteneklerini kaybederler. Yani bir genelleme yapmak için belli sayıda örneğe ihtiyaç duyarlar. Bu hassasiyeti ölçmek için yapılan deneyde, veri 1100 örnek kalıncaya kadar değiştirildiğinde f-ölçütü %60 değerinden %73’e çıkıyor. Ve bundan sonra bir doyum yakalıyor ve doyum

noktasında sabit kalıyor.

Bir diğer etken ise problem ile bu probleme uygulanan metod arasındaki uyumsuzluktur. Bu tez çalışmasına yönelik olarak 4.8 nolu tabloda gösterilen C-SVC, Voted Perceptron ve Naïve Bayes sınıflandırıcıları ele alınabilir. Bu algoritmalar için uygulanan parametrelerdeki doğrusal ve doğrusal olmayan seçimler üzerinden bir yorum yapılabilir. C-SVC ile alınan en kötü sonuç bu yöntem için uygulanan doğrusal formüldür. Bu formül aşağıdaki 4.32 nolu formüldür.

$$f(x) = (w^T x) + b \quad (4.32)$$

En iyi sonuç ise radyal tabanlı kullanılan aşağıdaki 4.33 nolu formüldür.

$$k(x, x) = \exp \frac{(\| x - x \|)^2}{2\sigma^2} \quad (4.33)$$

Weka yazılımındaki Voted Perceptron için kullanılan aşağıdaki 4.34 nolu formüldeki

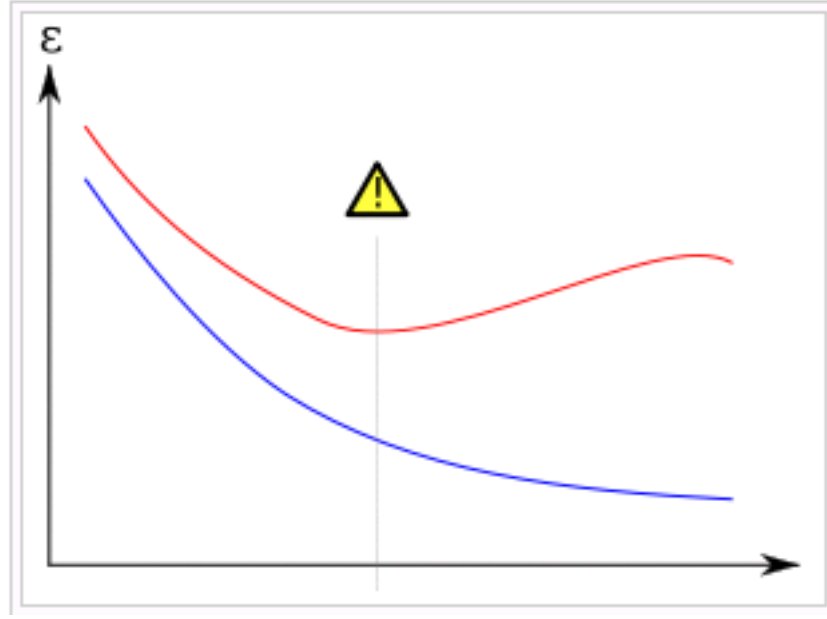
$$k(x, x') = (x \cdot x' + 1)^d \quad (4.34)$$

d değeri 1 seçildiğinde doğrusal bir çekirdek, daha büyük seçildiğinde doğrusal olmayan bir çekirdek kullanmış oluruz. Bu algortmada yine doğrusal olmayan çekirdek kullanıldığında başarı oranı artmıştır. Benzer bir özelliği Naïve Bayes için söyleyebiliriz. Bu algoritma normal dağılım ile uygulandığında doğ-

rusal bir sınıflandırıcı ve çekirdek yoğunluk tahmini (kernel density estimator) kullanıldığında doğrusal olmayan bir sınıflandırıcı sınıfına dahil olur. Bu örnekte de diğer örneklerde olduğu gibi doğrusal olmayan ayarlar çerçevesinde artgönderim sistemi biraz daha başarılı olmuştur. Yukarıdaki gözlemler veritabanının ya da derlemdeki gönderimsel ilişkilerin doğrusal bir şekilde değil de doğrusal olmayan bir biçimde olduğu sonucunu verir. Ve bu nedenle doğrusal olmayan yöntemlerin daha iyi sonuç vermesi olağandır.

Aşırı Uyum (Overfitting) ve Yetersiz Uyum (Underfitting)

Makine Öğrenmesi ve istatistik alanındaki en önemli sorulardan biri Aşırı Uyum ve Yetersiz Uyum ikiliğidir. Üretilen modelin kendisini oluşturan veriye aşırı bir şekilde bağımlı olması ve genellikle uzaklaşması Aşırı Uyuma yol açar. Veri içinde birçok istisna barındırır ve aynı şekilde çeşitli gürültüler içerebilir. Yani ele alınan veri birçok tutarsızlığını birlikte getirir. İçinde hatalı veriler barındırır. Verinin bütününe ulaşmaya ve onu modellemeye çalışmak yeni oluşacak verileri anlamayı zorlaştırır. Bu olguya veri madenciliği alanında İngilizce’de **overfitting** denir. Tükçe için yaygın bir karşılığı olmasa da bu kavram **aşırı uyum** olarak çevrilebilir. 4.7 nolu şekildeki eğriler öğrenme sürecindeki hata eğrileridir. Üstteki eğri sağlama ya da test kümesinin hata durumunu ifade ederken, alttaki eğri eğitim kümesine ilişkindir. Bu grafikteki eğitim kümesinde hatanın belli bir etkene göre azaldığı görülür. Bu etkenin ne olduğu bu tartışma açısından şimdilik önemli değildir. Sadece yaratılan modelin eğitim kümesi üzerindeki hatasının belli bir etken sebebiyle azaldığını düşünölsün. Bu azalmaya sağlama/test kümesi de belli bir noktaya kadar eşlik eder ve bu noktadan itibaren sağlama kümesindeki azalma durur ve artmaya başlar. İşte bu durma ve artma noktasının sonrası çalışmanın **Aşırı Uyum (overfitting)** olduğu bölgeyi ifade eder.



Şekil 4.7: Aşırı Uyma

Bu etkenlere örnek vermek için Karar Ağaçlarının budama işleminden bahsedilebilir. Karar Ağacı ne kadar az budanırsa eğitim kümesi üzerindeki başarısı o kadar artar. Sağlama kümesinde belli bir noktaya kadar budama yapılması iyi sonuç verir. Yani ağac budama onun veriye aşırı bağımlı olmasını engeller ve sağlama kümesi için daha iyi sonuç verir.

Çalışmada kullanılan veri dokuz değişken içerdiğinden doğrusal yöntemleri kullanan sınıflandırıcılar için uygun değildir. Doğrusal yöntemler uygulandığında veri ile yaratılan model arasında büyük bir sapma / kutup (bias) oluşur.

K-en yakın komşu sınıflandırıcısı ile Karar Ağacı algoritması ele alındığında, dikkat edilmesi gereken noktalardan biri de bu aşırı uyum sorunudur. Çünkü bu iki algoritmanın veriyi ifade etme gücü (expressive power) yüksektir. Çünkü eğer veri de herhangi bir muğlaklık yok ise verinin miktarına bakmaksızın eğitim verisini %100'e varan bir kesinlikle ifade edebilirler. Bu durumda gürültülü ve hatalı veriler içerdiğinden aşırı uymaya (overfitting) yatkındırlar. KNN algoritmasını K değerinin 1 yapılması, Karar Ağacı algoritmasının boosting özelliğinin

kullanımı ve budama yapılmaması bu algoritmaların ifade güçlerini (expressive power) artırır. Bu da gürültülü ve hatalı bilgilerin yarattığı aşırı uyma sorununu ortaya çıkarır. Bu nedenle 4.8 nolu tabloda bu iki algoritmaların bu ayarları için oluşan sonuç diğerlerine oranla daha düşük çıkmıştır. Hatta bu sonuçlar doğrusal sınıflandırıcıların performansı kadar düşüktür. Bu aşırı uyumun çözümü için ağacının budanması ve K değerinin artırılması gerekiyor. Biraz önce bahsedildiği gibi ve 4.7 nolu şekilde gösterildiği gibi hata oranını azaltmak uygun bir nokta belirlemek gerekiyor. Ağac için bu budama oranı genelde %25 olarak kabul edilir. K değeri 11 yapıldığında bu algoritma daha başarılı sonuç vermektedir. Ve bu sonuç ilginç bir şekilde doğrusal olmayan yöntemlerin sonuçlarına yaklaşmaktadır. Karmaşıklık ile aşırı sadelik performans anlamında birbirleriyle aynı kusurlara sahip. Karmaşıklığın aşırı uyuma yol açtığı, sadeliğin ise yetersiz uyum yarattığını bu deneylerde gözlemledik.

Özelliklerin Katkıları

Kullanılan özelliklerin artgönderim için ne kadar katkı sağladığı en önemli noktalardan biridir. Dolayısıyla her bir özelliğin katkısını kullanılan teknik açısından değerlendirmek gerekiyor. Bu değerlendirmeyi bilgi tabanlı yaklaşımlar için yapılmıştı. Şimdi öğrenme tabanlı yaklaşımlar açısından ne gibi bir sonuç elde edilecek ? Makine öğrenme algoritmaları kullanıldığında bu değerlendirme 10'lu çapraz sağlama değerlendirmesi içinde birçok sınıflandırıcı kullanılarak yapılabilir. 4.10 nolu tabloda herbir özelliğin katkısının ortalaması bulunur.

Bir diğer özellik gücü ölçme yöntemi bilgi kazancı (information gain) formülüdür. Bu formül Karar Ağacı oluşturma gibi sınıflandırıcıların içerdiği formülüdür. Özelliklerin güçleri hedef sınıfla olan bağlarından ortaya çıkar. Hedefin belirlenmesinde etkili olan özellikler daha fazla kazanca sahip olurlar. Bu formül

Tablo 4.10: Özelliklerin Katkı Değeri Ortalamaları / 10'lu Çapraz Sağlama

Özellik	Ortalama Katkısı
Adılın Durum Eki	85
Adılın Dilbilgisel Görevi	82
Adılın Açıklığı	62
Adılın Sözdizim Tipi	93
Gönderimsel İfadelerin Durum Eki	77
Gönderimsel İfadelerin Anlambilimsel Tipi	94
Gönderimsel İfadelerin Dilbilgisel Rolü	81
Uzaklık	100
Kişi-Sayı Uyumunu	0 ³

ile oluşturulan 4.11 nolu tablo büyükten küçüğe sıralanmıştır.

Tablo 4.11: Özelliklerin Bilgi Kazancı

Özellik	Bilgi Kazancı / bits
Uzaklık	0.1844
Gönderimsel İfadenin Anlambilimsel Tipi	0.1040
Gönderimsel İfadenin Dilbilgisel Rolü	0.0918
Gönderimsel İfadenin Durum Eki	0.0767
Adılın Durum Eki	0.0191
Adılın Dilbilgisel Rolü	0.0173
Adılın Sözdizimsel Tipi	0.013
Adılın Açıklığı (overtness)	0.0001
Kişi-Sayı Uyumunu	0

Bu iki tablodan çıkarılacak iki net sonuç var. Bunlardan biri adıl ile aday öncüller arasındaki uzaklık önemli bir unsurdur. Bu sonuç diğer diller içinde geçerlidir. Giriş kısmında bahsedilen yaklaşımların çoğunda uzaklığın gönderimsel ilişki üzerindeki etkisinden bahsedilir. Bir ikinci sonuç ise sayı/kişi uyumunun

gücünün 0 (sıfır) çıkmasıdır. Bu durum esasında bu özelliğin önemsiz olduğu anlamına gelmez. Çünkü eleme sürecinde sayı/kişi uyumu göstermeyen adaylar sistemden elenir ve söz konusu adıl için oluşan adayların hepsi zaten adıl için sayı/kişi uyumunu sağlar. Bu bakımdan sayı/kişi uyumu özelliği bu aşamada bir işe yaramıyormuş gibi görünse de asıl görevi bir önceki eleme aşamasındadır ve diğer özelliklere göre çok önemlidir. Bu anlamda en önemli özelliktir diyebiliriz. Diğer özelliklerin katkısı da önemsiz değildir. Anlambilimsel durum her zaman gönderimsel ilişkilerde etkisi vurgulanmış bir etkidir. Benzer bir şekilde Dilbilgisel Görevde tıpkı Merkezleme kuramında olduğu gibi çok önemli bir işleve sahiptir.

Artgönderim Çözümü Açısından Performans Değerleri

Artgönderim Çözümlemesi için makine öğrenmesi yaklaşımları kullanıldığında söylemdeki adılar ve öncüller ikili bir kayıt halinde ifade edilir ve ardından her bir kaydın doğru veya yanlış olduğu kaydedilir. Bu şekilde yapılmasının temel sebebi öğrenme algoritmalarının kayıtlar üzerinden ikili sınıflandırma yapabilme gücüdür. Her kaydın bir hedef sınıfı vardır bu sınıf ikili veya çoklu olabilir. Bu yöntem kullanıldığında makine öğrenme algoritması yeni gelen derlemi yine bu kayıtlar şeklinde ifade eder ve doğru yanlış şeklinde sınıflandırma yapar. İşte bu aşamada tekrar iki kayıtlardan eski derleme geri dönmek gerekiyor.

Bu noktadaki iki temel problem var:

1. Adılın eşleştiği öncüller ile oluşturduğu tüm kayıtları algoritmanın yanlış olarak etiketlemesi. Bu durumda söz konusu adıl için algoritma bir öncül önermemiş oluyor.
2. Adılın eşleştiği öncüllerden birden fazlası için doğru işaretin verilmesi. Bu

durumda algoritma adlı için birden fazla öncül önermiş oluyor. Ve bu durumda yine bir muğlaklık (ambiguity) oluşmaktadır.

İşte bu muğlaklığı çözmek için model aşağıdaki şekilde güncellendi ve sonuçlar bu açıdan değerlendirildi.

1. Eğer hiçbir aday önerilmemişse tüm aday öncüllerin en yakını o adının öncülü olarak atanır.
2. Eğer birden fazla aday önerilmişse o adayların adla en yakın olanı öncül olur.

Algoritma bu şekilde güncellediğinde 4.12 nolu tablo oluştu.

Tablo 4.12: Artgönderim Başarı Oranı

Model Ayarları	Başarı Oranı
Voted Perceptron - Üst=1	0.52
C-SVC - Doğrusal Çekirdek	0.54
Naïve Bayes- Normal Dağılım	0.61
J48 - Budamalı	0.62
Voted Perceptron - Üst = 2	0.63
Naïve Bayes - Çekirdek Tahim	0.63
IBk - k=11	0.66
J48 - Boosted ve Budamalı	0.69
C-SVC - Radyal Tabanlı Çekirdek	0.70
IBk - k=1	0.75

Bu tabloki sonuç şu şekilde özetlenebilir: Bir modelin ifade gücü (expressive power) ne kadar yüksekse, artgönderim çözümlemesinde o kadar başarılıdır. Tabloda göze çarpan bir diğer unsur kötü sonuca sahip alan doğrusal modellerdir:

1. Voted Perceptron - üs = 1

2. Doğrusal çekirdekli C-SVC
3. Normal Dağılımlı Naïve Bayes

Bunula beraber ifade gücü yüksek ayarlara sahip olan modeller şunlardır:

1. J48 - Budamalı
2. KNN - k=1

Yukarıdaki modeller daha başarılı olmuştur. Bu gözlem sınıflandırıcıların performansına tezat oluşturur. Çünkü ifade gücü çok arttıkça sistem aşırı uyumdan zarar görmekteydi ve başarı oranı belli bir noktadan sonra düşmekteydi. Bu farklı duruma şu şekilde açıklık getirmeye çalışalım.

Kullanılan Türkçe derlemde aday öncüller adıldan uzaklaştıkça gerçek öncül olma olasılıklarını kaybediyorlar. Adılın öncülü genelde ona en yakın öncüllerde yer alır. Bu çerçevede, muğlaklık olduğunda ve adaylardan en yakını seçildiğinde ifade gücü yüksek olan modellerin aşırı uyum riski azalmış olur.

Kritik Başarı Oranı (Critical Success Rate)

Mitkov artgönderim için dikkat edilmesi gereken bir performans ölçüm yaklaşımı önermiştir. Bu yaklaşıma göre eğer bir adıl için eleme işleminden sonra tek bir aday kalmışsa çözüm çok kolaydır. Dolayısıyla bu tip durumların çözümü önerilen sistemden çok durumun kendisinden kaynaklı çözülmesi basit bir örnektir. Yani sözkonusu derlem basit bir derlemdir. Bu tip örnekleri başarı hesabına katmaksızın, yani bu tip adılar dışında kalan adılar üzerinden bir başarı oranından bahsetmek gerekir. Çözümü kolay olmayan adılar için yapılan hesap sonunda oluşan performans oranına **kritik başarı oranı** adı verilmiştir. Bu

tez kapsamında sistemde elde edilen sonuçlar söz konusu kritik başarı oranına oldukça yakındır. Çünkü sistemde eleme işleminden sonra tek bir aday öncülü kalan adılların tüm adıllara oranı %3 gibi ihmal edilebilir küçüklüktedir. Bu anlamda her iki yaklaşımda (bilgi tabanlı ve makine öğrenme tabanlı) sunulan başarı oranları kritik başarı oranı olarak düşünülebilir. Bu oran günümüzde yapılan tüm Bilgisayarlı Artgönderim çalışmaları için kabul edilmiş bir orandır.

Bölüm 5

DENEY SONUÇLARININ DEĞERLENDİRİLMESİ

Bu çalışma kapsamında gerek öğrenme tabanlı yöntemler gerek bilgi tabanlı yöntemler Türkçe metinlere uygulandı. Değerlendirmeye makine öğrenme algoritmalarının sonuçlarıyla başlanabilir. Bu algoritmalarındaki süreç üç aşamadan oluşur. Birincisi her bir söylemdeki adılar ve onların olası aday öncülleri belirlendi. Her bir eşleşme veri tabanı kayıtları gibi ifade edilip, **evet/hayır** olmak üzere ikili sınıflandırıldı. Böylece üzerinde öğrenme algoritmaları kolaylıkla uygulanabilir duruma geldi. Bir diğer aşama öğrenme algoritmasının bu veriden bir model oluşturması oldu. Bu modeller, o modeli oluşumu için kullanılmayan başka söylemlerden oluşan veriler üzerinde denendi. Aslında öğrenme algoritmaları eşleşmelere evet/hayır diyerek gönderimsel ilişkileri belirlemeye çalışır. Ancak Makine Öğrenme Algoritmaları her bir adılın sadece bir tane öncül adaya sahip olması gerekliliği bilgisinden yoksundur. Bu nedenle sistemin bir üçüncü aşamaya ihtiyacı vardı. Bu aşamada eğer birden fazla aday öncül varsa ya da hiç aday yoksa en yakın adayın seçilmesi işlemi yürütüldü.

Sınıflandırma aşamasında değerlendirme sistemi için f-ölçütü kullandığında elde edilen sonuçlar %64 ve %74 arasında çıktı. Bu değerlerin yeterli olduğunu Kappa istatistiği aracılığı ile ifade edilebildi. Çünkü bu istatistik öne sürülen modelin rastgele bir modelden ne kadar başarılı olduğunu belirler. Bu bakımdan sayısal açıdan da oluşan modellerin belli bir başarıyı yakalayabildiğini görebilmekteyiz. Bu istatistiğe göre Kappa değeri 0.4'ten büyük çıkarsa önerilen modelin başarılı olduğu söylenebilmektedir. Önerilen algoritmaların çoğu bu değerlerin üstünde sonuç vermiştir.

Doğrusal olmayan ayarlara sahip öğrenme algoritmalarının, diğerlerine oranla daha başarılı olduğu görülmüştür. Bunun sebebi de derlemde oluşan kayıtların doğrusal bir yöntemle sınıflandırılmamasıdır. Aynı zamanda doğrusal olmayan sınıflandırıcılar ise belli bir seviyeden sonra aşırı uyumdan olumsuz etkilendiklerinden başarı oranları düşmektedir.

Makine Öğrenmesi sonrası artgönderim çözümü sırasında ise şu sonuçlar elde edildi. Bir modelin ifade gücü artar ise, başarı oranı artar. Herhangi bir aşırı uyuma yol açmaz. Bunun asıl sebebi birden fazla aday seçildiğinde en yakındaki adayın seçilmesidir. Derlemdeki adıl/öncül uzaklığının oldukça küçük olması sebebiyle bu aşırı uyum (overfitting) etkisi yok oldu.

Bir diğer gözlem ise artgönderim çözümünde uzaklığın ve kişi/sayı uyumunun çözümde çok etkili olduğudur. Bu önceki bölümlerde verilen tablolardan da görülebilir. Bu özellik diğer bilgi tabanlı yaklaşımlar için de güçlü bir özellik olarak ön plana çıkar.

Bilgi tabanlı ya da kural tabanlı yaklaşımlardan RAP temelli yaklaşım Türkçe alana uyguladığında elde edilen en iyi sonuç % 53 çıktı. Bu sonuç için algoritmadaki parametreler sürekli değiştirildi. Ve sonuçta 1114 adıldan 598'i doğru bir şekilde çözümlendi. Diğer yaklaşımlarla karşılaştırıldığında bu başarı oranının

daha düşük olduğu söylenebilir.

Bilgi tabanlı yaklaşımlardan Merkezleme Teorisini ele alan yaklaşım Türkçe alanına uyguladığında RAP temelli algoritmaya göre daha iyi bir sonuç alındı. Bu yaklaşımda 1114 adıldan 792'si başarılı bir şekilde çözümlenmiş oldu. Buradaki % 0.71 (792 /1114) başarı oranı RAP algoritmasına göre daha iyi bir sonuçtur. Bununla birlikte makine öğrenme algoritmalarına göre ise yeterli bir seviyede olduğu söylenebilir. Bu aşamada bilgi tabanlı ve makine öğrenme tabanlı yaklaşımların sonuçlarını tek bir tabloda ifade ederek değerlendirilebilir. 5.1 nolu tabloda hem makine öğrenme algoritmalarının hem bilgi tabanlı algoritmaların sonuçları listelenmiştir. Belirginlik temelli yaklaşımın Türkçe derlem üzerindeki başarı oranı **Algılayıcı** (Voted Perceptron) sınıflandırıcısı ile neredeyse aynıdır. Bu sınıflandırıcı diğer makine öğrenmesi algoritmalarıyla karşılaştırıldığında en kötü sonuca sahip olan algoritmadır. Daha önce de değinildiği gibi bu algoritma ifade gücünün zayıf olması sebebiyle derlemdeki durumları tam olarak ifade edememektedir. Üst değerinin 1 olması bu algoritmanın doğrusal bir sınıflandırıcı gibi davranmasına yol açmaktadır. Bu nedenle derlem üzerindeki başarı oranı düşüktür.

Bu noktada cevabı aranan soru şudur:

Peki Belirginlik Temelli Yaklaşım neden makine öğrenmesi algoritmalarından daha kötü bir sonuç verdi ? Ve bu algoritmalarından üs derecesi 1 olan, yani doğrusal bir öğrenme algoritması gibi davranan algılayıcı (Voted Perceptron) ile neredeyse aynı sonucu vermesinin nedeni nedir ?

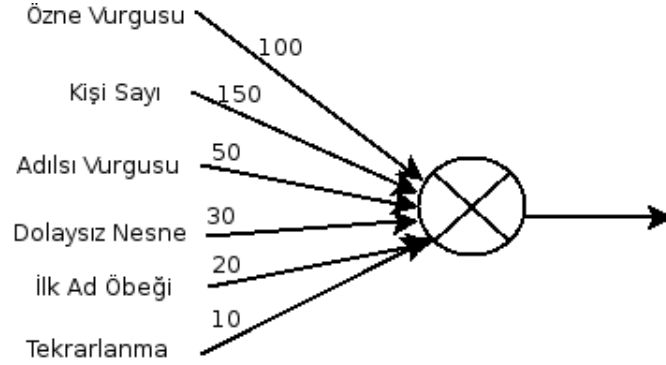
Makine Öğrenme algoritmalarının en önemli özelliği veritabanındaki gizli desenleri, kuralları, formülleri bulup çıkarabilmesidir. Dolayısı ile artgönderim

Tablo 5.1: Artgönderim Başarı Oranı

Makine Öğrenme Temelli Yaklaşımlar	
Model Ayarları	Başarı Oranı
Voted Perceptron - Üst=1	0.52
C-SVC - Doğrusal Çekirdek	0.54
Naïve Bayes- Normal Dağılım	0.61
J48 - Budamalı	0.62
Voted Perceptron - Üst= 2	0.63
Naïve Bayes - Çekirdek Tahim	0.63
IBk - k=11	0.66
J48 - Boosted ve Budamalı	0.69
C-SVC - Radyal Tabanlı Çekirdek	0.70
IBk - k=1	0.75
Bilgi Tabanlı Yaklaşımlar	
Belirginlik Temelli Yaklaşım (RAP)	0.53
Merkezleme Temelli Yaklaşım	0.71
Merkezleme Temelli (Adılsı Vurgusu)	0.72

problemine uygulanacak öğrenme algortimaları söylemdeki gönderimsel ilişkilerin hangi kurallara göre oluştuğunu bulabilme kapasitesine sahiptir. Şimdi bu aşamada belirginlik tabanlı yaklaşım içerdiği yöntemden dolayı doğrusal bir modele sahiptir. Yani girdi olarak aldığı özellik listesini belli ağırlık değerlerine göre çarparak bir sonuç üretmeye çalışır. Daha önceki bölümlerde ifade edildiği gibi algoritma gözden geçirildiğinde bu yöntemin bu manada basit ve doğrusal bir yapısı vardır ve sadece derlemdeki doğrusal bilgileri bulabilir. Makine Öğrenme algortimaları da, eğer derlemde mevcut ise bu basit ve doğrusal kuralları rahatlıkla bulur. Ancak öğrenme algortimaları belirginlik temelli yaklaşımın aksine doğrusal olmayan kurallarıda bulabilir. Eğer derlem doğrusal yöntemle sınıflandırılmıyorsa, doğrusal olmayan makine öğrenme yöntemleri bu kuralları bulup çıkarma yeteneğine sahiptir. Yani belirginlik temelli yaklaşımındaki algoritma makine öğrenme algortimaları tarafından zaten öğrenilebilir ve hatta belirginlik

tabanının ulaşamayacağı desenleri de derlemden bulup çıkarabilir. Belirginlik tabanlı yaklaşım tam da Doğrusal Voted Perceptron ($\alpha=1$) gibi çalışır. 4.3.2 nolu bölümde tanımlanan Belirginlik Temelli algoritma birinci dereceden bir algılayıcı gibi ifade edilebilir. Örneğin 4.3.2 nolu tablodaki ağırlıkları kullanarak yaratılan model 5.1 nolu şekildeki gibi ifade edilebilir.



Şekil 5.1: Belirginlik Temelli Yaklaşım Gösterimi

Bundan dolayı uyguladığımız belirginlik temelli yaklaşım öğrenme algoritmaları tarafından rahatlıkla öğrenilebilir. Veri doğrusal sınıflandırıcılarla ifade edilemez ise belirginlik tabanlı yöntem doğrusal olduğundan veriyi doğru ifade edemez. Sonuç olarak belirginlik tabanlı yaklaşımın doğrusal olmayan derlemdeki gizli desenlere ulaşma yeteneği yoktur. Bu yaklaşımda kullanılan ağırlıklandırmadaki formül, Doğrusal Voted Perceptron ile birebir aynıdır. Algılayıcı (Voted Perceptron) derlemdeki bilgileri kullanarak en iyi ağırlık değerlerini kendi bulurken, belirginlik tabanlı yaklaşımda bu ağırlıklandırmalar elle verilir. Bazı denemelerle en iyi sonucu gözlemleyene kadar ağırlıkları değiştirmek durumunda kalınabilir. İşte bu yapısal benzerlik sebebiyle doğrusal algılayıcı (Voted Perceptron) ile belirginlik tabanlı yöntemin sonuçları birbirine oldukça yakındır. Ama yine de doğrusal olarak sınıflandırılmayan bir veri üzerinde elde edilen sonuç çok başarılı değildir.

Öte yandan Merkezleme yaklaşımının sonucu hem belirginlik tabanlı yakla-

şıma göre hem de birçok öğrenme algoritmasına göre daha iyi sonuç vermiştir. Bunun en önemli sebebi Türkçe'nin Merkezleme Teorisine uygun bir söylem yapısı oluşturduğudur. Bu yaklaşımda kullanılan sıralama algoritması dilbilgisel roller üzerinden yapılmıştır. Bu roller konu ve odak oluşturmada çok belirleyici olduğundan Türkçe derlem üzerinde başarılı olmuştur. Türkçe söylem davranışındaki merkezin sürekli sabit kalma eğilimi merkezleme teorisinin başarısını artırmıştır. Aslında bu durum sadece Türkçe'ye özgü değildir. Diğer dillerde de bu durum geçerli olabilir. Bu şimdilik bu çalışmanın kapsamı dışındadır. Türkçe'nin bu merkezleme kurallarına yatkın bir davranış sergilediği görülmüştür. Bu durum belki sadece hikayelerden oluşturulan derlem sebebiyle oluşmuş olabilir. Merkezdeki söylem varlığı sisteme girdikten sonra uzun süre merkezi işgal etmesi bu teorinin daha başarılı olmasını sağlamıştır. Zaten 4.6 nolu Karar Gücü tablosuna bakıldığında **ilk ad öbeği vurgusu** etkeninin çok önemli bir güce sahip olduğunu görülür. Çünkü sisteme ilk dahil olan söylem varlığı en azından Türkçe hikaye yapısında merkezde kalma eğilimindedir. Bu nedenle Merkezleme Algoritması bizim çalışma alanımızda yüksek bir başarı kaydetmiştir. Örneğin aşağıdaki Dostoyevski'nin Yeraltıdan Notlar eserinden alınmış aşağıdaki metnine bakılabilir.

Başımı tiksintiyle öteye çevirdim. Artık konuşmam, soğukkanlı felsefe yapmak şeklinden çıkmıştı. Yavaş yavaş neler söylediğimin farkında oluyor, heyecanlanıyordum. Köşemde tek başıma otura otura içimde doğan fikirleri dökmek için sabırsızlanıyordum. İçimde bir alevin parladığını hissettim. Önümde bir gaye belirmişti...

Bu metindeki birinci tekil şahıs (ben) boş artgönderim olarak tüm metinde tekrarlanmaktadır. Bazen söylemin merkezi değişse de ilk ad öbeği olarak sisteme giren $\odot$ (ben) boş artgönderimi metin boyunca sürekli kendini merkezde

boş artgönderim olarak tekrarlar. Bu da merkezleme terosinin DEVAM geçişinin sürekli tercih edilmesi manasına gelir. Türkçe hikaye metnindeki bu özellik Merkezleme Teorisi ile yapılan deneyin başarılı çıkmasını sağlamıştır.

Tüm bu denemeler çerçevesinde Türkçe derlemde hangi özelliklerin gönderimsel ilişkiler üzerinde etkili olduğu birçok metrik sistem üzerinden tartışılabilir. 4.5 nolu Bağlı Önem tablosunda en önemli etkenler **sayı/kişi** uyumu ve **özne vurgusu** olmuştur. Kullanılan diğer özellikler belli bir seviyenin altında kalmıştır. Bununla birlikte 4.6 nolu Karar Gücü tablosunda **sayı/kişi** ve **özne vurgusu** dışında en önemli özellik **İlk Ad Öbeği** olmuştur. Bunu sebebi kullanılan Türkçe öykülerin karakteristiğinden kaynaklanabilir. Yukarıda bahsedildiği gibi sisteme dahil olan söylem varlığı sürekli tekrarlanarak söylem boyunca odağı ve konuyu işgal edebiliyor. Makine Öğrenme deneylerinde üretilen 4.10 nolu tabloda en önemli etken olarak **uzaklık bilgisi** öne çıkmaktadır. Uzaklık bilgisi Belirginlik tabanlı yöntemde özellik olarak kullanılmadığından *karar gücü* ve *bağlı önem* gibi tablolarda karşımıza çıkmadı. Ancak orada kullanılan algoritmada yine önemli bir etkidir. Veri Madenciliği alanında sıklıkla kullanılan bir diğer özellik gücü ölçme yöntemi bilgi kazancı yöntemidir. Daha önce 4.11 nolu tabloda verilen sonuçlarda ise dört özellik ön plana çıkmıştır: Uzaklık, anlambilimsel bilgi, dilbilgisel rol ve durum eki bilgisidir.

Bu sonuçlara bakıldığında tüm yöntemler için ortak öneme sahip özellikler uzaklık, **sayı kişi** uyumu ve dilbilgisel rol olarak sıralanabilir. Bu bakımdan oluşturulacak bir sistem sadece bu bilgiler üzerine kurulabilir. Bu bilgiler az bilgili bir sistemi ifade etmektedir. Bu çalışmadaki derlem üzerinde az bilgili bir artgönderim çözümlemesi başarılı bir şekilde uygulanabilir.

Bu aşamda uygulanan tüm metodlar genel bir tabloda belli açılardan değerlendirilebilir. 5.2 nolu tabloda bilgi tabanlı ve makine öğrenme tabanlı algo-

ritmalarının bazı özellikleri listelenmiştir: algoritmalarının öğrenme hızı, karar verme hızı, bilgi gereksinimi ve başarı oranı.

Tablo 5.2: Tüm Yaklaşımların Genel Görünümü

Yöntem	Öğrenme Hızı	Karar Hızı	Bilgi Gereksinimi	Başarı Oranı
Bilgi Tabanlı Yaklaşımlar				
Belirginlik Tabanlı Yöntem	Yok	Çok Hızlı	Normal	0.53
Merkezleme Yaklaşımı	Yok	Hızlı	Az	0.71
Öğrenme Tabanlı Yaklaşımlar				
IBk (k=11)	Yok	Çok Yavaş	Normal	0.66
J48 - Budamalı	Normal	Hızlı	Normal	0.69
IBk (k=1)	Yok	Çok Yavaş	Normal	0.75

Yukarıdaki tablonun da gösterdiği gibi bilgi tabanlı yaklaşımların herhangi bir öğrenme süreci mevcut değildir. Çünkü bilgi tabanlı yöntemler dilbilimsel teorilerden oluşturulmuştur. İstatistiksel bir yöntem içermezler. En önemli bilgi tabanlı yöntemlerden biri olan belirginlik tabanlı yöntemin karar verme hızı yüksektir. Çünkü içerdiği formül doğrusal bir toplamadan oluşur. Doğrusal bir algoritmaya sahip olduğundan, daha önce de belirttiğimiz gibi doğrusal olmayan derlemimiz üzerindeki başarı oranı da düşüktür. Bilgi gereksinimi ise normal düzeydedir.

Bir diğer yöntem olan Merkezleme Teorisinin karar verme hızı yüksektir ancak belirginlik tabanlı yönteme göre biraz yavaştır. Çünkü merkezleme teorisi basit bir toplama yapmaz. Bir çok sıralama ve senaryo seçim süreci içerir. Bu yönüyle belirginlik temelli yaklaşıma göre daha karmaşıktır. Öte yandan Merkezleme teorisinin ihtiyaç duyduğu bilgi sadece isim halleridir. Bu anlamda bilgiye en az ihtiyaç duyan yöntem bu yöntemdir. Diğer tüm yöntemler biçim-

bilimsel, dilbilgisel, anlambilimsel gibi farklı farklı bilgiye ihtiyaç duyarlar. Bununla birlikte merkezleme teorisinin başarı oranı belirginlik tabanlı yöntemle göre daha yüksektir. Çünkü daha önce belirtildiği gibi Türkçe hikayelerindeki yapısal durum Merkezleme Teorisini desteklemektedir.

Makine öğrenmesi algoritmalarından IBk (k-en yakın komşu) algoritmasının ilk versiyonunda k değeri 11 olarak kullanılmıştır. Bu algoritmanın öğrenme süreci yoktur. Çünkü yöntem yeni bir durum geldiğinde veri tabanını tarar. Bu bakımdan karar verme hızı oldukça yavaştır. K değeri 1 olduğunda ise karar verme hızı biraz daha yükselebilir ama yine de K-en yakın komşu algoritmalarının genel bir özelliği olan karar vermedeki yavaşlığı sabittir. J48 Karar Ağacı ise bir öğrenme sürecine ihtiyaç duyar. Ancak Karar Ağaçlarının karar verme hızı k-en yakın komşu algoritmasına göre daha hızlıdır. Karar Ağacının başarı oranı IBk (k=1)'den yüksek, ancak IBk(k=11)'den düşüktür.

IBk(k=1) algoritması tablodaki en başarılı yöntemdir. Ancak tasarlanacak bir sisteme uygun olup olmadığı oluşturulacak sistemin gereksinimine göre değişir. Çünkü bu algoritmanın karar verme hızı çok düşüktür. Öte yandan Karar Ağacının başarısı ona göre düşük olsa da karar verme hızı daha yüksektir. Hızlı karar verilmesi gereken sistemler için karar ağacı yapısı tercih edilebilir. Eğer oluşturulacak sistemin bilgi gereksinimi çok kısıtlıysa bu durumda tercih edilmesi gereken yöntem merkezleme yöntemidir. Çünkü merkezleme yöntemi hem başarı hem az bilgili olma açısından avantajlı durumdadır. Bu yöntem öğrenme süreci içermez, karar hızı yüksek, bilgi gereksinimi en az ve başarı oranı yüksektir.

Tüm bu tez kapsamında sunulan yaklaşımların sonuçları, bu alanda yapılmış diğer çalışmalarla karşılaştırılabilir. Ancak sağlıklı bir karşılaştırmanın önünde bazı problemler vardır. Bu problemler şu şekilde sıralanabilir:

1. Karşılaştırılan iki sistem farklı diller için yazılmış olabilir. İngilizce için yapılan bir çalışmayı Japonca için yapılmış başka bir çalışmayla birebir karşılaştırmak anlamsız olabilir.
2. Karşılaştırılan iki sistem aynı dil için oluşturulmuş olsa bile farklı derlemeler kullanıyor olabilir. Her derlemin kendine göre bir karmaşıklığı olduğundan ortaya çıkan başarı oranlarının birbiriyle karşılaştırılması yine anlamsız olacaktır.
3. Kullanılan derlemin büyüklüğü de karşılaştırmayı etkileyecek bir durumdur. Literatürdeki kimi çalışmalar 300 adıl içerirken kimi çalışmalar 5000 adıl içerebiliyor. Bu anlamda derlemin büyüklüğü de yaratılan sistemin güvenilirliğini belirler. Bu da göz önünde bulundurulması gereken başka bir kriterdir.
4. Artgönderim çözümleme problemindeki kapsam bir diğer kriterdir. Çünkü bazı çalışmalar sadece adılar üzerinde çalışırken bazı sistemler tüm gönderimsel ilişkiler üzerinde çalışabilmektedir. Örnek olarak bu tez kapsamında sadece açık adılar, dönüşlü adılar, boş artgönderimler, karşılıklı artgönderimler incelendi. Sözlüksel ad öbeği artgönderimleri kapsam dışında bırakıldı.
5. Artgönderim Çözümleme Sistemlerindeki bir diğer önemli unsur sistemin otomatik olup olmadığıdır. Çünkü gerçek ve güncel uygulamalar tam otomatik bir alanda çalışmak isterler. Artgönderim çalışmalarında derlemi elle işaretlemek sıklıkla başvuru bir yöntem. Ancak gerçek sistemlerde elle işaret vermek gibi bir şey söz konusu olmadığından bu durum bir dezavantaj yaratmaktadır. Ancak makine öğrenmesi çalışmalarında derlemin gönderimsel ilişkileri elle belirlenmek zorunda. Çünkü bir öğrenme süreci mevcuttur. Dolayısıyla oluşturulan sistemlerin ne kadar otomatik olduğu

önemlidir. Karşılaştırma yapılırken bu etkeni de göz önünde bulundurmak gerekiyor.

6. Az bilgili Artgönderim Çözümleme Sistemleri son zamanlarda sıklıkla tercih edilmektedir. Az bilgili / çok bilgili ikililiği her zaman göz önünde bulundurulacaktır. Gelecekte bilgi çok kolay erişilebilirse bu durumda çok bilgili sistemlere bir yönelim olabilir. Günümüzde bu erişim kolay olmadığından ve uygulanabilir gerçek sistemler oluşturulmaya çalışıldığından bu sistemlerin bilgiye bağımlılığının en aza indirilmesi ve daha hızlı çalıştırılması hedeflenmektedir. Sonuç olarak sistemler karşılaştırılırken az bilgililik etkeni de göz önünde bulundurulmalıdır.

Şimdi bu tezde bahsedilen diğer artgönderim çalışmalarının sonuçları Tablo 5.3'de gösterilmiştir.

Tablo 5.3'de de görüldüğü gibi artgönderim problemi için birçok dile yönelik çok sayıda çalışma yapılmıştır. Bu tablo yapılan çalışmaların bir kısmını temsil etmektedir. Ancak bunlar literatürdeki en önemli çalışmalar arasından seçilmiştir. Çalışmaların başarı oranları ve f-ölçütleri %60 ile %90 arasında değişmektedir. Bu sonuçlar mükemmel bir sistem için yeterli değildir. Dolayısıyla artgönderim çözümlemesi için daha çok çalışma yapılması gerekmektedir. Her ne kadar yukarıda sıralanan sebeplerden dolayı birebir karşılaştırma anlamsız olsa da çalışma sonuçlarının diğer çalışma sonuçlarına yakın olması sebebiyle bu çalışmada tasarlanan artgönderim sistemlerinin bu anlamda yeterli olduğu söylenebilir. Ama yine de Türkçe için daha birçok çalışma ve deney yapılması gerekiyor. Bu çalışma ve yukarıda sıralanan çalışmalar bize gelecek için bir takım ipuçları vermektedir. Bu manada değerlendirme yapmak faydalı olacaktır.

Tablo 5.3: Artgönderim Çalışmaların Başarı Oranları

Çalışma	Dil	Yöntem	Sonuç
[Kennedy ve Boguraev, 1996]	İngilizce	Bilgi Tabanlı	% 75
[Baldwin, 1997]	İngilizce	Bilgi Tabanlı	%92 kesinlik ve %64 anma
[Mitkov, 2002]	İngilizce	Bilgi Tabanlı Yarı Otomatik	% 85-90
[Mitkov, 2002]	İngilizce	Bilgi Tabanlı Tam Otomatik	% 61.55
[Mitkov, 2002]	Japonca Adaptasyon	Bilgi Tabanlı Yarı Otomatik	% 75.8
[Mitkov, 2002]	Bulgarca Adaptasyon	Bilgi Tabanlı Yarı Otomatik	% 72.6
[Aone ve Bennet, 1995.]	Japonca	Öğrenme Tabanlı	%70.20/%88.55/%77.27 (anma/kesinlik/f-ölçütü)
[McCarthy ve Lehnert, 1995.]	İngilizce	Öğrenme Tabanlı	%85.4/%87.6/%86.5 (anma/kesinlik/f-ölçütü)
[Soon vd., 2001.]	İngilizce	Öğrenme Tabanlı	%52/%68/%60 (anma/kesinlik/f-ölçütü)
[Yüksel ve Bozşahin, 2002]	Türkçe	Bilgi Tabanlı	% 70
[Tüfekci ve Kılıçaslan, 2005]	Türkçe	Bilgi Tabanlı	% 84
[Küçük, 2005]	Türkçe	Bilgi Tabanlı	%85/%88/%87 (anma/kesinlik/f-ölçütü)
Bizim Çalışma	Türkçe	Belirginlik Tabanlı	% 53
Bizim Çalışma	Türkçe	Merkezleme Yaklaşımı	% 71
Bizim Çalışma	Türkçe	J48 - Karar Ağacı	% 69
Bizim Çalışma	Türkçe	IBk (k=1)	% 75

Bölüm 6

SONUÇ

6.1 Amaç

Bu tez çalışması kapsamında Türkçe metinlerdeki artgönderimsel ilişkilerin bilgisayarlı çözümlenmesi araştırılmıştır. Artgönderim çözümlemesi bilgisayarlı dilbilim araştırma alanının önemli bir alt başlığı olduğundan birçok sistemin vazgeçilmez unsurudur. Türkçe için yürütülen bilgisayarlı dilbilim çalışmalarının azlığı ve bu alanlardaki çalışmalara duyulan ihtiyaç sebebiyle Türkçe için artgönderim çözümlemesi bu tez kapsamında araştırılmıştır. Kapsam olarak adısal artgönderim, dönüşlü artgönderim, boş artgönderim ve karşılıklı artgönderim tipleri ele alınmıştır. Deneylerin düzgün yapılması amacıyla yaklaşık 10 000 kelime kapasiteli ve 1114 adıl içeren bir derlem oluşturulmuştur. Deney sonuçlarını değerlendirebilmek için derlemdeki gönderimsel ilişkiler elle işaretlenmiştir.

6.2 Yöntem

En önemli hususlardan biri bu çözüm sistemini oluştururken hangi yaklaşımların kullanılacağıdır. Bu çalışma kapsamında birçok deney yapma olanağı bulunmuş, çeşitli yöntemler artgönderim probleminin çözümü için denenmiştir. Her alandaki bilgisayarlı sistemler için iki yaklaşım uygulanabilir: Kural tabanlı yaklaşımlar, Öğrenme Tabanlı yaklaşımlar. Bu iki yaklaşımdan birincisi olan kural tabanlı ya da bilgi tabanlı yöntem daha önce yapılmış olan teorik çalışmaları ele alır ve buna göre bir sistem oluşturur. Bu yaklaşım bir diğer ifade ile tündengelim olarak bilinir. İkinci yaklaşım olan öğrenme tabanlı yaklaşım ise tümevarım olarak da ifade edilebilir. Çünkü birçok örnek durumdan özel kurallar bulunmaya çalışılır. Makine Öğrenmesi olarak da bilinen bu yöntemler istatistiksel bilgileri kullanarak veriden bir model oluşturmaya çalışırlar ve bir genelleme ararlar.

6.3 Sonuçlar ve Bulgular

Bu tez kapsamında bu iki yaklaşımın nasıl bir sonuç vereceği araştırılmış ve bu yaklaşımla çeşitli algoritmalar oluşturulmuştur. Bilgi tabanlı yaklaşımın önemli bir yöntemi olan **belirginlik tabanlı** yöntem derlem üzerinde denenmiş ve %52 oranında bir başarı elde edilmiştir. Bir diğer bilgi tabanlı yaklaşım ise Merkezleme Algoritmasıdır. Bu algoritma daha önce de belirtildiği gibi söylemin merkezinin davranışına ilişkin çeşitli kurallar içerir. Merkezdeki ögenin kendini merkezde tutma eğilimi prensibi üzerine kurulmuş bu yöntem Türkçe derlem üzerinde sınıandığında ise %71 gibi bir sonuç elde edilmiştir. Öte yandan adılısı öğelerin çok daha belirgin olması ihtimaliyle bu algoritma güncellenmiş ve %72 gibi bir başarı oranı elde edilmiştir. Bu noktada ise adılısı öğelerin belirginliğinin

diğer ögelere göre çok daha fazla olduğu gözlemlenmiştir. Bilgi tabanlı yaklaşımın bu iki önemli algoritmasından merkezleme yönteminin belirginlik tabanlı yönteme göre daha başarılı olduğu gözlemlenmiştir. Bu deneyden Türkçe metinlerin merkezleme teorisinin vurguladığı kurallara yatkın olması sonucu çıkarılmış ve bu bulgu çeşitli örneklerle güçlendirilmiştir. Belirginlik tabanlı yöntemin doğrusal bir toplama işlemine dayanması ele alınan derlem üzerinde etkisiz kalmasına yol açmıştır. Bu yöntem makine öğrenmesi yöntemlerinden üs derecesi 1 olan algılayıcı ile benzeştirilmiş ve diğer yöntemlere göre daha başarısız olma sebebi doğrusallık üzerinden ifade edilmiştir.

Makine öğrenmesi algoritmalarından sadece beş tanesi çeşitli ayarlarla denenmiştir: Karar Ağaçları, K-en Yakın Komşu algoritması, Destek Vektör Makineleri, Naïve Bayes algoritmalar ve Algılayıcılar (Voted Perceptron). Bu algoritmalarından ilk ikisinin ifade gücü (expressive power) arttıkça, belli bir noktadan sonra aşırı uyum (overfitting) sorunu ürettiği gözlemlenmiştir. Bununla birlikte ifade gücü azaldıkça yetersiz uyum (underfitting) sorunu yarattığı da gözlemlenmiştir. Diğer üç algoritma ise doğrusal ayarlarla kullanıldıklarında derlem üzerinde başarılı olamamıştır. Bu ayarlar doğrusal olmayacak şekilde düzenlendiğinde ise başarı oranları yüksek çıkmıştır. Bu açıdan üzerinde çalışılan Türkçe derlemin doğrusal sınıflandırıcılarla ifade edilemeyeceği ve sınıflandırılmayacağı sonucu çıkarılmıştır. Doğrusal ayara sahip bir algılayıcının (Voted Perceptron) performansının belirginlik tabanlı yaklaşımla aynı çıkmasının yöntemlerin yapısal benzerliklerinden kaynaklandığı ve bu yapısal benzerliğin Türkçe artgönderim problemi için yeterli bir çözüm yöntemi oluşturamadığı gösterilmiştir.

Makine öğrenmesindeki elde edilen başarılar genelde anma, kesinlik ve f-ölçütü (recall, precision, f-measure) metrikleriyle değerlendirilir. Ancak makine öğrenmesi modelinin ürettiği sonuçların artgönderim çözümlemesi için bir aşağı-

madan daha geçirilmesi gerekmiştir. Çünkü makine öğrenme algoritmaları söyleme veya artgönderime özel olan kuralları bilemeyeceğinden bazı düzenlemeler yapılmıştır. Öğrenme sonrası (post-processing) süreç olarak tanımlanan bu aşamada öğrenme için kullanılan anma, kesinlik, f-ölçütü metriklerinden başarı oranı metriğine geçilmiştir. Bu geçiş sırasında "ifade gücünün çok olması aşırı uyuma (overfitting) yol açar" biçimdeki olumsuz etkinin yok olduğu gözlemlenmiştir. Muğlaklık durumunda başvuru *en yakın adayın öncül olarak seçilmesi*, bu olumsuz etkiyi azaltmıştır.

Öğrenme sonrası aşamada elde edilen sonuçlara bakıldığında en iyi sonucun K-en yakın komşu algoritmasının $k=1$ ayarı ile üretildiği gözlemlenmiştir. Bu algoritma %75 başarı oranı kaydederken, bu algoritmaya en yakın algoritma ise %71 başarı oranıyla Merkezleme Algoritması olmuştur. Diğer algoritmalarından Karar Ağacı %69 başarı oranı sağlamış ve K-en yakın komşu algoritması $k=11$ ayarı ile %66 başarı oranına sahip olmuştur. Birinci dereceden doğrusal ayara sahip Algılayıcı (Voted Perceptron) ile yine doğrusal bir algoritmaya sahip belirginlik tabanlı yöntemin benzer ve düşük bir başarı oranına sahip olduğu gözlemlenmiştir. Çünkü bu iki yöntem yapısal olarak da birbirlerine benzemektedirler. Bu algoritmalar ile en çok %52 gibi bir sonuç elde edilmiştir.

Artönderim Çözümlemesinde kullanılan özelliklerin hangilerinin gönderimsel ilişkilerde etkin olduğu çeşitli ölçüm yöntemleriyle sınanmıştır. Bu sayısal ölçüm tekniklerine dayanarak sayı/kişi uyumunun, dilbilgisel rollerin ve adıl ile öncül arasındaki uzaklığın etkili özellikler olduğu gösterilmiştir. Bunların yanında ilk ad öbeği vurgusunun Türkçe söylemde önemli olduğu saptanmıştır ve bu önem merkezleme algoritmasının başarısını da etkilediği kaydedilmiştir.

İstatistiksel ve öğrenme tabanlı yaklaşımların önemi bilimin her alanında kendisini hissettirmiştir. Dolayısıyla, Artgönderim Sistemlerinde de bu yakla-

şımlar önemli bir konum edinmiştir. Ancak yine de bilgi tabanlı yaklaşımlar halen daha bilgisayarlı dilbilim alanında çok büyük öneme sahiptirler. İstatistiksel yöntemler dildeki örnek söylemlere bakarak belli desenleri veya kuralları yakalamaya çalışırlar. Öte yandan bilgi tabanlı yöntemler ise bu desenleri ya da kuralları istatistiksel yöntemlere dayanarak değil de teorik bilgiler ışığında ortaya çıkarmaya çalışırlar. Bu iki yaklaşım daha önce de belirttiğimiz gibi tümevarım ve tümdengelim ikililiğinde yer alan alternatiflerdir. Bu iki yaklaşımı Türkçe Artgönderim çözümü için birbirlerini destekler ve benzer tepkiler verirler. Bir yaklaşımın diğer yaklaşıma göre belirli bir üstünlüğünün olmadığı görülmüştür. Makine öğrenmesi algoritmalarından en başarılı olanı k-en yakın komşu algoritması olmuşken, bilgi tabanlı yöntemden Merkezleme Algoritması en başarılı yöntem olmuştur. Bu iki farklı yaklaşımın sonuçları itibariyle birbirine yakın olması yukarıda özetlenen durumu destekler niteliktedir. Belki de makine öğrenme algoritmaları bilgi tabanlı algoritmaların kurallarını kendi yöntemleriyle derlemde çıkarabilmişlerdir.

Kaynakça

- [Aone ve Bennet, 1995.] Evaluating automated and manual acquisition of anaphora resolution rules. In: Proceedings of ACL95, 122-129.
- [Boser vd., 1992] A training algorithm for optimal margin classifier. 5th Annual ACM Workshop on COLT, Pittsburgh, PA. ACM Press.
- [Baldwin, 1997] Cogniac: High Precision Coreference with limited knowledge and Linguistics Resource. Operational factors in practical, robust anaphora resolution for unrestricted texts, pages 38-45.
- [Brennan vd., 1987] A centering approach to pronouns. In *The 25th Annual Meeting of the Association of Computational Linguistics*. Stanford.
- [Çeltek ve Oktar, 2004] Türkçe Sözlü Söylemde Artgönderim Örüntüleri *Dilbilim Araştırmaları.2004. Boğaziçi Üniversitesi Yayınevi*.
- [Chomsky, 1981] Lectures on Government and Binding, Dordrecht Foris
- [Enç, 1986] Topic Switching and pronominal subjects in Turkish *Studies in Turkish Linguistics. John Benjamin Publishing Company*.
- [Erguvanlı-Taylan, 1986] Pronominal versus zero representation of anaphora in Turkish. *Studies in Turkish Linguistics. John Benjamin Publishing Company*.

- [Fox, 1987] Discourse Structure and Anaphora: Written and Conversational English. Cambridge. *Cambridge University Press*.
- [Givon, 1983] Topic Continuity in Discourse: A Quantitative Cross-Language Study. *John Benjamins Publishing*.
- [Grosz, 1995] Centering: A framework for modeling the local coherence of discourse. In *Computational Linguistics*.
- [Grosz ve Sidner, 1986] Attention, intentions, and the structure of discourse. In *Computational Linguistics*.
- [Ho vd., 2006] Measures of Geometrical Complexity in Classification Problems. Data Complexity in Pattern Recognition, Springer-Verlag, 3-25
- [Hobbs, 1978] Resolving pronoun references. *Lingua* 44, 339-352.
- [Holen, 2007] Automatic Anaphora Resolution for Norwegian. Anaphora: Analysis, Algorithms, and Applications. 6th DAARC 2007 Selected Papers, Springer.
- [Kennedy ve Boguraev, 1996] Anaphora for everyone: pronominal anaphora resolution without a parser Proceeding of EMNLP-2, Providence, Rhode Island, USA.
- [Küçük, 2005] A Knowledge-Poor Pronoun Resolution System For Turkish. MSc Thesis. The Graduate School Of Natural And Applied Sciences, Middle East Technical University
- [Küçük ve Turhan, 2007] A knowledge-poor pronoun resolution system for turkish. In *6. Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*.

- [Landis ve Koch, 1977] The measurement of observer agreement for categorical data. *Biometrics*. Vol.33, 159-174.
- [Lappin ve Leass, 1994] An algorithm for pronominal anaphora resolution. In *Computational Linguistics*.
- [McCarthy ve Lehnert, 1995.] Using decision trees for coreference resolution. In *IJCAI*, 1050-1055.
- [Mitkov, 1998] Robust pronoun resolution with limited knowledge. In *COLING-ACL*.
- [Mitkov, 2002] *Anaphora Resolution*. Pearson Education.
- [Oktar ve Yağcıoğlu, 1997] *Türkçe Söylem Yapısı ve Artgönderim*. 8. Uluslararası Türkdilbilimi Konferansı Bildirileri
- [Quinlan, 1993.] C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
- [Soon vd., 2001.] A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4), 521-544.
- [Tın ve Akman, 1992] Anaphora in Turkish: A Computational Framework In: *Proceedings of 6th International Turkish Linguistics Conference*, Anadolu University, Eskişehir.
- [Tın ve Akman, 1994] Situated Processing of Pronominal Anaphora In: *Proceedings of Second Conference for Natural Language Processing (KONVENS'94)*, University of Vienna, Austria, 369 ? 378.
- [Turan, 1996] *Null Vs. Overt Subjects in Turkish Discourse: A Centering Analysis, Ph.D. Dissetation*. University pf Pennysylvania, Philadelphia.

- [Turhan-Yöndem ve Şehitoğlu, 1997] Resolution of Dropped Pronouns In the Proceedings of the ESSLLI'97 Student Session.
- [Tüfekci ve Kılıçaslan, 2005] A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm. WEC (5), 13-17.
- [Tüfekci ve Kılıçaslan, 2007] Syntax-based pronoun resolution system for turkish. In *6. Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007*.
- [Vapnik ve Lerner, 1963] Patter Recognition using Generalized Portrait Method. Automation and Remote Control. 24:6
- [Yıldırım ve Kılıçaslan, 2005] A computational model for anaphora resolution in turkish via centering theory. In *IJCI2005*.
- [Yıldırım ve Kılıçaslan, 2007] A machine learning approach to personal pronoun resolution in turkish. In *FLAIRS20*. AAAI.
- [Yüksel ve Bozşahin, 2002] Contextually Appropriate Reference Generation Natural Language Engineering 8 (1). Cambridge University Press. 69-89