Bürünsel, Sözcüksel ve Biçimbilgisel Bilgiyi Kullanan Co-Training ile Türkçe Konuşma Dilinin Otomatik Cümle Bölütlemesi

Program Kodu: 1001

Proje No: 111E228

Proje Yürütücüsü:
Doç. Dr. Ümit Güz

Araştırmacı(lar):
Doç. Dr. Hakan Gürkan

NİSAN 2015
İSTANBUL

Önsöz

Bu sonuç raporu, TÜBİTAK - ARDEB - EEEAG - 1001 - Bilimsel ve Teknolojik Araştırma Projelerini Destekleme Programı tarafından desteklenen 111E228 numaralı, “Bürünsel, Sözcüksel ve Biçimbilgisel Bilgiyi Kullanan Co-Training ile Türkçe Konuşma Dilinin Otomatik Cümle Bölütlemesi (Co-training using Prosodic, Lexical and Morphological Information for Automatic Sentence Segmentation of Turkish Spoken Language)” başlıklı proje kapsamında yapılan çalışmaları içermektedir.

Projemize verdiği destekten dolayı TÜBİTAK Elektrik-Elektronik ve Enformatik Araştırma Grubu’na, çalışmalarımızın gerçekleştirilmesine olanak sağlayan Işık Üniversitesi’ne ve emeklerinden dolayı Proje Araştırmacısı Doç. Dr. Hakan Gürkan’a, bu proje kapsamında tez danışmanlıklarını yürüttüğümüz yüksek lisans ve doktora öğrencimiz Doğan Dalva’ya (Yüksek Lisans Tezi: Automatic Speech Recognition for Turkish Spoken Language, Doktora tezi: Co-training using Prosodic, Lexical and Morphological Information for Automatic Sentence Segmentation of Turkish Spoken Language, 2014-devam ediyor), yüksek lisans öğrencimiz İzel D. Revidi’ye (Prosodic, Morphological and Lexical Feature Extraction of Turkish Broadcast News Data, 2013-2014) ve projemizi olumlu eleştirileri ve yönlendirmeleri ile daha nitelikli hale getirilmesinde katkıda bulunan proje izleyicisi değerli hocamıza ve proje sorumlusu EEEAG grup uzmanı sayın Dr. Onur Jane’ ye teşekkür ederiz.

TÜBİTAK

İçindekiler

TÜBİTAK

Şekil Listesi

TÜBİTAK

Tablo Listesi

Sözlük

10-kez çapraz doğrulama	10-fold cross validation
Açık kaynak kodlu araçlar/yazılımlar	Open source tools/softwares
Adlandırılmış varlık	Named entity
Adlandırılmış varlık tanıma	Named entity recognition
Akış şeması	Flow chart
Alan dışı	Out-of-domain
Alan içi	In-domain
Alan/bölge	Domain
Anlamsal	Semantic
Ayrık saklı markov modelleri	DHMMs - Discrete Hidden Markov Models
Ayrışık özellik seti	Disjoint feature set
Bağlamsal	Contextual
Bakış	View
Başarım değerlendirme ölçütleri	Performance evaluation metrics
Beklenti enbüyükleme	EM-Expectation maximization
Belirsizlik	Ambiguity
Biçimbilgisel	Morphological
Biçimbilgisel açıklaştırma	Morphological disambiguation
Biçimbirim/anlambirim	Morpheme
Bilgi çıkarımı	Information extraction
Bilginin geri kazanımı	Information retrieval
Bilinmeyen	Unknown
Bir önceki kelime	Previous word
Bir sonraki kelime	Next word
Bitişken biçimbilgisi	Agglutinative morphology
Bürün	Prosody
Bürünsel/ezgisel	Prosodic
Cümle bölütleme	Sentence segmentation
Cümle öğe etiketleri	Part of Speech Tags
Cümle öğe sıralamasının serbest olarak kullanılabildiği dil	Free-constituent language
Cümle öğeleri	Part of speech
Cümle sınırı	Sentence boundary
Cümle sınırı değil	Non-sentence boundary
Çağrı merkezi	Call center
Çerçeve	Frame/segment
Çerçeve zaman işaretleri	Segment time marks
Çoklu zayıf sınıflandırıcılar	Multiple weak classifiers
Çözümleme	Parsing
Değişinti	Variance
Destek vektör makineleri	Support vector machines
Doğal özellik seti	Natural view set
Doğrusal	Linear

Duraksama süresi	Pause duration
Eğitim veri seti	Training data set
Elle etiketlenmiş veri	Manually labeled data
Enbüyüklemek	Maximization
Enerji dağılımı	Energy distribution
Eşik	Threshold
Etiketleme	Annotation
Etiketlenmemiş veri	Unlabeled data
Fazlaca yeterli	Redundantly sufficient
Geliştirim veri seti	Development data set/Held out data set
Gerçek durak	Real pause
Gerçek pozitif	True pozitive
Gereksiz	Redundant
Gereksiz yeterlilik	Redundant sufficiency
Gereksizlik	Redundancy
Geri getirme	Recall
Grafiksel kullanıcı arayüzü	Graphical User Interface
Gürbüzlük	Robustness
Güvenilir etiketlenmiş örnekler	Confidently labeled examples
İki durumlu biçimbilgisel özellik	Binary morphological feature
İlgilenilen kelime	Current word
İşlenmemiş	Raw
Karar ağaç yapıları	Decision trees
Kelime	Word
Kelime anlam açıklaştırma	Word sense disambiguation
Kelime dizisi	Word stream
Kelime güvenilirlik değerleri	Word confidence scores
Kelime ifade belirsizliği	Word sense ambiguity
Kelime karışıklık ağı	Word confusion network
Kendi kendine eğitme	Self-training
Kesinlik	Precision
Kod taslağı	Pseude code
Konu bölütleme	Topic segmentation
Konuşma akışı	Speech stream
Konuşma tanıyıcısı	Speech recognizer
Konuşmacı	Speaker
Konuşmacı ayrımı	Speaker diarization
Konuşmacıdan bağımsız	Speaker independent
Konuşmanın bir bölümünün etiketlenmesi	Part of speech tagging
Koşulsal bağımsızlık	Conditionally independent
Kök	Root
Makine ile dilden dile çeviri	Machine translation
Makine öğrenmesi	Machine learning
Makine tarafından etiketlenmiş örnekler	Machine-labeled examples
Metin özetleme	Text summarization

Metine çevirme/çevriyazı	Transcription
NIST hata oranı	NIST error rate
Noktalama	Punctuation
Otomatik etiketlenmiş veri	Automatically labeled data
Otomatik konuşma tanıma	Automatic speech recognition
Öğreticili öğrenme algoritması	Supervised learning algorithm
Önceki	Prior
Örnek	Example
Özellik	Feature
Özellik çıkarımı	Feature extraction
Özellik gözlemleri	Feature observations
Özne-nesne-yüklem	SOV-subject-object-verb
Perde	Pitch
Perde dağılımı	Pitch distribution
Saklı Markov Modeli	Hidden Markov Model
Serbestiyet derecesi	Degree of freedom
Sesbirim	Phoneme
Sesli	Voiced
Sessiz	Unvoiced
Seyrek veri setleri	Sparse data set
Sınıflandırıcı	Classifier
Sondan ekleme	Suffixation
Sonsal olasılık	Posterior probability
Söylem	Discourse
Sözcüksel	Lexical
Sözdizimsel	Syntactic
Sözdizimsel çözümleme	Syntactic parsing
Temel frekans	Fundamental frequency
Test seti	Test set
Tonlamaya ilişkin-çekimsel	Inflectional
Toplantı verisi	Meeting data
Türetsel	Derivational
Uzlaşma	Agreement
Uzlaşmama	Disagreement
Veri	Data
Veri toplama	Data collection
Yanlış negatif	False negative
Yanlış pozitif	False positive
Yarı öğreticili öğrenme	Semi-supervised learning
Yayın	Broadcast
Yayın haberleri	Broadcast news
Yeniden sıralama	Reranking
Yüksek güvenilirlik değerleri	Highest confindence scores
Zenginleştirme	Enriching

Dictionary

10-fold cross validation	10-kez çapraz doğrulama
Agglutinative morphology	Bitişken biçimbilgisi
Agreement	Uzlaşma
Ambiguity	Belirsizlik
Annotation	Etiketleme
Automatic speech recognition	Otomatik konuşma tanıma
Automatically labeled data	Otomatik etiketlenmiş veri
Binary morphological feature	İki durumlu biçimbilgisel özellik
Broadcast	Yayın
Broadcast news	Yayın haberleri
Call center	Çağrı merkezi
Classifier	Sınıflandırıcı
Conditionally independent	Koşulsal bağımsızlık
Confidently labeled examples	Güvenilir etiketlenmiş örnekler
Contextual	Bağlamsal
Current word	İlgilenilen kelime
Data	Veri
Data collection	Veri toplama
Decision trees	Karar ağaç yapıları
Degree of freedom	Serbestiyet derecesi
Derivational	Türetsel
Development data set/Held out data set	Geliştirim veri seti
DHMMs - Discrete Hidden Markov Models	Ayrık saklı markov modelleri
Disagreement	Uzlaşmama
Discourse	Söylem
Disjoint feature set	Ayrışık özellik seti
Domain	Alan/bölge
EM-Expectation maximization	Beklenti enbüyükleme
Energy distribution	Enerji dağılımı
Enriching	Zenginleştirme
Example	Örnek
False negative	Yanlış negatif
False positive	Yanlış pozitif
Feature	Özellik
Feature extraction	Özellik çıkarımı
Feature observations	Özellik gözlemleri
Flow chart	Akış şeması
Frame/segment	Çerçeve
Free-constituent language	Cümle öğe sıralamasının serbest olarak kullanılabildiği dil
Fundamental frequency	Temel frekans
Graphical User Interface	Grafiksel kullanıcı arayüzü
Hidden Markov Model	Saklı Markov Modeli

Highest confindence scores	Yüksek güvenilirlik değerleri
In-domain	Alan içi
Inflectional	Tonlamaya ilişkin-çekimsel
Information extraction	Bilgi çıkarımı
Information retrieval	Bilginin geri kazanımı
Lexical	Sözcüksel
Linear	Doğrusal
Machine learning	Makine öğrenmesi
Machine translation	Makine ile dilden dile çeviri
Machine-labeled examples	Makine tarafından etiketlenmiş örnekler
Manually labeled data	Elle etiketlenmiş veri
Maximization	Enbüyüklemek
Meeting data	Toplantı verisi
Morpheme	Biçimbirim/anlambirim
Morphological	Biçimbilgisel
Morphological disambiguation	Biçimbilgisel açıklaştırma
Multiple weak classifiers	Çoklu zayıf sınıflandırıcılar
Named entity	Adlandırılmış varlık
Named entity recognition	Adlandırılmış varlık tanıma
Natural view set	Doğal özellik seti
Next word	Bir sonraki kelime
NIST error rate	NIST hata oranı
Non-sentence boundary	Cümle sınırı değil
Open source tools/softwares	Açık kaynak kodlu araçlar/yazılımlar
Out-of-domain	Alan dışı
Parsing	Çözümleme
Part of speech	Cümle öğeleri
Part of speech tagging	Konuşmanın bir bölümünün etiketlenmesi
Part of Speech Tags	Cümle öğe etiketleri
Pause duration	Duraksama süresi
Performance evaluation metrics	Başarım değerlendirme ölçütleri
Phoneme	Sesbirim
Pitch	Perde
Pitch distribution	Perde dağılımı
Posterior probability	Sonsal olasılık
Precision	Kesinlik
Previous word	Bir önceki kelime
Prior	Önceki
Prosodic	Bürünsel/ezgisel
Prosody	Bürün
Pseude code	Kod taslağı
Punctuation	Noktalama
Raw	İşlenmemiş
Real pause	Gerçek durak
Recall	Geri getirme

Redundancy	Gereksizlik
Redundant	Gereksiz
Redundant sufficiency	Gereksiz yeterlilik
Redundantly sufficient	Fazlaca yeterli
Reranking	Yeniden sıralama
Robustness	Gürbüzlük
Root	Kök
Segment time marks	Çerçeve zaman işaretleri
Self-training	Kendi kendine eğitme
Semantic	Anlamsal
Semi-supervised learning	Yarı öğreticili öğrenme
Sentence boundary	Cümle sınırı
Sentence segmentation	Cümle bölütleme
SOV-subject-object-verb	Özne-nesne-yüklem
Sparse data set	Seyrek veri setleri
Speaker	Konuşmacı
Speaker diarization	Konuşmacı ayrımı
Speaker independent	Konuşmacıdan bağımsız
Speech recognizer	Konuşma tanıyıcısı
Speech stream	Konuşma akışı
Suffixation	Sondan ekleme
Supervised learning algorithm	Öğreticili öğrenme algoritması
Support vector machines	Destek vektör makineleri
Syntactic	Sözdizimsel
Syntactic parsing	Sözdizimsel çözümleme
Test set	Test seti
Text summarization	Metin özetleme
Threshold	Eşik
Topic segmentation	Konu bölütleme
Training data set	Eğitim veri seti
Transcription	Metine çevirme/çevriyazı
True pozitive	Gerçek pozitif
Unknown	Bilinmeyen
Unlabeled data	Etiketlenmemiş veri
Unvoiced	Sessiz
Variance	Değişinti
View	Bakış
Voiced	Sesli
Word	Kelime
Word confidence scores	Kelime güvenilirlik değerleri
Word confusion network	Kelime karışıklık ağı
Word sense ambiguity	Kelime ifade belirsizliği
Word sense disambiguation	Kelime anlam açıklaştırma
Word stream	Kelime dizisi

Özet

Co-training, web sayfası sınıflandırması, kelime anlam açıklaştırma ve adlandırılmış varlık tanıma gibi pek çok sınıflandırma işlevinde başarı ile kullanılan oldukça etkili bir makine öğrenme algoritmasıdır. Co-training, elle etiketlenmiş eğitim veri setine, etiketlenmemiş büyük miktarlardaki veriyi belirli miktarlarda etiketleyerek katmak suretiyle öğreticili öğrenme algoritmalarının performansını arttıran bir yarı öğreticili öğrenme metodudur. Co-training algoritmaları etiketlenmiş giriş verisine ilişkin farklı bakışlar üzerinde eğitilmiş iki veya daha fazla sınıflandırıcının üretilmesi ve daha sonra bu sınıflandırıcıların etiketlenmemiş veriyi ayrı ayrı etiketlemesi için kullanıldığı algoritmalardır. Otomatik olarak en güvenilir biçimde etiketlenmiş örnekler daha sonra insanlar tarafından elle etiketlenmiş veriye katılmaktadır. Bu işlem pekçok defa devam ettirilmektedir. Bu projede konuşma verisine ilişkin bürünsel, sözcüksel ve biçimbilgisel bilgilerin bakış olarak kullanıldığı co-training ile cümle bölütlemenin gerçekleştirilmesi ele alınmıştır.

Cümle Bölütleme işlevi standart konuşma tanıyıcılarının çıkışından elde edilen işlenmemiş kelime dizisi biçimindeki veriyi zenginleştirmeyi amaçlayan bir işlemdir. Bu işlemin rolü, kelime dizisi biçiminde olan verinin cümle ünitelerine ayrılmasını sağlamaktır. Cümle Bölütleme konuşma anlamaya kadar olan süreçte ilk adımdır. Cümle bölütleme işlevi, çözümleme, makine çevirimi, bilgi çıkarımı gibi cümle bölütlemenin yapıldığının varsayıldığı konuşma işlemenin daha ileri uygulamaları için bir ön adım olarak gerçekleştirilmektedir. Cümle sınırları belirlendikten sonra bu cümleler üzerinde daha ileri düzeydeki sözdizimsel ve/veya anlamsal analizler gerçekleştirilebilmektedir.

Bu projede konuşma özellikleri (bürünsel, sözcüksel ve biçimbilgisel) ayrışık ve doğal özellik seti olarak ele alınmış ve bu özellik setlerinin co-training algoritması ile kullanılması ile baseline sistemin performansının arttırılmasına çalışılmıştır.

Ayrıca, co-training için uzlaşma ve uzlaşmama adı verilen farklı öğrenme stratejileri de araştırılmıştır. Buna ek olarak, self-combined adını verdiğimiz ve kendi kendine eğitme ile co-training yaklaşımlarının bir araya getirildiği bir yaklaşım da öne sürülmüştür.

Abstract

Co-training is a very effective machine learning technique that has been used successfully in several classification tasks like web page classification, word sense disambiguation, and named entity recognition. Co-training is a semi-supervised learning method that aims to improve performance of a supervised learning algorithm by incorporating large amounts of unlabeled data into the training data set. Co-training algorithms work by generating two or more classifiers trained on different views of the input labeled data that are then used to label the unlabeled data separately. The most confidently labeled examples of the automatically labeled data can then be added to the set of manually labeled data. The process may continue for several iterations. In this project, we have described the application of the co-training method for sentence segmentation where we used the prosodic, lexical and morphological information as the views of the data.

Sentence segmentation from speech is part of a process that aims at enriching the unstructured stream of words that are the output of standard speech recognizers. Its role is to find the sentence units in this stream of words. Sentence segmentation is a preliminary step toward speech understanding. It is of particular importance for speech related applications, as most of the further processing steps, such as parsing, machine translation and information extraction, assume the presence of sentence boundaries.

In this project, we consider the speech features (prosodic, lexical and morphological) as disjoint and natural feature sets or views and we try to improve performance of the baseline by using these feature sets with the co-training algorithm.

Furthermore we have tried to investigate the different learning strategies for the co-training such as agreement and disagreement. In addition to these strategies it has been proposed that a new approach that we called self-combined which is the mixed version of the self-training and co-training approaches.

1. GİRİŞ

1.1 Projenin Amaç ve Kapsamı

Bu projede genel olarak, Türkiye Türkçesinin bürünsel/ezgisel, sözcüksel ve biçimbilgisel özelliklerinin çıkarılması ve bu özelliklerin en etkin yarı öğreticili algoritmalardan biri olan Co-training ile cümle bölütlemesinde kullanılması amaçlanmıştır. Böylece Türkçe konuşma diline ilişkin yüksek performanslı bir cümle bölütleme sisteminin oluşturulması hedeflenmiştir.

Elde edilen sistem çok az etiketlenmiş veri ile yüksek miktarlarda etiketlenmemiş veriyi mümkün olabilecek en yüksek güvenilirlikte etiketleyerek zaman alıcı ve emek yoğun bir işlevi yerine getirdiği gibi, cümle bölütlemenin büyük bir doğrulukla yapılmış olması ile de daha ileri araştırma ve uygulamaların (konu bölütleme, özetleme, bilginin geri kazanımı vb.) başarımını da arttıracaktır.

Konuşma işaretlerinin otomatik konuşma sistemleri tarafından basit kelime dizilerine dönüştürülmesi sonucu konuşma ile ilgili bürünsel özellikler (zamanlama ve perde yapılarına ilişkin bilgiler örneğin duraklar, vurgular, duygunun aksettirilmesi) kaybolmaktadır. Tüm bu özelliklerden yoksun olan metnin gerek insanlar tarafından okunması gerekse makineler tarafından işlenmesi oldukça zordur. Tüm bu özelliklerin yeniden kazandırılması (zenginleştirilmesi), insanların okudukları metni daha doğal bir biçimde algılamasını ve makinelerin daha doğru bir biçimde işlem yapmasını sağlamaktadır. Özellikle konuşma işareti incelendiğinde, içerisinde varolan bürünsel bilginin konuşmanın doğal akışı sırasında oldukça sık kullanıldığı ve konuşmayı tek düzelikten arındırarak zenginleştirdiği görülmektedir. Yine konuşmanın doğal akışı sırasında bu bürünsel bilginin yada özelliklerin, bir cümlenin bitirilip yeni bir cümleye başlandığı, yada bir konudan başka bir konuya geçildiğinde de kendisini gösterdiği görülmektedir. Bu projenin en önemli amaçlarından biri, bu özelliklerin yeniden kazandırılması olduğu gibi bu özelliklerin yukarıda belirtilen cümle ve konu bölütlemesi için de kullanılmasını sağlamaktır.

Bu projenin en önemli özelliklerinden biri metin tabanlı çalışmalardan farklı olarak konuşma işaretinin bizatihi kendisinin kullanılacak olması başka bir deyiş ile konuşma tabanlı olmasıdır. Bu nedenle çok fazla miktarda ve içerik olarak çok çeşitli alanlardan konuşma verileri elde edilerek incelenmiş ve bu veriler kullanılarak Türkçenin bürünsel sözcüksel ve biçimbilgisel özellikleri çıkarılmıştır. Daha sonra Türkçenin bu özellikleri ile, özellikle bilginin

veri tabanlarından otomatik olarak taranması, bilgi çıkarımı, bilginin geri kazanımı, metin özetleme ve makine ile dilden dile çeviri gibi uygulamaların temel adımı olan cümle bölütleme için kullanılmıştır. Böylece gerek insanların yukarıda belirtilen uygulamaları kendi başlarına yapabilmek için gerekli olan emek, zaman ve maliyetleri azaltmak, gerekse makineler tarafından yapılacak daha ileri işlemler ve uygulamalar için çok daha az ve sınıflandırılmış verileri kullanmaları mümkün olabilecektir. Bilgi, bilgiye erişim, bilginin kullanımı ve iletişim alanındaki hızlı gelişim, özellikle iletişimden gelen doğası ve iletişim uygulamaları ile olan sıkı ilişkisi itibari ile konuşma teknolojilerine ve uygulamalarına yönelik araştırmaların hızla gelişimine yolaçmaktadır. Bu alanlarda özellikle İngilizce başta olmak üzere farklı dillerde uygulamaların geliştirilmesi, gerek yukarıda belirtilen nedenlerden dolayı bu alanlarda yapılması gereken araştırmaların bir gereksinim yada zorunluluk haline gelmesi, gerekse uluslararası alanda yaşanan rekabet ve söz sahibi olma nedenlerinden dolayı oldukça önem taşımaktadır. Ayrıca, bu ve benzer uygulamaların Türkçe için yapılması ise ayrı bir önem arzetmektedir.

Yukarıda genel amaçlarına ilişkin bilgilerin verildiği bu proje genel kapsamı itibari ile, Türkçenin kendine özgü bürünsel sözcüksel ve biçimbilgisel özelliklerinin çıkarılması ve bu özelliklerin daha ileri dil işleme uygulamalarında gerçekleştirilmesi gereken ilk adım olan cümle bölütleme uygulamalarında kullanılmasını kapsamaktadır.

Projenin ilk aşamasında Türkçenin bürünsel açıdan yapılacak analizleri sonucu bürünsel özelliklerin çıkarılmasının yanısıra Türkçeye ilişkin başka analizler (sözcüksel, sözdizimsel, biçimbilgisel) de yapılmakta ve bulunan özellikler ile bürünsel özelliklerin ilişkisi irdelenerek özellikle sınıflandırma aşamalarında cümle bölütleme uygulamalarında performansın arttırılması sağlanmaktadır. Böylece özellikle Türkçe için geliştirilmiş ve daha ileri dil işleme uygulamalarında kullanılabilecek bir altyapı hazırlanması sağlanmıştır. Ayrıca pek çok farklı kaynaktan elde edilen veriler ile Türkçe konuşmaya ilişkin veritabanları oluşturularak, uygulanan yöntemlerin gerektirdiği analizlerin yapılması sonucunda bürünsel ve diğer özellikler ve bunların birbirleri ile ilişkilerini içeren veritabanları oluşturulmuştur. Elde edilen özellik veritabanları ile Co-training yöntemine ilişkin geliştirilen farklı öğrenme stratejileri (kendi kendine eğitme, uzlaşma, uzlaşmama ve self-combined) cümle bölütleme üzerinde uygulanmış ve yöntemlerin tümüne ilişkin çeşitli başarım ölçütleri baz alınarak değerlendirmeler yapılmıştır. Böylece özellikle Türkçe konuşma verileri için cümle bölütleme uygulamalarına yönelik olarak başarımda etken olan yada en yüksek performansı veren co-training stratejileri ile bürünsel, sözcüksel ve biçimbilgisel özelliklerin ortaya çıkarılması sağlanmıştır.

Bu projenin en önemli kısmını; elde edilecek bürünsel, sözcüksel ve biçimbilgisel özelliklerin yanısıra bu özelliklerin yarı öğreticili bir algoritma olan co-training yöntemi ile cümle bölütleme üzerinde kullanılması oluşturmaktadır. Bu çalışma literatürde gerek Türkçe konuşma diline ilişkin özellik setlerinin co-training algoritması ile birlikte kullanılacak olması açısından gerekse co-training yaklaşımının Türkçe cümle bölütleme alanında uygulanması açısından ilk yapılan çalışma olma niteliğindedir.

Cümle bölütleme için kullanacağımız Co-training yönteminde ilk defa literatüre İngilizce dili için yaptığımız çalışmalarla kazandırdığımız farklı stratejiler kullanılmaktadır. Bu stratejiler kendi kendine eğitme dışındaki uzlaşma, uzlaşmama ve self-combined stratejileridir. Ayrıca elde ettiğimiz sonuçlar ile İngilizce dili için en iyi performansı veren strateji-özellik set(ler)i ile Türkçe dili için elde edilen en iyi strateji-özellik set(ler)i bulma olanağı ile birlikte bu iki dile ilişkin elde edilecek sonuçlar ile yeni analizlerin yapılması da mümkün olabilecektir. Bu projede önerilen yöntemler ile Türkçe konuşma verilerinin pekçok yönden analiz edilmesinin sağlanması ve önemli bulgulara ulaşılması amaçlanmıştır. Literatürde özellikle cümle ve konu bölütleme alanında yapılan araştırmalarda çoğunlukla dile ilişkin sözcüksel bilginin kullanılmasına yönelik yöntemlerin geliştirildiği görülmektedir. Bu projede ise cümle bölütleme için sözcüksel bilgi ile birlikte bürünsel ve biçimbilgisel bilgilerin de çıkarılması ve kullanılması sağlanmaktadır.

Özellikle Türkçe konuşma verilerine ilişkin özellik setlerinin çıkarılması ve bu özelliklerin cümle bölütleme gibi dil işleme alanında pek çok uygulamanın ilk adımına uygulanmış olması projenin en özgün yönlerinden birini oluşturmaktadır. Ayrıca belirtilen Türkçe konuşma veya audio işaretlerin oldukça başarılı bir dilden bağımsız otomatik konuşma tanıyıcısından (SRI Decipher) geçirilmiş olması önerilen yöntemin performansını ve güvenilirliğini arttıran bir özelliktir. Bu işlemler SRI Decipher konuşma tanıyıcısı üzerinde gerçekleştirildiği gibi bağımsız açık kaynak kodlu yazılımlar (Hidden Markov Toolkit (HTK)) üzerinde de gerçekleştirildiğinden gerek telif hakkı, gerekse daha sonradan yapılacak çalışmalarda oluşabilecek yazılım bağımlılığı sözkonusu olmamaktadır. Ayrıca belirtilen açık kaynak kodlu yazılımların kullanılması, proje araştırmacıları ve başka araştırmacılar tarafından daha sonradan sistem üzerinde yapılabilecek geliştirimler ve farklı ileri uygulamalar (konu bölütleme vb.) için de uygun ve ortak kullanıma açık bir taban oluşturmaktadır.

Bu projenin bir kısmında özellikle Türkçenin biçimbilgisel yapısının analizi konusunda daha önce gerçekleştirilen araçlar da kullanılmaktadır. Bu projede, Türkçenin yapısından kaynaklanan nedenler ile ilgili olarak büyük miktarlarda özellik setleri ile çalışılacağından

karar ağaç yapıları yerine özellikle boosting sınıflandırıcılar tercih edilmiştir. Sınıflandırma işleminde özellik olarak kelime n-gramlar da kullanılmıştır. Ayrıca farklı bir yaklaşım olarak da cümle bölütleme işlemlerinde herbir cümle veya konu sınırı için sınır olma ve sınır olmama durumlarına göre ayrı ayrı sınıflandırmanın yapılması ve bu yapılırken de bağlamsal özelliklerin kullanılmasıdır. Bu amaçla özellikle boosting sınıflandırıcıları kullanılmaktadır.

Türkçe için gerçekleştirilen bu proje ile, elde edilen özellik veritabanları ve elde edilen cümle bölütleme yaklaşımlarının daha sonra tarafımızdan ve diğer araştırmacılar tarafından yapılacak daha ileri dil işleme uygulamaları için bir temel teşkil etmesi amaçlanmaktadır.

1.2 Temel Kavramlar

Aşağıdaki alt bölümlerde bundan sonraki bölümlerde değinilecek olan temel kavramlar ile yapılacak işlemlere yönelik olarak bilinmesi gereken bazı önbilgilere yer verilmektedir.

1.2.1 Otomatik Konuşma Tanıma (ASR- Automatic Speech Recognition)

Aşağıda yeralan örnekte de görüleceği üzere, insanlar tarafından gerçekleştirilen konuşma sinyalinden metine çevirme işlemi-çevriyazı sonunda elde edilen sonuç ile otomatik konuşma sisteminin (konuşma tanıyıcısının) girişinden uygulanan konuşma sinyalinin metine dönüştürülmesi sonucunda elde edilen çıkış arasında bazı farklılıklar oluşmaktadır. Bilindiği üzere günümüzde kullanılmakta olan otomatik konuşma tanıma sistemlerinin çıkışından elde edilen kelime dizileri biçimindeki metinler, noktalama işaretleri, büyük küçük harf ayrımı, paragraf başlangıcı ve bitişi, konuşmacı değişikliğini gösterir belirteçler gibi pekçok bilgiden eksik bir biçimde elde edilirler. Bütün bu eksiklikler aynı zamanda cümlelerin bölütlenmemiş olduğu gerçeğine de işaret eder.

İnsan çevriyazımı (human transcription):
but uh i'm i i i think that you know i mean we always uh i mean i've i've had a a lot of good experiences with uh with many many people especially where they've had uh extended family and i and an- i i kind of see that that you know perhaps you know we may need to like get close to the family environment and and get down to the values of you know i mean uh it's money seems to be too big of an issue wi- with with with with with what's going on today

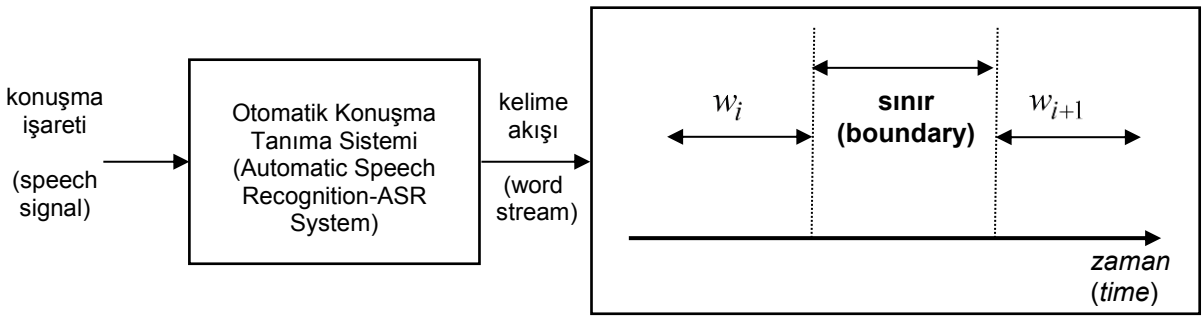
Otomatik Konuşma Tanıma sistemi çıkışı (ASR output/machine transcription):
but ~~um that that~~ {uh i'm i i} i think that you know ~~we~~ {i mean} we always uh i mean i've i've had ~~it there~~ {a} a lot of good experiences with ~~the~~ {uh} with many many people especially ~~with have~~ {where they've} had extended family ~~night~~ and i and {an- i} i kind of see that that you know perhaps you know we may need to like ~~you're~~ {get} close to the family environment and ~~in~~ {and} get down to the values of you know i mean ~~no and~~ {uh it's} money seems to be too big of an issue ~~we would~~ {wi- with with with} with with ~~really was we would~~ what's going on today

Çoğu zaman insanlar tarafından el ile yapılan çevriyazıların işlenmesinin ve okunmasının dahi zor olabileceği durumlar düşünüldüğünde makineler tarafından gerçekleştirilen bu eksikliğin öneminin oldukça büyük olduğu görülecektir. Özellikle insanlar tarafından konuşma işaretlerinin aslına uygun ve yukarıda belirtilen tüm eksiklikleri karşılayacak bir biçimde metne çevrilmesi, emek yoğun, dikkat gerektiren ve oldukça pahalı bir işlevdir. Bu nedenle bu işlevlerin makineler tarafından etkin bir biçimde yapılmasına yönelik çalışmalar hız kazanmıştır. Bu bağlamda, özellikle cümle bölütleme konusu bu alanda yapılan pek çok ileri araştırmaya (makine ile dilden dile çeviri, bilgi çıkarımı, konu belirleme, özetleme vb.) temel oluşturması ve bu araştırmaların ilk adımını oluşturması açısından oldukça önemlidir.

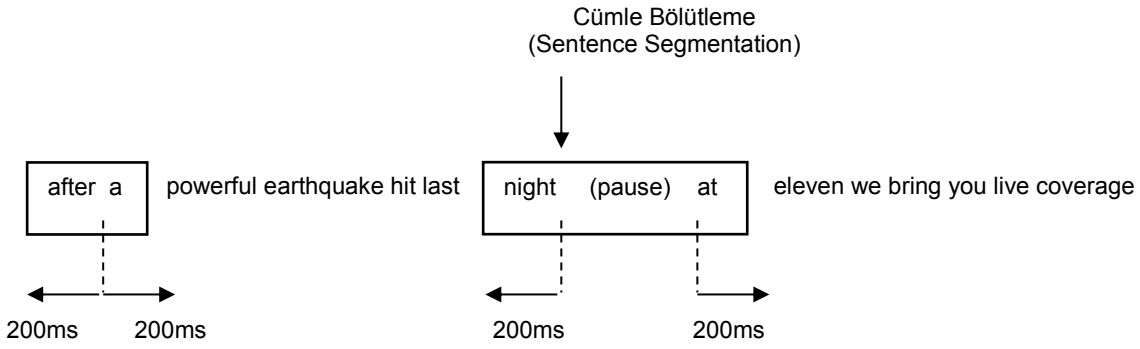
1.2.2 Cümle Bölütleme (Sentence Segmentation)

Cümle bölütlemesi başka bir ifade ile konuşmanın cümlelere ayrılması, standart otomatik konuşma tanıyıcılarından çıkan yada bu tanıyıcılar tarafından üretilen ve üzerinde bir işlem yapılmamış/işlenmemiş kelime akışının zenginleştirilmesi amacıyla yapılan bir işlemdir. Cümle bölütleme işlemine ilişkin blok şema Şekil 1.1 de verilmiştir. Cümle bölütleme, cümle sınırlarının önceden belirlenmiş olmasını gerektiren yada varolduğunu varsayan çözümleme, makine çevirimi, bilgi çıkarımı gibi konuşma işleme ile ilgili daha ileri işleme adımlarında özel bir öneme sahiptir. Burada adı belirtilen tüm ileri işleme adımları ve daha pek çok uygulamada cümle bölütlemesi ilk yapılacak işlemdir. Cümle bölütlemesi için farklı bilgi kaynakları dikkate alınmaktadır. En önemli bilgi kaynakları, otomatik konuşma tanıma biriminden gelen kelimeler dizisi ve komşu kelimeler arasındaki duraksama süreleridir. Şekil 1.2 de otomatik konuşma tanıyıcısından kelime dizisi olarak çıkan metinde, duraksama süresinin cümle sınırı tespitinde ipucu olarak kullanıldığı bir örnek yeralmaktadır.

Cümle bölütleme problemi, verilen bir özellik seti üzerinden sınıfın sonsal olasılık değerinin kestirimine göre herbir kelime sınırının bir sınıf etiketi (cümle sınırı olan veya cümle sınırı olmayan) ile ilişkilendirildiği bir ikili sınıflandırma problemi olarak ele alınabilir.

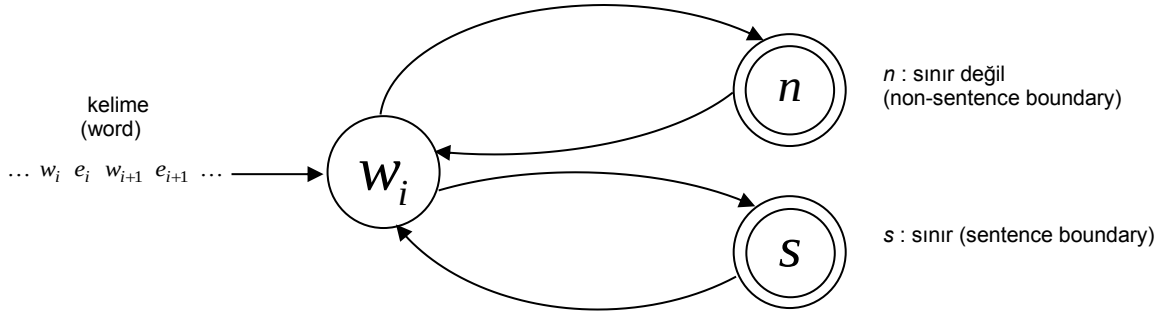


Şekil 1.1. Cümle bölütleme blok şeması



Şekil 1.2. Herbir kelime sınırı için özellik çıkarım bölgeleri

Cümle bölütleme için, sözcüksel ve duraksal özelliklerin birleşimi yada sözcüksel, duraksal veya diğer bürünsel özellikler kullanılabilir. Sınıflandırma için, boosting, maximum entropy ve destek vektör makineleri (SVM-support vector machines) sınıflandırıcıları gibi farklı sınıflandırıcıları kullanmak mümkündür. Ayrıca belirtmek gerekmektedir ki, cümle bölütlemesi konu bölütlemesi için gerekli olan ilk adımdır. Yukarıda değinildiği üzere, cümle bölütlemesi "cümle sınırı" (*s*) ve "cümle sınırı değil" (*n*) kararlarının sınıf olarak atandığı bir ikili sınır sınıflandırma problemi olarak ele alınabilir. Şekil 1.3 de bu probleme ilişkin Bigram model gösterilmektedir. Verilen bir kelime akışı veya dizisi ($\{w_1, \ ..., \ w_N\}$) için amaç, sınırlar ($\{e_1, \ ..., \ e_N\}, e \in \{s,n\}$) için sınıfların kestirimidir. Burada ($s_i, i = 1,..., N$), w_i ve w_{i+1} arasındaki sınırdır. Genellikle bu, sonsal olasılığın $P(s_i = k \mid o_i), \ k \in \{s,n\}$ kestirimi için sınıflandırıcının eğitilmesi ile yapılır. Burada o_i, ler s_i kelime sınırı için özellik gözlemleridir. İdeal durumda, sınıflandırıcının kararı, en yüksek olasılık değerine $P(s_i = k \mid o_i)$ sahip sınıftır. Bununla birlikte, cümle bölütleme işlevinde cümle sınırı için olasılık $P(s_i = s \mid o_i)$, bir eşik değeri ile karşılaştırılır. Bu eşik değerinin üzerindeki olasılık değerlerinde karar cümle sınırı olduğu yönünde, aksi durumda ise cümle sınırı olmadığı yönünde verilmektedir. Daha sonradan da görüleceği üzere, farklı sınıflandırıcılar ve farklı değerlendirme ölçütleri için en iyi (optimal) eşik değeri farklıdır.



Şekil 1.3. Bigram model

2. LİTERATÜR ÖZETİ

Türkiye'de, gerek belli başlı araştırma grupları, gerekse bireysel olarak özellikle doğal dil işleme, konuşma işleme, otomatik konuşma tanıma, konuşmacı tanıma, konuşmacı doğrulama, makine öğrenmesi, makine ile dilden dile çeviri yapılması gibi konularda Boğaziçi Üniversitesi, Sabancı Üniversitesi, Ortadoğu Teknik Üniversitesi, Bilkent Üniversitesi, Koç Üniversitesi ve diğer üniversitelerde araştırmalar yapılmaktadır. Özellikle Türkçe Dil İşleme konusunda da yukarıda belirtilen üniversitelerlerde ve bazı özel araştırma merkezleri ve şirketlerde de çalışmalar yapılmakta ve uygulamalar geliştirilmektedir. Özellikle doğal dil işleme konusunda dünyadaki çalışmaların yoğunlaşmasına paralel olarak Türkiye'de bu konu önem kazanmaktadır.

Türkçenin biçimbilimsel ve dilbilimsel yapısına ilişkin belli başlı çalışmalar mevcut olmakla birlikte, özel olarak Türkçenin bürünsel özelliklerinin çıkarılması ve kullanılması konusunda yapılan çalışmalar oldukça sınırlıdır.

2.1 Değişik Dillere İlişkin Bürünsel ve Biçimbilgisel Özelliklerin Çıkarılması ve Cümle Bölütleme Konusunda Yapılan Çalışmalar

Türkçenin biçimbilimsel yapısı ile ilgili olarak Türkçe kelime yapılarının iki tabanlı biçimbilgisel açıklanması, Türkçe metinler için derlem etiketleyicileri, Türkçede cümlelerin ayrılmasında kullanılan yapılar gibi metin tabanlı çalışmalar Oflazer (1994) ve Oflazer vd. (1995) tarafından gerçekleştirilmiştir.

Türkçe ile ilgili dilbilimsel olarak bazı bürünsel özelliklerin incelendiği Oskay vd. (2001), Levi (2001), Levi (2002), Sayli ve Arslan (2003) ve Levi (2005) tarafından yayınlanmış yayınlar bulunmaktadır.

Literatürde, özellikle dünya üzerinde birçok insanın konuştuğu ortak bir dil olması nedeniyle İngilizce ve yine sayı olarak çok fazla insanın konuştuğu Arapça ve Çince dillerinde cümle bölütleme ve konu bölütleme ile bürünsel özelliklerin kullanıldığı bazı çalışmalar mevcuttur.

Konuşma akışının kendisinin ele alındığı cümle bölütleme konusundaki ilk çalışmalar konuşma işaretinin kendisinde varolan ancak yazıya uyarlanamayan özellikle vurgu, durak gibi özelliklerin kullanılmasına yöneliktir. Hirschberg ve Grosz (1992) vurgu ve söylem yapısı arasındaki ilişki konusunda çalışmış ve vurgu özellikleri ile etiketlenen söylem yapıları arasında bir bağ olduğunu ortaya koymuştur.

Hirschberg ve Nakatani (1996), söylemin, metnin ve konuşmanın birlikte kullanılması ile yapılan bölütlemenin sadece metin kullanılarak yapılan bölütlemeden daha iyi olduğunu göstermişlerdir.

Özellikle konuşma işaretlerinde cümle sınırlarının tespiti (ve benzer biçimde noktalama, işaretleme) otomatik konuşma tanıyıcısı çıkışının zenginleştirilmesi olarak ele alınmış ve bu konuda Shriberg vd. (2000), Gotoh ve Renals (2000), Liu vd. (2005) ve Roark vd. (2006) tarafından çalışmalar yapılmıştır. Önceki yaklaşımlarda bu işlev için hem metin hem de bürünsel özelliklerin yararlanıldığı farklı sınıflandırıcılar (örneğin, saklı Markov modelleri (HMM), maximum entropy gibi) kullanılarak değerlendirmeler yapılmıştır. Bu konuda yapılan diğer bir proje ise Amerikan Savunma Bakanlığı'na bağlı çalışan İleri Savunma Araştırmaları Projeleri Ajansı (DARPA-Defense Advanced Research Projects Agency) EARS (Effective, Affordable, Reusable Speech-to-Text) programı için geliştirilmiştir. Bu projede, konuşmanın, otomatik bir biçimde üretilen cümle sınırları, konuşmacının konuşmasına devam ederken durakladığı ve bir sonraki kelime veya kalıbı düşünürken arada çıkardığı “uhh, ah” gibi sesleri ve doldurulmuş kelimeler gibi yapısal bilgileri de içeren zengin bir biçimde yazıya dönüştürülmesi konusunda özel çabalar harcanmıştır.

Liu vd. (2005) tarafından yapılan çalışmada hem telefon konuşma verileri üzerinde hem de radyo haber yayınlarına ilişkin konuşma verileri üzerinde farklı modelleme yaklaşımları (saklı Markov modelleri, maximum entropy ve conditional random fields) ve çeşitli bürünsel ve metinsel özellikler değerlendirilmiştir. Roark vd. (2006) tarafından yapılan çalışmada bir yeniden sıralama tekniği geliştirilmiş ve referans olarak kullandığı Liu vd. (2005) tarafından verilen yöntemden daha yüksek bir cümle sınırı belirleme performansı göstermiştir.

Cümle bölütleme yöntemleri telefon konuşma verilerine uygulandıktan ve olumlu sonuçlar alındıktan sonra benzer yöntemler Ang vd. (2005) ve Zimmermann vd. (2005) tarafından bu defa içerisinde birçok kişinin yeraldığı toplantı konuşma derlemleri için de uygulanmıştır.

Cümle sınırlarının bulunduğu ve çoğunlukla metin tabanlı benzer yaklaşımlar Kolar vd. (2004) tarafından saklı Markov modellerinin kullanılması ile Çekçeye, Zong ve Ren (2003) ve Liu ve Zong (2003) tarafından maximum entropy sınıflandırıcıları kullanılarak da Çinceye uyarlanmışlardır.

Son zamanlarda yapılan çalışmalarda destek vektör mekineleri, boosting ve maximum entropy sınıflandırıcılar ile entropy tabanlı sınıflandırıcıların birleştirilmeleri ve bunların

çıkışlarının da saklı dil modelleri ile bütünleştirilmeleri ile İngilizce, Arapça ve Çince için oldukça belirgin geliştirimler sağlanmıştır.

Konu bölütleme ile ilgili önceki çalışmalar, genellikle sözcüksel özelliklerin kullanımında yoğunlaşmıştır. Buna örnek olarak kelime benzerliklerinin kullanıldığı çalışmalar Kozima (1993), ipucu kalıplarının kullanıldığı çalışmalar Passonneau ve Litman (1997), sözcüksel pencerelerin kosinüs benzerliklerinin kullanıldığı çalışmalar Hearst (1997) ve adaptif dil modellemenin kullanıldığı çalışmalar Beeferman vd. (1999) verilebilir.

Shriberg vd. (2000) ve Tür vd. (2001) tarafından radyo haber yayınlarına ilişkin konuşma verilerinin konu bölütlemesi alanında başka bir deyiş ile akustik özelliklerin kullanıldığı konu bölütlemesi konusunda çalışmalar yapılmıştır.

Tür vd. (2001) tarafından konuşmanın konu ünitelerine otomatik olarak bölütlenmesi için hem bürünsel hem de sözcüksel özelliklerin kullanıldığı olasılıksal bir model sunulmuştur. Bu yaklaşım, saklı Markov modelleri, istatistiksel dil modellerini ve bürün-tabanlı karar ağaç yapılarını birleştiren bir yaklaşımdır. Sözcüksel bilgi, konuşma tanıyıcısından, bürünsel özellikler ise otomatik bir biçimde konuşma dalga biçimlerinden elde edilmektedir. Bu çalışmanın sonucunda, bürünsel modelin tek başına kelime tabanlı bölütlemeden üstün olduğu ve hem bürünsel hem de kelime tabanlı bilgi kaynaklarının birleştirilmesi ile çok daha iyi bir model oluşturulabileceği gösterilmiştir.

Bürünsel özelliklerin kelime bilgisi ile birleştirilerek Tür vd. (1999) tarafından bilgi çıkarımı ve Stolcke vd. (1999) tarafından konu bölütlemesinde kullanıldığı çalışmalar da mevcuttur.

2.2 Yarı Öğreticili Öğrenme Alanındaki Çalışmalar

Yarı öğreticili öğrenmenin amacı istatistiksel modelleri eğitmek için gerekli olan veri miktarını azaltmaktır. Gerçek yaşam uygulamalarının çoğunda verilerin toplanması etiketlenmesinden daha kolaydır. Örneğin, yayın haberlerinden audio işaretlerin kaydedilmesi, yazılması ve bazı etiketler ile etiketlenmesinden daha kolaydır.

Yarı öğreticili öğrenme yöntemleri, az miktarda etiketlenmiş veri ve oransal olarak daha yüksek miktarda etiketlenmemiş verinin bulunduğunu varsayar. Bu durumda amaç etiketlenmemiş veriyi kullanarak sistemin başarımını geliştirmektir. Bütün yarı öğreticili öğrenme yöntemlerinde öncelikle varolan etiketlenmiş veri kullanılarak bir başlangıç sınıflandırıcı eğitilmektedir. Daha sonra temel düşünce bu sınıflandırıcının etiketlenmemiş

veriyi kullanarak etiketleme yapması ve makine tarafından etiketlenmiş bu yeni örneklerin sınıflandırıcı başarımını arttırmasıdır. Bu işlemin yinelemeli ve her yinelemede az miktarda etiketlenen örneklerin katılması ile yapılması etiketlenmemiş verinin artımsal bir biçimde kullanılmasını sağlamaktadır. Nigam ve Ghani (2000) tarafından yapılan çalışmada bu durumun etiketlenmemiş verinin tümünün kullanılması işleminden çok farklı olduğu ve modellerin davranışını belirgin bir biçimde değiştirdiği gösterilmektedir.

2.2.1 Kendi Kendine Eğitme (Self-training) Konusunda Yapılan Çalışmalar

Nigam ve Ghani (2000) tarfından yapılan çalışmada da gösterildiği üzere kendi kendine eğitme en popüler yarı öğreticili öğrenme metodlarından biridir. Kendi kendine eğitmede verilen model verinin etiketlenmemiş kısmının sınıflarını kestirmektedir. Daha sonra otomatik olarak sınıflandırılan/etiketlenen örnekler eğitme setine dahil edilmekte ve model yeniden eğitilmektedir. Bu işlem yinelemeli bir biçimde sürdürülmektedir. Yanlış bir biçimde sınıflandırılmış örnekler tarafından getirilen gürültünün ortadan kaldırılması için sadece yüksek güvenilirlik ile sınıflandırılmış örnekler kullanılır.

Kendi kendine eğitme, tipik olarak konuşma ve konuşmacı işleme sistemleri tarafından kullanılan öğreticili olmayan model adaptasyonu ile çok yakından ilgilidir. Gauvain ve Lee (1994) tarafından yapılan çalışmada oldukça popüler bir adaptasyon yaklaşımı olan maximum aposteriori adaptation (MAP) denenmiştir. Bazı basitleştirmeler ile MAP adaptasyonu önceki modelin ağırlıklandırılmış doğrusal interpolasyonu na indirgenebilmiş ve model otomatik olarak sınıflandırılmış örnekler ile eğitilmiştir. Bu kendi kendine eğitmeden başka bir şey değildir. Bu yaklaşım öğreticisiz akustik ve dil modeli adaptasyonu ile konuşmacı adaptasyonuna uygulanmıştır. Bacchiani ve Roark (2003) yinelemeli öğreticisiz dil modeli adaptasyonunu sesli email çevriyazılarına uygulamışlardır. Tür vd. (2004) öğreticisiz dil modeli adaptasyonunu yeni bir çağrı merkezi diyalog sistemi için uygulamıştır. Gretter ve Riccardi (2001) öğreticisiz dil modeli adaptasyonu boyunca kelime karışıklık ağlarından elde edilen kelime güvenilirlik değerlerini kullanmışlardır.

Doğal dil işleme için kendi kendine eğitme, Wang vd. (2007), Mihalcea (2004) ve McClosky vd. (2006) tarafından yapılan çalışmalarda, konuşmanın bir bölümünün etiketlenmesi, kelime anlam açıklaştırma ve sözdizimsel çözümleme konularına uygulanmıştır.

2.2.2 Co-training Konusunda Yapılan Çalışmalar

Geleneksel tek bakış veya özellik kullanan makine öğrenme konseptinin tersine çok bakışlı yaklaşımın özellik seti her biri ayırtedilebilir ve herbiri kendi kendine öğrenebilir yeterlikte iki

veya daha fazla farklı bakıştan oluşmaktadır. Bu iki yaklaşımın temel farkı çok bakış kullanan yöntemde bunlar birbirini eğitirken tek bakışlı yöntemde algoritma kendi kendini eğitir.

Co-training oldukça etkili bir yarı öğreticili makine öğrenme algoritmasıdır. Bu algoritma göreceli olarak az miktarda olan etiketlenmiş bir veriden yararlanarak etiketlenmemiş çok miktardaki veriyi azar azar etiketleyen ve bunu yaparken de çoklu zayıf sınıflandırıcıları kullanan bir algoritmadır. Co-training yaklaşımındaki temel varsayım, sözgelimi probleme özgü olarak kullanılan iki adet zayıf sınıflandırıcıdan birinin, diğer sınıflandırıcının etiketlemeyi başaramadığı örnekleri etiketlemesidir. Böylelikle sınıflandırıcılar karşılıklı olarak birbirlerini eğitmektedirler. Cümle bölütleme probleminin, co-training yöntemine ilişkin iki temel gerekliliği karşıladığı gerçeği değerlendirildiğinde, bu yöntem için oldukça uygun bir problem olduğu görülecektir. Co-training metodunun uygulanması için sağlanması gereken iki gereklilik veya koşul, veri setinin iki ayrı ve doğal bakış veya özellik setinden oluşmasıdır. Bu projede, cümle bölütlemesi için kullanılan iki doğal bakış veya özellik seti; bürünsel, sözcüksel ve biçimbilgisel özellik setlerinin ikili kombinasyonlarıdır.

Co-training yaklaşımı ilk defa Blum ve Mitchell (1998) ve Mitchell (1999) tarafından öne sürülmüş ve uygulanmıştır. Yöntemdeki ana amaç etiketlenmemiş veri ile birlikte çoklu bakış kullanarak az sayıda etiketlenmiş veriden yola çıkarak etiketlenmiş veri sayısını arttırmak veya genişletmektir. Daha özel olarak, herbir örneğin çoklu belirli bakışlarının varlığı aynı görev için farklı modelleri eğitmek için kullanılabilir ve daha sonra herbir sınıflandırıcının etiketlenmemiş örnekler üzerindeki kestirimleri diğer sınıflandırıcının eğitim setinin genişletilmesinde kullanılabilir. Blum ve Mitchell' in gerçekleştirdiği işlem internet ortamında birçok bilgisayar mühendisliği bölümlerinden toplanmış çok sayıdaki web sayfasından alınan akademik ders sayfalarını tanımlayabilmekti. Blum ve Mitchell'in co-training uygulamasında iki doğal özellik seti bulunmaktaydı. Bunlar; ders web sayfasında bulunan kelimeler ve bu web sayfalarına işaret eden linklerde kullanılan kelimelerdi. Bunlara "bakış (view)" adını vermişlerdi. Bu işlem için her iki örnekler bakışı da öğrenme için yeterli varsayılmışlardır. Blum ve Mitchell teorilerinde, iki bakış sınıflandırma için tekbaşlarına yeterli ve verilen sınıf içerisinde koşulsal olarak bağımsızlarsa, co-training in "probably approximately correct (PAC) learnable" olduğunu göstermişlerdir. Blum ve Mitchell'in sonuçları birleştirilmiş sınıflandırıcıların hata oranının %11 den %5 e düşürüldüğünü göstermiştir.

Birçok farklı alanda ve uygulamalarda co-training algoritmasının performansı ve etkisinin arttırılması üzerinde çaba sarfedilmiş ve araştırmalar yapılmıştır. Abney (2002) trafından yapılan çalışmada bağımsız olma varsayımı yada koşulunun çözülebileceği yada serbest

bırakılabileceği ve daha zayıf bağımsızlık koşulu altında dahi co-training algoritmasının etkili olabileceği gösterilmiştir. Bu çalışmada, etiketlenmemiş veri üzerinde uzlaşmayı enbüyüklemek için bir greedy algoritmasının işletilmesi temel alınmıştır. Bu yöntem adlandırılmış varlık sınıflandırılması için yapılan co-training deneylerinde sonuçların geliştirilmesine neden olmuştur. Bu çalışmada gösterilmiştir ki, zayıf bağımsızlıkla iki sınıflandırıcı arasındaki uzlaşmama oranı co-training hata oranındaki en üst sınırdır.

Kiritchenko ve Matwin tarafından yapılan çalışmada, co-training yaklaşımı e-mail sınıflandırma problemine uygulanmıştır. Bu çalışmada, co-training algoritmasının performansının kullanılan öğrenme algoritmasının performansına duyarlı olduğu bulunmuştur. Naive Bayes ile yapılan co-training'in daha iyi bir performans ile sonuçlanmadığı gözlenmiştir. Bununla birlikte, bu durum destek vektör makineleri kullanıldığı durum ile aynı değildir. Yazarlar bu durumu, büyük miktarlardaki seyrek veri setleri ile ilgili olan Naive Bayes'lerin yetersizliğine bağlamışlardır. Bu açıklama, aynı zamanda özellik seçiminden sonra elde edilen belirgin bir biçimde iyileştirilmiş sonuçlar ile de doğrulanmıştır.

Nigam ve Ghani (2000) tarafından yapılan çalışmalar co-training yönteminin koşulsal bağımsızlık ve fazla yada gereksiz yeterlilik varsayımlarına olan duyarlılığının araştırılması ile ilgili olan çalışmalardır. İlk deneyde, co-training Blum ve Mitchell (1998) tarafından yapılan çalışmada yeralan web sayfası veritabanına uygulanmıştır. Bu çalışmanın sonucunda elde edilen sonuçlar göstermiştir ki, Naive Bayes'i kullanan co-training metodu özelliklerde doğal bir ayrım olsa dahi beklenti enbüyükleme (EM) den daha iyi sonuçlar vermemektedir. Hem beklenti enbüyükleme hem de Naive Bayes'i kullanan co-training algoritması başlangıç sınıflandırıcıyı %10 düzeyinde geliştirmekte yada iyileştirmektedir. İkinci deney, iki özellik setinin gerçek bir biçimde koşulsal olarak bağımsız olması amacına uygun olarak yari yapay olarak yaratılmış bir veritabanı üzerinde gerçekleştirilmiştir. Buna ek olarak, Naive Bayes çok küçük bir hata oranı üretebilmesi için herbir veri seti üzerinde ayrı ayrı eğitildiklerinden dolayı fazlaca yeterli koşulu da sağlanmış olmaktadır. Bu durumda, Naive Bayes'i kullanan co-training'in beklenti enbüyüklemeden ve daha da önemlisi etiketlenmiş tüm örnekler kullanılarak eğitilmiş Naive Bayes'den dahi iyi sonuçlar verdiği bulunmuştur. Üçüncü deneylerinde, co-training yöntemini özellik setlerinin doğal olarak ayrılmış olmadığı veri setleri üzerinde test etmişlerdir. Bu iki özellik seti, veri setlerine ilişkin tüm özelliklerin tamamen rastgele bir biçimde iki farklı gruba ayrılması veya atanması biçiminde gerçekleştirilmiştir. Bu işlem iki veri seti için denenmiştir. Birinci veri seti özelliklerin oldukça açık bir fazlalığa yada gereksizliğe sahip olduğu durum, diğeri gereksizlik düzeyinin belirsiz olduğu yada bilinmediği ve özelliklerde doğal bir ayrımın bulunmadığı duruma ilişkin veri

setleridir. Bu deneyden elde edilen sonuçlar göstermiştir ki, özellik setlerindeki fazla gereksiz bilginin varlığı, co-training algoritmasına beklenti enbüyükleme algoritması ile karşılaştırıldığında daha büyük bir avantaj vermektedir. Bu deneyler aynı zamanda beklenildiği üzere, co-training yönteminin koşulsal bağımsızlık ve gereksiz yeterlilik varsayımlarına oldukça bağımlı olduğunu doğrulamıştır. Bununla birlikte, co-training yönteminin bu iki varsayımı üzerinde bir zorlama yapılsa dahi co-training'in sınıflandırıcıların performanslarının geliştirilmesi yönündeki başarımının yine de oldukça iyi olduğu söylenebilir.

Wang vd. (2007) co-training için farklı bakışlar yerine farklı sınıflandırma algoritmalarını temel almışlardır. Örneğin etiketleme ve çözümleme için maximum entropy ve saklı Markov modelleri (HMM) kullanılmıştır. Co-training sonuçları kendi kendine eğitme sonuçları ile karşılaştırılmıştır. Benzer biçimde co-training Mihalcea (2004) tarafından kelime anlam açıklaştırma için kullanılmış ve kendi kendine eğitme ile karşılaştırılmıştır. Guz vd. (2007) Co-training'i bürünsel ve sözcüksel bilgiyi kullanarak İngilizce veriler için cümle bölütleme alanında uygulamışlardır.

3. VERİ TOPLAMA (DATA COLLECTION) VE ETİKETLEME (ANNOTATION)

Veri toplama ve etiketleme konusunda Boğaziçi Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, BUSİM Speech Processing Group[1] olanaklarından yararlanılmıştır. Özellikle öğretim üyesi Doç. Dr. Murat Saraçlar ve araştırma ekibinde bulunan Erinç Dikici ile proje yürütücüsünün yüksek lisans ve doktora tez danışmanlıklarını yaptığı İzel D. Revidi ve Doğan Dalva adlı öğrencilerden destek alınmıştır.

Söz konusu veriler Boğaziçi Üniversitesi'nde bulunan BUSİM Speech Processing Group tarafından kaydedilen ve belirli projelerde üzerinde çalışılmakta olunan, Amerika'nın Sesi (Voice of America) Türkçe Yayın Bölümü[2] tarafından hazırlanan Türkçe radyo yayını haberlerine ilişkin konuşma verilerinden oluşmaktadır. Bu veriler belirli günlerde kaydedilmiş herbiri 30 dakikalık ve wav formatında kaydedilmiş toplam 21 saat uzunluğunda olan 42 adet dosyadan oluşmaktadır. Dosyalarda farklı konuşmacılara ilişkin toplam 104458 kelime ve 6881 cümle sınırı bulunmaktadır. Sözü edilen konuşma verilerine ilişkin aşağıda detayları ile verilen değişik wav, stm ve ctm uzantılı formatlarda dosyalar hazırlanmıştır.

a) Amerika'nın Sesi Türkçe Yayın Bölümü'nün günlük sabah ve akşam saatlerinde yapmış olduğu haber yayınlardan kaydedilen konuşma verileri audio (wav) formatında 16kHz örnekleme frekansında ve 16bit doğrusal PCM de örneklenmişlerdir.

b) Audio verilerin STM (segment time marks) baska bir deyiş ile herbir zaman aralığına (çerçeve) ilişkin başlangıç ve bitiş zamanları ile ona karşı gelen yazılı metni içeren versiyonudur. Bu formatta hazırlanmış dosyalar aynı zamanda kaydedilen dosyanın adını, konuşmacıya ilişkin bilgileri (speaker-id, native-non native, male-female), konuşma kaydının arka plan bilgisi vb. bilgileri de içermektedir. Ayrıca bu dosyalar kişiler tarafından mümkün olabildiği ölçüde noktalama işaretlerine uygun olarak metne dönüştürülmüş dosyalardır. Referans çevriyazı dosyaları olarak da adlandırılan bu dosyalar otomatik konuşma tanıyıcısına ilişkin çıkış dosyası ile karşılaştırılarak otomatik konuşma sisteminin performansının ölçülmesinde de kullanılmaktadır.

[1] http://www.busim.ee.boun.edu.tr/

[2] http://www.voanews.com/turkish/

Aşağıda, belirtilen audio kayıtlara ilişkin hazırlanan stm uzantılı dosyalardan bir örnek verilmektedir.

```
;; Transcriber export by tstm.tcl,v 1.21 on Fri Nov 23 04:58:11 PM EET 2007 with encoding ISO-
8859-9
;; transcribed by , version 3 of 071105
;;
;; CATEGORY "0" "" ""
;; LABEL "O" "Overall" "Overall"
;;
;; CATEGORY "1" "Hub4 Focus Conditions" ""
;; LABEL "F0" "Baseline//Broadcast//Speech" ""
;; LABEL "F1" "Spontaneous//Broadcast//Speech" ""
;; LABEL "F2" "Speech Over//Telephone//Channels" ""
;; LABEL "F3" "Speech in the//Presence of//Background Music" ""
;; LABEL "F4" "Speech Under//Degraded//Acoustic Conditions" ""
;; LABEL "F5" "Speech from//Non-Native//Speakers" ""
;; LABEL "FX" "All other speech" ""
;; CATEGORY "2" "Speaker Sex" ""
;; LABEL "female" "Female" ""
;; LABEL "male"   "Male" ""
;; LABEL "unknown"   "Unknown" ""
;; CATEGORY "3" "Topic" ""
;; LABEL "ozet" "Ozet" ""
;; LABEL "spor" "Spor" ""
;; LABEL "hava" "Hava_Durumu" ""
;; LABEL "isitme" "Isitme_Engelliler" ""
;; LABEL "demec" "Demec" ""
;; LABEL "ekonomi" "Ekonomi" ""
;; LABEL "haberler" "Haberler" ""
;; LABEL "unknown"   "Unknown" ""

FM1028_0108_063000 1 excluded_region 0.000 1.800 <o,,unknown>
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 1.800 10.450 <o,f0,female,Haber> geCen
hafta toplanan yeni kongrede CoGunluGu oluSturan demokrat partili Uyeler Iraka daha fazla
asker gOnderilmesine karSI CIkIyor.
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 10.450 18.265 <o,f0,female,Haber>
temsilciler meclisi baSkanI nancy pelosi Iraktaki mevcut askerlere daha fazla Odenek
ayrIlmasInI destekliklerini bildirdi.
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 18.265 23.575 <o,f0,female,Haber> bununla
birlikte pelosi baSkan bushdan ek asker gOnderilmesini OngOren planI
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 23.575 28.619 <o,f0,female,Haber> ve
istediGi tahsisat hakkInda gerekCeler gOstermesi gerektiGini bildirdi.
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 28.619 35.825 <o,f0,female,Haber> pelosi
amerikan halkInIn sonu belli olmayan bir savaSI desteklemeye mecbur edilemeyeceGini de
belirtti.
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 35.825 40.553 <o,f0,female,Haber>
amerikan anayasasI baSkana askeri kararlar alma konusunda yetki veriyor
FM1028_0108_063000 1 FM1028_0108_063000_VOA_Spiker_1 40.553 45.550 <o,f0,female,Haber> ancak
savunma harcamalarInIn arttIrIlmasI kongrenin yetkisine giriyor.
```

c) Audio verilere ilişkin kelime ve sesbirim tabanlı CTM (conversation time mark) dosyaları. Bu dosyalar dosyanın adı, kanal sayısı ve kelime tabanlı ise her bir kelimenin başlangıç süresini, kelimenin başlangıç ve bitiş süreleri arasındaki toplam süre ile sözkonusu kelimenin metin bilgisini, sesbirim tabanlı ise herbir kelimeyi oluşturan sesbirimlerin başlangıç ve toplam süreleri ile sözkonusu sesbirimlerin metin bilgisini içermektedir.

Aşağıda, yukarıda verilen stm uzantılı dosyaya ilişkin olarak hazırlanan ctm uzantılı dosyalardan bir örnek verilmektedir.

```
FM1028_0108_063000 1 1.80 0.36 Z1
FM1028_0108_063000 1 2.16 0.31 geCen
```

```
FM1028_0108_063000 1 2.47 0.30 hafta
FM1028_0108_063000 1 2.77 0.59 toplanan
FM1028_0108_063000 1 3.36 0.26 yeni
FM1028_0108_063000 1 3.62 0.56 kongrede
FM1028_0108_063000 1 4.18 0.55 CoGunluGu
FM1028_0108_063000 1 4.73 0.47 oluSturan
FM1028_0108_063000 1 5.20 0.41 demokrat
FM1028_0108_063000 1 5.61 0.04 Z1
FM1028_0108_063000 1 5.65 0.37 partili
FM1028_0108_063000 1 6.02 0.46 Uyeler
FM1028_0108_063000 1 6.48 0.42 Z1
FM1028_0108_063000 1 6.90 0.55 Iraka
FM1028_0108_063000 1 7.45 0.23 Z1
FM1028_0108_063000 1 7.68 0.28 daha
FM1028_0108_063000 1 7.96 0.38 fazla
FM1028_0108_063000 1 8.34 0.36 asker
FM1028_0108_063000 1 8.70 0.74 gOnderilmesine
FM1028_0108_063000 1 9.44 0.32 karSI
FM1028_0108_063000 1 9.76 0.01 Z1
FM1028_0108_063000 1 9.77 0.43 CIkIyor
FM1028_0108_063000 1 10.20 0.25 Z1
FM1028_0108_063000 1 10.45 0.42 Z1
FM1028_0108_063000 1 10.87 0.63 temsilciler
FM1028_0108_063000 1 11.50 0.37 meclisi
FM1028_0108_063000 1 11.87 0.40 baSkanI
FM1028_0108_063000 1 12.27 0.36 $nancy
FM1028_0108_063000 1 12.63 0.60 pelosi
FM1028_0108_063000 1 13.23 0.39 Z1
FM1028_0108_063000 1 13.62 0.54 Iraktaki
FM1028_0108_063000 1 14.16 0.36 mevcut
FM1028_0108_063000 1 14.52 0.63 askerlere
FM1028_0108_063000 1 15.15 0.23 daha
FM1028_0108_063000 1 15.38 0.29 fazla
FM1028_0108_063000 1 15.67 0.38 Odenek
FM1028_0108_063000 1 16.05 0.69 ayrIlmasInI
FM1028_0108_063000 1 16.74 0.92 destekledIklerini
FM1028_0108_063000 1 17.66 0.51 bildirdi
FM1028_0108_063000 1 18.17 0.10 Z1
FM1028_0108_063000 1 18.27 0.31 Z1
FM1028_0108_063000 1 18.57 0.36 bununla
FM1028_0108_063000 1 18.94 0.01 Z1
FM1028_0108_063000 1 18.95 0.39 birlikte
FM1028_0108_063000 1 19.34 0.59 pelosi
FM1028_0108_063000 1 19.93 0.29 Z1
FM1028_0108_063000 1 20.21 0.46 baSkan
FM1028_0108_063000 1 20.68 0.48 $bushdan
FM1028_0108_063000 1 21.16 0.03 Z1
FM1028_0108_063000 1 21.19 0.24 ek
FM1028_0108_063000 1 21.43 0.39 asker
FM1028_0108_063000 1 21.82 0.79 gOnderilmesini
FM1028_0108_063000 1 22.61 0.45 OngOren
FM1028_0108_063000 1 23.05 0.50 planI
FM1028_0108_063000 1 23.55 0.01 Z1
FM1028_0108_063000 1 23.57 0.31 Z1
FM1028_0108_063000 1 23.88 0.28 ve
FM1028_0108_063000 1 24.16 0.46 istediGi
FM1028_0108_063000 1 24.62 0.64 tahsisat
FM1028_0108_063000 1 25.27 0.49 hakkInda
FM1028_0108_063000 1 25.75 0.23 Z1
FM1028_0108_063000 1 25.98 0.68 gerekCeler
FM1028_0108_063000 1 26.66 0.74 gOstermesi
FM1028_0108_063000 1 27.41 0.56 gerektiGini
FM1028_0108_063000 1 27.96 0.46 bildirdi
FM1028_0108_063000 1 28.42 0.20 Z1
FM1028_0108_063000 1 28.62 0.37 Z1
FM1028_0108_063000 1 28.99 0.53 pelosi
FM1028_0108_063000 1 29.52 0.51 amerikan
FM1028_0108_063000 1 30.03 0.59 halkInIn
FM1028_0108_063000 1 30.62 0.17 Z1
FM1028_0108_063000 1 30.79 0.37 sonu
FM1028_0108_063000 1 31.16 0.26 belli
FM1028_0108_063000 1 31.42 0.51 olmayan
FM1028_0108_063000 1 31.93 0.15 bir
FM1028_0108_063000 1 32.08 0.38 savaSI
FM1028_0108_063000 1 32.46 0.78 desteklemeye
```

```
FM1028_0108_063000 1 33.24 0.24 Z1
FM1028_0108_063000 1 33.48 0.60 mecbur
FM1028_0108_063000 1 34.08 0.90 edilemeyeceGini
FM1028_0108_063000 1 34.98 0.12 de
FM1028_0108_063000 1 35.10 0.52 belirtti
FM1028_0108_063000 1 35.62 0.21 Z1
FM1028_0108_063000 1 35.83 0.41 Z1
FM1028_0108_063000 1 36.23 0.48 amerikan
FM1028_0108_063000 1 36.72 0.68 anayasasI
FM1028_0108_063000 1 37.40 0.59 baSkana
FM1028_0108_063000 1 37.98 0.01 Z1
FM1028_0108_063000 1 38.00 0.48 askeri
FM1028_0108_063000 1 38.48 0.49 kararlar
FM1028_0108_063000 1 38.97 0.26 alma
FM1028_0108_063000 1 39.23 0.47 konusunda
FM1028_0108_063000 1 39.70 0.33 yetki
FM1028_0108_063000 1 40.03 0.45 veriyor
FM1028_0108_063000 1 40.48 0.07 Z1
FM1028_0108_063000 1 40.55 0.28 Z1
FM1028_0108_063000 1 40.83 0.37 ancak
FM1028_0108_063000 1 41.20 0.35 savunma
FM1028_0108_063000 1 41.55 0.76 harcamalarInIn
FM1028_0108_063000 1 42.31 0.82 arttIrIlmasI
FM1028_0108_063000 1 43.13 0.28 Z1
FM1028_0108_063000 1 43.41 0.63 kongrenin
FM1028_0108_063000 1 44.04 0.49 yetkisine
FM1028_0108_063000 1 44.53 0.41 giriyor
FM1028_0108_063000 1 44.94 0.61 Z1
FM1028_0108_063000 1 45.55 0.12 Z1
```

Elde edilen belirli sayıdaki konuşma verisine ilişkin dosyalar üzerinde çalışılarak bunların mümkün olduğunca birbirleri ile eşleştirilmeleri ve bazı hataların giderilmesine çalışılmıştır. Bu sayede birbirleri ile eşleştirilmiş 1) wav, 2) stm, 3) kelime tabanlı ctm ve 4) sesbirim tabanlı ctm formatında hazırlanmış tüm dosyaları mevcut olan toplamda 42 adet farklı kayıta ilişkin dosyalar elde edilmiştir. Bu dosyalardan wav dosyalarının herbiri 30 dakikalık ve tek kanal olarak kayıt edilmiş haber içerikli konuşma kayıtlarını içermekte olup, verilerin toplamı 21 saatlik bir veriye karşı gelmektedir. Bir perl kodu ile özellikle stm dosyalarındaki cümlelere ilişkin kelimeler tek tek elde edilerek ctm dosyalarındaki kelimeler ile hizalanmışlardır. Tespit edilen hatalar Boğaziçi Üniversitesi BUSİM Speech Processing Group da çalışan araştırmacılar ile de paylaşılmış ve bazı dosyaların yeniden elde edilmesi sağlanmıştır. Projeye ilişkin tüm işlemler bu dosyalar üzerinde gerçekleştirilmiştir.

Öncelikle referans dosyaların (stm) kullanılması ile herbir dosyada yeralan herbir kelime için ilgili kelimenin cümle sınırı olup olmadığını belirleyen bir perl kodu yazılmıştır. Bu işlemde, stm dosyalarında yeralan kelimelerin sonunda nokta gibi cümle sonunu belirten noktalama işaretlerinin bulunmaması durumunda, kelime sınırı, cümle sınırı olmadığını gösteren n etiketi ile (n: non-sentence boundary), bulunması durumunda ise cümle sınırı olduğunu gösteren s (s: sentence boundary) etiketi ile etiketlenmişlerdir.

Yapılan bu işlemler sonucunda belirlenen 42 adet dosyada toplam 104458 adet kelime ve 6881 cümle sınırı olduğu tespit edilmiştir. Daha sonra, yapmış olduğumuz sözlüksel,

biçimbilgisel ve bürünsel özellik çıkarımı işlemlerinden sonra, elde edilen dosyaların büyük bir kısmı eğitim seti, bir kısmı geliştirim seti bir kısmı ise test set olarak ayrılmaktadır. Özellikle cümle bölütleme sistemimizin performansı, farklı co-training stratejilerinin kullanıldığı algoritmalar yardımı ile development set üzerinde optimize edilmekte ve test set üzerinde ölçülmektedir.

4. BİÇİMBİLGİSEL, SÖZCÜKSEL VE BÜRÜNSEL ÖZELLİKLERİN ÇIKARILMASI

4.1 Türkçenin Biçimbilgisel Analizi (Morphological Analysis) ve Biçimbilgisel Özelliklerin (Morphological Features) Çıkarılması

Kelimelerin biçimbilgisel analizini yapabilmek için Oflazer (1994) ve Oflazer vd. (1995) tarafından geliştirilen biçimbilgisel analiz yapan araçlar kullanılmıştır. Bu araçlar aynı zamanda biçimbirim/anlambirim ve tonlamaya ilişkin grup sınırlarını da işaretleyebilmektedir. Bunlar özellikle biçimbirim tabanlı özelliklerin çıkarılması için yararlı olabileceklerdir. Otomatik konuşma tanıyıcı çıkışındaki sesbirimlerin (phoneme) başlangıç ve bitiş zamanlarından biçimbirimlerin başlangıç ve bitiş zamanları hesaplanmaktadır. Bazı kelimeler belirsizlikten dolayı birden fazla biçimbilgisel analize sahiptirler. Örneğin “bakan” kelimesi devlet kademesindeki “bakan” anlamı ile yorumlanabileceği gibi bakmak fiilini gerçekleştiren özne (bak+an) olarak da yorumlanabilmektedir. Özelliklerin hesaplanması sırasında belirsizliğin korunmasına devam edilmiş ve tüm çoğul versiyonlar için biçimbilgisel özellikler kestirilmiştir. Bu özelliklere ilişkin veritabanı, daha sonra örneğin diğer yararlı bilgilerin de birleştirilebileceği bürünsel özelliklerin kullanılması ile biçimbilgisel açıklaştırma için kullanılabilecektir.

Türkçe cümle öğe sıralamasının serbest olarak kullanılabildiği bir dildir. Bununla birlikte cümle öğelerinin tipik sırası özellikle haber tarzında konuşma türlerinde özne-nesne-yüklem biçimindedir.

Türkçede “Çocuk yemek yedi” cümlesini ele aldığımızda İngilizcedeki karşılığı “The child ate the meal” olacaktır. Doğru bir biçimde yapılan biçimbilgisel analiz sonucunda aşağıda belirtilen biçimbilgisel yapı elde edilecektir.

çocuk: Noun+A3sg+Pnon+Nom (the child)
yemek: Noun+ A3sg+Pnon+Nom (the meal)
yedi: Verb+Pos(+dH)+Past+A3sg (ate)

Türkçe, çekimsel ve türetsel sondan eklemelerin yapılabildiği bitişken biçimbilgisine sahiptir. Bir kökten çok sayıda yeni kelime formu türetilebilir. Aşağıda verilen örnekte olduğu gibi, herhangi bir yüklem, isim veya sıfatın sonuna sadece 3 adet biçimbirim in sondan eklenmesi ile yeni kelime formları türetilebilmektedir.

Türkçede biçimbilgisel bilgi genel olarak aşağıda verilen biçimde gösterilebilir.

Root + IG_1 + ^DB + IG_2 + ^DB + ... + ^DB + IG_n

Bu gösterimde çekimsel gruplar (IG-Inflectional Groups) türetsel sınırları göstermekte ve türetsel sınırlar (DB-Derivational Boundaries) ^DB ile işaretlenmektedirler. Bir kelimenin kök ve türetsel elemanları farklı IG ler ile temsil edilebilmektedir. Herbir IG_i uygun bir çekimsel özellik dizisini göstermektedir. Bu çekimsel özelliklerin bazıları aşağıdaki gibi listelenebilirler.

+Adj: adjective (sıfat)
+Noun: noun (isim)
+Verb: verb (yüklem)
+A3sg: 3rd person singular agreement (3. tekil şahıs uyumu)
+P1sg: 1st person singular possesive agreement (1.tekil şahıs iyelik (mülkiyet) uyumu)
+Pnon: no possessive agreement (iyelik uyumu bulunmaması)
+Nom: nominative case (yalın durum)
+Past: past tense (geçmiş zaman)
+Fut: future tense (gelecek zaman)
+FutPart: future participle (gelecek zaman ortaçı)

Bir örnek olarak “yapabileceğim” kelimesini analiz edelim. Yapabileceğim kelimesi;
(yap) + (abil) + (ecek) + (im)
biçiminde biçimbirimlerine ayrılabilmektedir.

Bu kelimeye ilişkin potansiyel 3 farklı biçimbilgisel analiz yapılabilir. Bunlar;

1) (yap) yap+Verb+Pos(+yAbil)^DB**+Verb**+Able(+yAcAk)+Fut(+yHm)+A1sg
   [I will be able to do it]
2) (yap) yap+Verb+Pos(+yAbil)^DB+Verb+Able(+yAcAk)^DB**+Adj**+FutPart(+Hm)+P1sg
   [The (thing that) I will be able to do]
3) (yap)
   yap+Verb+Pos(+yAbil)^DB+Verb+Able(+yAcAk)^DB**+Noun**+FutPart+A3sg(+Hm)+P1sg+Nom
   [The one I will be able to do]

Bu örnekte, kök bir yüklem olmakla birlikte herbir analizin en sonunda yeralan çekimsel gruplar (IG) sırası ile yüklem (Verb), sıfat (Adj) ve isimdir (Noun).

Türkçede yeralan kelimeler yapı itibari ile biçimbilgisel olarak ayrılabilecek olası çok sayıda birimlerden (cümlenin öğeleri) (POS-Part of Speech Tags) oluşabildiğinden istatistiksel modeller için oldukça ilginç bir problem ortaya koymaktadır. Bu özellik, Türkçenin oldukça üretken ve türetsel bir biçimbilgisel yapıya sahip olmasından ileri gelmektedir.

Biçimbilgisel özellikler dizayn edilirken, sözkonusu biçimbilgisel analiz yapan araçların herbir kelime için verdiği olası tüm biçimbilgisel olarak ayrılmış kısımlar kullanılmaktadır (Bkz. Yukarıda verilen örnek). Geliştirdiğimiz en önemli biçimbilgisel özelliklerden bir tanesi, modellerimizi basitleştirmek için kullandığımız ve herbir kelimenin en sonunda yeralan çekimsel gruplardır. Bu gruplar cümledeki enson kategoriyi işaret etmektedir. Kelime-ifade belirsizliğini çözümlemeksizin herbir kelimeye ilişkin en son çekimsel grup ile birlikte biçimbilgisel olası cümle öğeleri de elde edilmiştir.

Cümle öğelerinin etiketlerinin özellik değeri olarak, eğer kelime birden fazla ayrıştırılmış kısım (parse) içeriyorsa bilinmeyen olarak işaretlenmektedir. Ayrıca, bir kelimenin herhangi bir biçimbilgisel ayrışımının final kategorisine göre yüklem olup olmadığını denetleyen bir baska iki durumlu biçimbilgisel özellik daha çıkarılmıştır. Böylece bu sayede Türkçenin SOV (subject-object-verb) doğasının avantajının kullanılması öngörülmektedir. Türkçenin bu yapısı cümle bölütlemede büyük bir avantaj yaratmaktadır. Bu özelliklerden başka, herbir kelimenin son üç harfini içeren bir başka biçimbilgisel özellik daha çıkarılmıştır (pseudo morphological features). İngilizcedeki “ed” son ekine benzer biçimde Türkçede de belirli son ekler yüklem kategorisine işaret edebilmektedir. Bu durum, cümle bölütleme işleminde, cümle sınırlarının belirlenmesinde önemli bir ipucu olarak kullanılmaktadır. Özellikle yüklem bilgisi Türkçenin biçimbilgisel analizinin getirdiği dilbilimsel açıdan en önemli özelliktir. Her ne kadar Türkçe, kelime sırasının serbestçe kullanılabildiği bir dil olsa da, en sık kullanılan öğe sıralaması SOV biçiminde olmaktadır. Özellikle radyo haber yayınlarında ve Türkçenin formal kullanımında öğe sıralamasının daima SOV biçiminde olduğu görülmektedir. Bu nedenle, özellikle, biçimbilgisel analizlerden biri yüklem formunda ise bu durum cümle sonunu işaret eden en önemli belirteçlerden biri olmaktadır. Dolayısı ile bu ipucu yada özellik, cümle sınırlarının belirlenmesi, başka bir ifade ile cümle bölütlemesi için oldukça yararlı bir bilgiyi taşımaktadır. Geliştirilen ve deneylerde kullanılan biçimbilgisel özellikler aşağıda listelenmiştir.

Biçimbilgisel Özellikler (morphological features)
lastMarkerA3sg: 1 ,0. lastMarkerNom: 1, 0. lastIGhasVerb: 1, 0. lastPOS: Adj, Adverb, Conj, Det, Dup, Interj, Noun, Num, Postp, Pron, Ques, Verb. PrevLast3: text. CurrentLasr3: text. NextLast3: text. PrevCurrentLast3: text. CurrentNextLast3: text. PrevCurrentNextLast3: text.

4.2 Türkçe Konuşma Verilerinin Sözcüksel Özelliklerinin (Lexical features) Çıkarılması

Word *n*-grams: Verilen bir kelime dizisinden maksimum *n* adet kelimeden oluşan alt diziler yada kelime kombinasyonlarıdır. Örneğin "word 3-grams" word triagrams olarak adlandırılır ve tekli, ikili, üçlü kelime gruplarını içerir. Kullandığımız sözcüksel özellikler herbir kelime sınırı için oluşturulmuş 6 adet kelime *n*-gram özellikleridir. Bunlardan 3 adet unigram, 2 adet bigram ve 1 adet trigram oluşturulmuştur. *n*-gramlar cümle sınırı olup olmadığı ile ilgilenilen kelime, bir sonraki kelime ve bir önceki kelimenin yanyana getirilmiş kombinasyonları ile oluşturulmaktadır (Cuendet vd., 2007). Perl dilinde yazılan bir kod ile Türkçe konuşma verilerine ilişkin tüm kelimelerin sözcüksel özellikleri aşağıda verilen formatta çıkarılmıştır.

6 adet sözcüksel özellik aşağıda verilmektedir.
Unigram lar : {bir önceki (previous)}, {şimdiki (current)}, {bir sonraki (next)}
Bigram lar : {şimdiki, bir sonraki}, {bir önceki, şimdiki}
Trigram lar : {bir önceki, şimdiki, bir sonraki}

Sözcüksel özellikler aşağıda verilen formatta hazırlanmışlardır.
s, *n* : cümle sınırı, cümle sınırı değil (sentence : s, nonsentence boundary : n)
w : güncel yada ilgilenilen kelime (current word : w_i)
wn : bir sonraki kelime (next word : w_{i+1})
wp : bir önceki kelime (previous word : w_{i-1})
wwn : guncel kelime ve bir sonraki kelime (current word and next word : w_i , w_{i+1})
wwp : guncel kelime ve bir önceki kelime (current word and previous word : w_i , w_{i-1})
wpwwn :bir önceki kelime, güncel kelime, birsonraki kelime (previous word, current word, next word: w_{i-1} , w_i , w_{i+1})

Burada w_i ilgilenilen cümle sınırından önceki son kelimeyi işaret eder.

Sözcüksel Özellikler (lexical features) (word *n*-grams)
s,n. w: text. wp: text. wpwwn: text. wn: text. wwn:text. wpw: text. Veya gerçek formatta; w,wp,wpwwn,wn,wwn,wpw, cümle sınırı (sentence boundary) (s veya n).

Türkçe konuşma verileri için oluşturulmuş Örnek bir cümleye ilişkin sözcüksel özelliklerin çıkarılması ve cümle sınırlarının işaretlenmesi aşağıda belirtildiği gibi yapılmaktadır.

Örnek Cümle: “Burası Amerikanın sesi Türkçe yayın bölümü Washington.”

burasI amerikanIn sesi tUrkCe yayIn bOlUmU $washington.	
w,wp,wpwwn,wn,wwn,wpw	cümle sınırı (s/n)
burasI,-,- burasI amerikanIn,amerikanIn,burasI amerikanIn, - burasI,	n
amerikanIn,burasI,burasI amerikanIn sesi,sesi,amerikanIn sesi, burasI amerikanIn,	n
sesi,amerikanIn,amerikanIn sesi tUrkCe,tUrkCe,sesi tUrkCe, amerikanIn sesi,	n
tUrkCe,sesi,sesi tUrkCe yayIn,yayIn,tUrkCe yayIn, sesi tUrkCe,	n
yayIn,tUrkCe,tUrkCe yayIn bOlUmU,bOlUmU,yayIn bOlUmU, tUrkCe yayIn,	n
bOlUmU,yayIn,yayIn bOlUmU $washington,$washington,bOlUmU $washington, yayIn bOlUmU,	n
$washington,bOlUmU,bOlUmU $washington bugUn,bugUn,$washington bugUn, bOlUmU $washington,	s

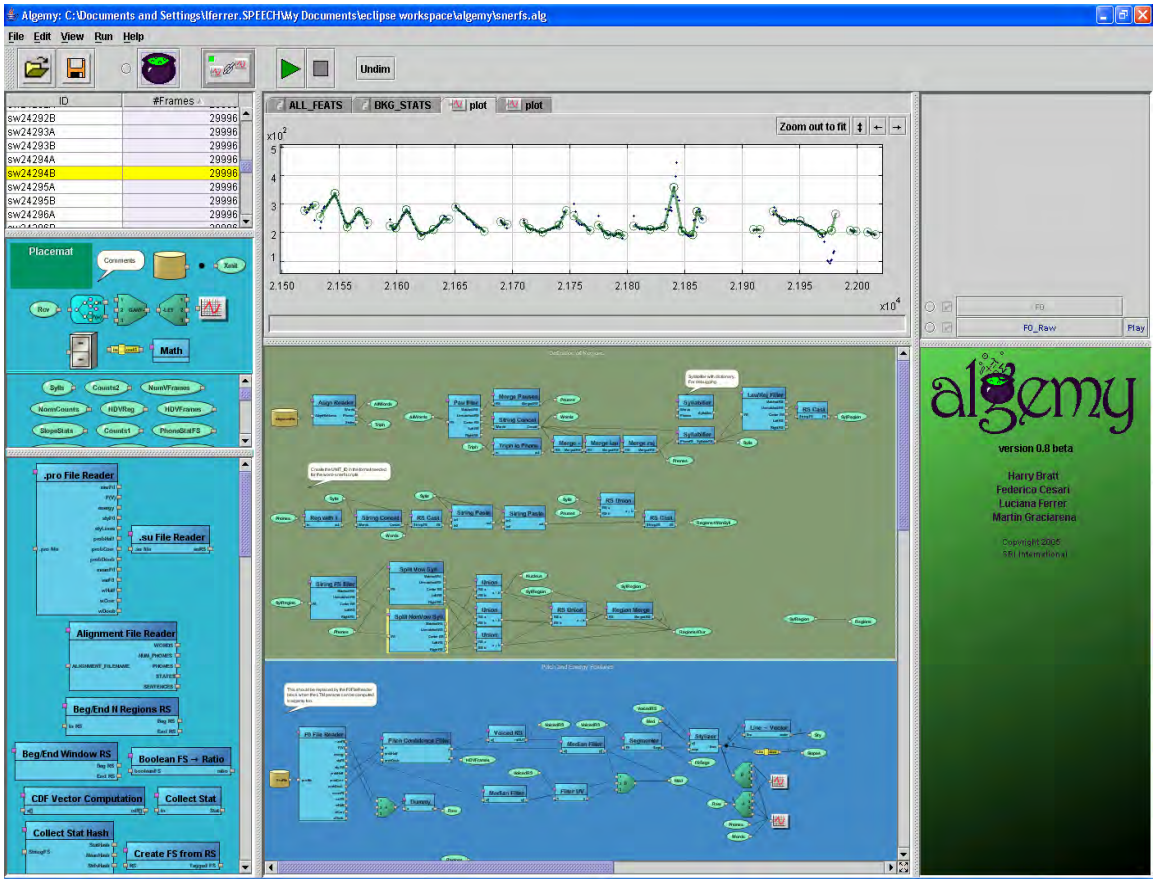
4.3 Türkçe Konuşma Verilerinin Bürünsel Özelliklerin Hesaplanması (Computation of Prosodic Features)

Bürün genel olarak bir dilin vurgu, perde, tonlama, duraksama, vb. özelliklerini içerir. Bu özellikler, anlam farkı yaratmaktan çok, niyete veya işleve işaret eder. Soru sorulurken son hecelerin yüksek bir tonlamayla ifade edilmesi, keskin bir vurgunun kızgınlık, yüksek bir tınının kaygı ifade edebilmesi buna birkaç örnektir. Konuşma işaretlerinin otomatik konuşma sistemleri tarafından basit kelime dizilerine dönüştürülmesi sonucu konuşma ile ilgili zamanlama ve perde yapılarına ilişkin bilgiler veya özellikler kaybolur. Bu yapılar (ve bu yapılara bağlı kelimelerden bağımsız diğer özellikler) bürün olarak adlandırılır. Tüm dillerde bürün, yapısal, anlamsal ve işlevsel bilgiyi taşımaktadır. Bürünsel özellikler veya ipuçları dilin söylem yapısı ile ilişkili olduklarından özellikle çeşitli bilgi çıkarılması işlevlerinde önemli bir rol oynamaktadır. Dilbiliminde, okuma ve anında yapılan monologların analizinde ve buna bağlı işlevlerde cümleler ve paragraflar gibi bilgi birimlerinin bürünsel olarak işaretlenmedikleri gösterilmiştir. İngilizcede ve ona benzer veya bağlı dillerde bürünsel belirteçler olarak duraklar, perde aralığındaki ve genliğindeki değişim, genel perde eğimi, ezgi ve sınır ton dağılımı ve konuşma oranındaki değişim gibi özellikler sayılabilir. Örneğin cümle veya paragraf sınırları ile konu sınırlarının her ikisi de bazı uzun durakların, tonda

meydana gelen bir düşmenin ve perdenin yeniden düzenlenmesinin olduğu durumlarda işaretlenir. Bunlar sınır belirleme veya işaretleme için önemli göstergelerdir. Ayrıca, bürünsel özellikler doğası gereği göreceli bir biçimde kelimenin kimliğinden etkilenmediğinden, başka bir ifade ile kelimelerden bağımsız olduklarından otomatik konuşma sistemi çıkışını baz alan sözlüksel bilgi çıkarım yöntemlerinin gürbüzlüğünü geliştirmektedir. Kullanılan bürünsel özelliklerden bazılarına ilişkin özet bilgiler aşağıda verilmiştir.

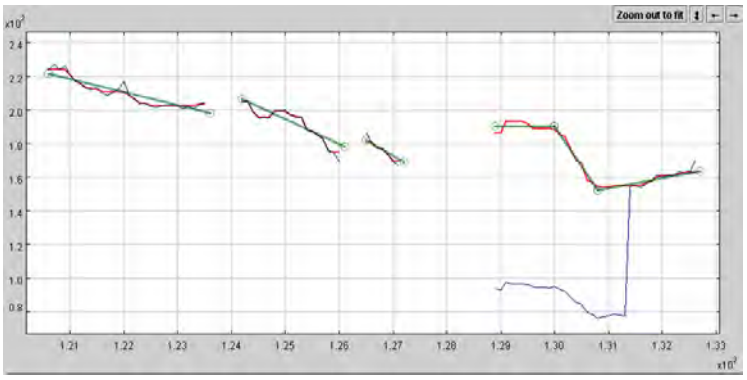
4.3.1 Bürünsel Özelliklerin ALGEMY (SRI-International'ın bürünsel özellikleri hesaplama yazılımı) ile Hesaplanması

Bürünsel özelliklerin tamamı SRI-International tarafından geliştirilen ve Algemy adı verilen grafiksel kullanıcı arayüzü yapısındaki bir yazılım ile gerçekleştirilmiştir. Algemy ile bazı bürünsel özelliklerin çıkarılması Ferrer (2002) ve Fung (2011) tarafından yapılmıştır. Şekil 4.1 de Algemy grafiksel kullanıcı arayüzeyi gösterilmektedir. Ancak bu projenin amaçlarından biri de bürünsel özelliklerin açık kaynak kodlu sistemler ile gerçekleştirilmesidir. Bu amaçla bürünsel özelliklerin tamamı aynı zamanda Huang vd. (2006) tarafından geliştirilen açık kaynak kodlu PRAAT tabanlı Purdue Prosodic Feature Extraction Tool yazılımı ile de elde edilmiştir.

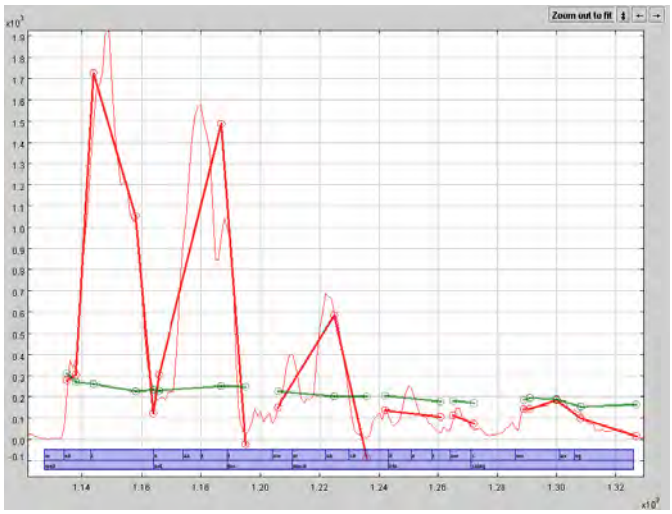


Şekil 4.1. Algemy grafiksel kullanıcı arayüzeyi

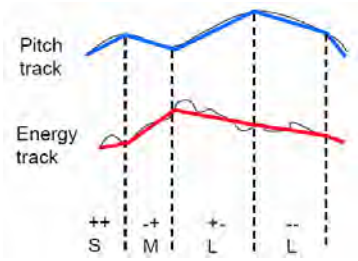
Bürünsel özelliklerin belirlenmesinde en basit yaklaşım perde ve enerji dağılımlarının modellenmesidir. Özellikler herbir çerçeve için, perde nin logaritması log(pitch), enerjinin logaritması log-energy ve sesli çerçeveler için delta özellikleridir. Bu özellikler UBM-GMM (Universal Background Model-Gaussian Mixture Models) ile modellenmektedir. Daha az basit olan bir yaklaşım da, perde enerji ve işaret sürelerinin modellenmesidir. Bu yaklaşımda perde ve sayısal olarak süreleri ile etiketlenmiş perde ve enerji izlerinin eğimlerini açıklayan bir semboller dizisi yaratmaktır. Bu özellikler ayrık saklı markov modelleri ile (DHMMs) ile modellenmektedirler. Şekil 4.2, Şekil 4.3 ve Şekil 4.4 de sırası ile biçimlendirilmiş perde, biçimlendirilmiş enerji ve bunlara ilişkin izler ve eğimleri yeralmaktadır.



Şekil 4.2. Biçimlendirilmiş perde (stylized pitch)



Şekil 4.3. Biçimlendirilmiş enerji (stylized energy)



Şekil 4.4. Perde ve enerji izleri ile eğimleri

Temel bazı bürünsel özellikler (Base prosodic features):

PAUSE_DUR: Kelime sınırındaki durağın süresidir. Eğer bu sınır dalga biçimleri arasındaki bir sınır ise, toplam durma süresi, sınırdan önceki dalga biçiminde yeralan son kelimenin sonundan (aynı konuşmacıya ilişkin konuşmalar bazında) sınırdan sonraki dalga biçiminde yeralan ilk kelimenin başlangıcına kadar olan süredir. Eğer önceki (sonraki) dalga biçimi tümüyle sessiz bir dalga biçimi ise o anki dalga biçimi için ilk (son) durma süresi öncekinin ilk çerçevesinden (bir sonrakinin son çerçevesine kadar) alınır.

PATTERN BOUNDARY: PATTERN_NEXT_WORD ile birleştirilen PATTERN_WORD ün son "f" veya "r" sinin PATTERN_NEXT_WORD ün ilk "f" veya "r" si arasındaki sınır (PATTERN_WORD: Bu özellik; bir alçalan eğim, bir sessiz bölge ve bir sınırdan önceki kelimedeki yükselen eğimi temsil eden "f", "uv" ve "r" dizisi ile tanımlanır. Bu dizi min_frame_length den daha az ise stilize edilmiş f0 dosyasındaki f ler, "r" ler veya "uv" ler atlanır. Daha uzun diziler sadece bir "f", "r" ve "uv" ile temsil edilir. Farklı eğimli "f" veya "r", dizileri "ff" veya "rr" ile temsil ettirilir).

SLOPE_DIFF: Son non_zero (min_frame_length den uzun) kelime eğimi ile bir sonraki kelimenin ilk non_zero (min_frame_length den uzun) eğimi arasındaki farktır. Eğer min_frame_length zamanlarından daha çok oluşan non_zero eğime sahip olmayan kelimeler var ise bu özellik "X" değerine sahip olur.

PAU_DUR_PREV: Bir önceki sınıra ilişkin değerler olup, PAU_DUR açıklamasının aynısı geçerlidir.

rp: gerçek durak.

f0 (fundamental frequency) temelli bazı özellikler (f0 derived features):

Önceki ve sonraki kelimelerin maksimum ve minimum değerlerinin oranının logaritmasına ilişkin özellikler:

F0K_WRD_DIFF_HIHI_N=log(MAX_PWLFIT_F0 / MAX_PWLFIT_F0_NEXT)

F0K_WRD_DIFF_HILO_N=log(MAX_PWLFIT_F0 / MIN_PWLFIT_F0_NEXT)

F0K_WRD_DIFF_LOLO_N=log(MIN_PWLFIT_F0 / MIN_PWLFIT_F0_NEXT)

F0K_WRD_DIFF_LOHI_N=log(MIN_PWLFIT_F0 / MAX_PWLFIT_F0_NEXT)

F0K_WRD_DIFF_MNMN_N=log(MEAN_PWLFIT_F0 / MEAN_PWLFIT_F0_NEXT)
Bu ifadelerdeki MAX_PWLFIT_F0, MAX_PWLFIT_F0_NEXT, MIN_PWLFIT_F0, MIN_PWLFIT_F0_NEXT, MEAN_PWLFIT_F0 ve MEAN_PWLFIT_F0_NEXT değerlerinin tümü F0 (temel frekans) değerleridir.

Uç kelimelerdeki pitch değerlerinin oranının logaritmasına ilişkin bazı özellikler:
F0K_WRD_DIFF_BEGBEG=log(FIRST_PWLFIT_F0 / FIRST_PWLFIT_F0_NEXT)
F0K_WRD_DIFF_ENDBEG=log(LAST_PWLFIT_F0 / FIRST_PWLFIT_F0_NEXT)
F0K_INWRD_DIFF=log(FIRST_PWLFIT_F0 / LAST_PWLFIT_F0)
Bu ifadelerdeki FIRST_PWLFIT_F0, LAST_PWLFIT_F0, FIRST_PWLFIT_F0_NEXT değerleri F0 (temel frekans) değerleridir.

Eğimdeki normalizasyonlara ilişkin özellikler:
SLOPE _DIFF_N: SLOPE_DIFF / SPKR_FEAT_SD_SLOPE
LAST_SLOPE_N: LAST_SLOPE / LAST_PWLFIT_F0
Bu ifadelerdeki SLOPE_DIFF, LAST_SLOPE ve LAST_PWLFIT_F0 f0 özellikleridir. SPKR_FEAT_SD_SLOPE ise konuşmacı temelli özelliklerden biridir.

Bürünsel Özellikler (prosodic features) den bazıları
s,n. PAUSE_DUR: continuous. PATTERN_BOUNDARY: X, f+f, f+r, r+f, r+r. ENERGY_PATTERN_BOUNDARY: X, f+f, f+r, r+f, r+r. SLOPE_DIFF: continuous. ENERGY_SLOPE_DIFF: continuous. SLOPE_LAST: continuous. ENERGY_SLOPE_LAST: continuous. SLOPE_LAST_N: continuous. ENERGY_SLOPE_LAST_N: continuous. WRD_F0K_DIFF_HIHI_N: continuous. WRD_ENERGY_DIFF_HIHI_N: continuous. WRD_F0K_DIFF_HILO_N: continuous. WRD_ENERGY_DIFF_HILO_N: continuous. WRD_F0K_DIFF_LOLO_N: continuous. WRD_ENERGY_DIFF_LOLO_N:continuous. WRD_F0K_DIFF_LOHI_N: continuous. WRD_ENERGY_DIFF_LOHI_N: continuous. F0K_WIN_DIFF_HIHI_N: continuous. ENERGY_WIN_DIFF_HIHI_N: continuous. F0K_WIN_DIFF_HILO_N: continuous. ENERGY_WIN_DIFF_HILO_N: continuous. F0K_WIN_DIFF_LOLO_N: continuous. ENERGY_WIN_DIFF_LOLO_N:continuous. F0K_WIN_DIFF_LOHI_N: continuous. ENERGY_WIN_DIFF_LOHI_N: continuous. WRD_F0K_DIFF_MNMN_N: continuous. WRD_ENERGY_DIFF_MNMN_N: continuous. F0K_WRD_DIFF_BEGBEG: continuous. ENERGY_WRD_DIFF_BEGBEG: continuous. F0K_WRD_DIFF_ENDBEG: continuous.

```
ENERGY_WRD_DIFF_ENDBEG: continuous.
F0K_INWRD_DIFF:continuous.
ENERGY_INWRD_DIFF:  continuous.
PAU_DUR_PREV:continuous.
.........
.........
rp: continuous.
boundary (s veya  n).
```

4.3.2 Bürünsel Özelliklerin Açık Kaynak Kodlu PRAAT Tabanlı Purdue Prosodic Feature Extraction Tool ile Çıkarılması

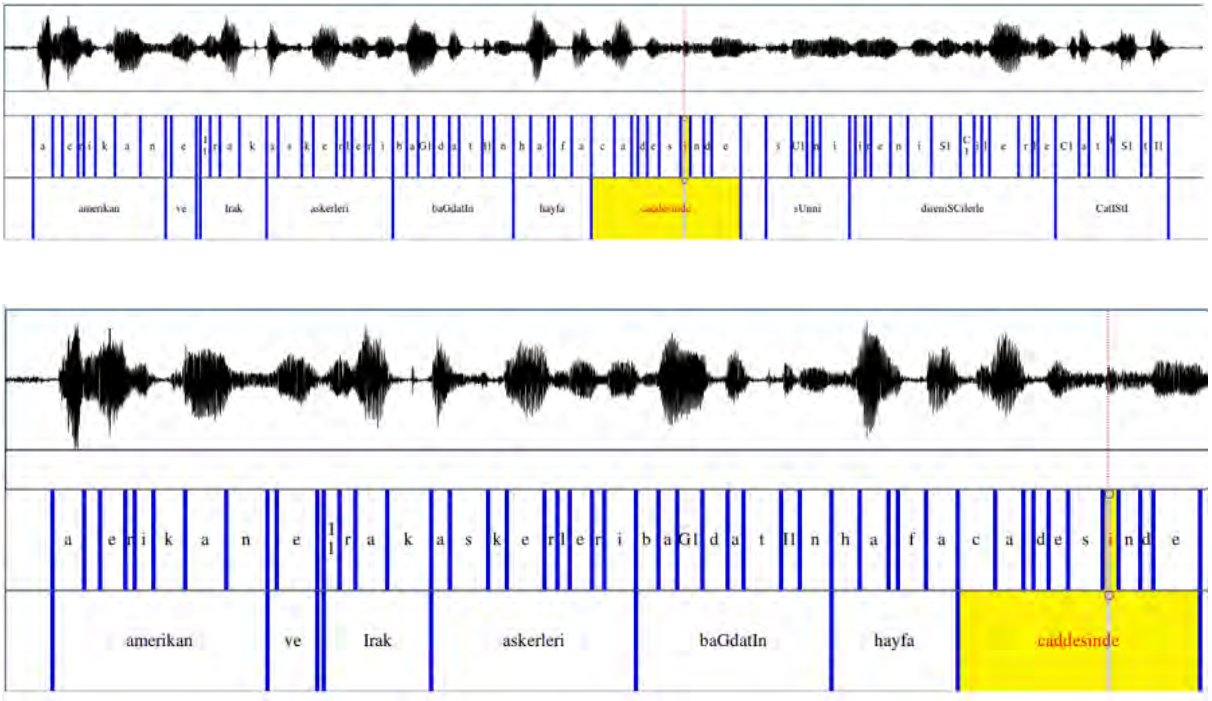
Bürünsel özellikleri çıkarmak amacı ile kullandığımız SRI-International Speech Technology and Research (STAR) Laboratory tarafından geliştirilen Algemy yazılımının yanısıra açık kaynak kodlu ve Praat tabanlı bir yazılım olan ve Purdue Üniversitesi'nde geliştirilen Purdue Prosodic Feature Extraction Tool dan yararlanılmıştır. Bu bağlamda Algemy ile çıkarılan bürünsel özelliklerin büyük bir kısmı Purdue Prosodic Feature Extraction Tool ile de elde edilmiştir. Esasen bu yazılım SRI yazılımından esinlenerek hazırlanmıştır. Praat tabanlı Purdue Prosodic Feature Extraction Tool ile bürünsel özelliklerin çıkarılması için wav formatında hazırlanmış audio dosyalar ile bu dosyalara ilişkin kelime ve sesbirim hizalarına gereksinim vardır. Giriş olarak audio veri ve zamanda hizalanmış kelime ve sesbirimler verildiğinde yazılım öncelikle bazı temel eleman setlerini (örneğin; ham perde, biçimlendirilmiş perde, sesli/sessiz bölütlemesi, duraklar, F0 ve enerji ile ilişkili bileşenler) çıkarmaktadır. Daha sonra süre istatistikleri (örneğin; durak süresinin, sesbirim süresinin ve son uyak süresinin ortalama ve varyans değerleri), F0 a ilişkin istatistikler (örneğin; logaritmik F0 değerlerinin ortalama değeri ve değişinti) ve enerji ile ilişkili istatistikler hesaplanmaktadır. Süre, F0 ve enerji bilgisi ve istatistiklerin de yeralması ile her bir kelime sınırı için bürünsel özellikler çıkarılmaktadır. Aşağıdaki çizelgede elde edilen hangi temel elemanların hangi özelliklerin çıkarılmasında rol oynadıkları görülmektedir. Örneğin; kelime hizalaması, süre, F0 ve Enerji özelliklerinin tümünün hesaplanmasında kullanılırken, sesli sessiz bölütlemesi sadece F0 özelliklerinin hesaplanmasında kullanılmaktadır.

Dalva (2012), Revidi (2014), Dalva vd. (2014a, 2014b) tarafından Praat ile geliştirilen yazılımlar ve Purdue Prosodic Feature Extraction Tool ile, Türkçe diline yapılan uyarlamalar sayesinde bürünsel bilginin Türkçe konuşma verileri için çıkarılması gerçekleştirilmiştir. Geliştirilen sistem ile yukarıda detayları ile verilen tüm bürünsel özellikler kelime bazında çıkarılmıştır. Bürünsel bilgiler Türkçe yayın konuşma verileri içerisinde yeralan herbir konuşmacı için (konuşmacı bazlı) ayrı ayrı çıkarılmaktadır.

Praat tabanlı Purdue Prosodic Feature Extraction Tool ile Bürünsel bilginin çıkarılmasına yönelik adımlar aşağıda belirtilmiştir. Tablo 4.1 de bürünsel özelliklerin hesaplanmasında kullanılan temel elemanlar gösterilmektedir. Şekil 4.5 de Praat ile bürünsel özelliklerin çıkarılmasını gösteren blok şema verilmektedir.

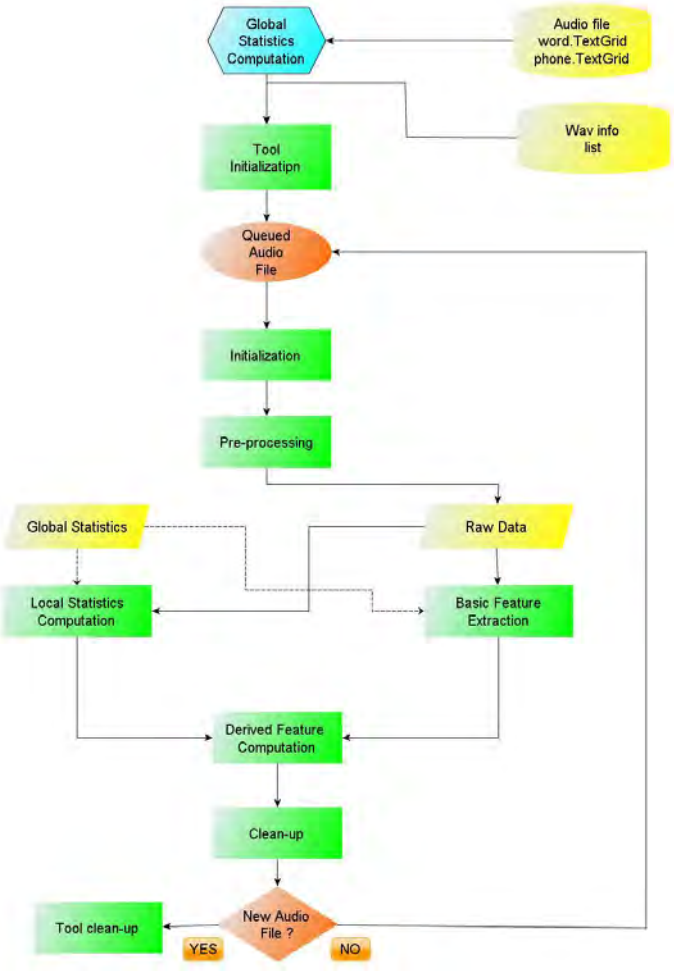
Tablo 4.1. Bürünsel özelliklerin (süre, F0 ve enerji) hesaplanmasında kullanılan temel elemanlar

	Süre Özellikleri	F0 Özellikleri	Enerji Özellikleri
Kelime	√	√	√
Phone	√	×	×
Sesli harf	√	×	×
Uyak	√	×	×
Sesli/Sessiz	×	√	×
Ham Pitch	×	√	×
Biçimlendirilmiş Pitch	×	√	×
Pitch Eğimi	×	√	×
Ham Enerji	×	×	√
Biçimlendirimiş Enerji	×	×	√
Enerji Eğimi	×	×	√



Şekil 4.5. Praat ile bürünsel özelliklerin çıkarılması (zorlanmış hizalamalar ve audio dosyaları ile)

Purdue Prosodic Feature Extraction Tool un bürünsel özelliklerin çıkarılmasında yeralan iki temel adım Şekil 4.6 da verildiği gibidir. Burada ilk adım temel özelliklerin ve istatistik tablolarının hazırlandığı Global istatistiklerin hesaplandığı adımdır. İkinci adım ise, temel özellikler ve istatistik tablolarının hesaplanması ile bürünsel özelliklerin çıkarıldığı adımdır.



Şekil 4.6. Praat ile bürünsel özelliklerin hesaplanması adımları

5. CO-TRAINING İLE CÜMLE BÖLÜTLEME (SENTENCE SEGMENTATION WITH CO-TRAINING)

5.1 Cümle Bölütleme, Boosting ve Boosting Sınıflandırıcılar (AdaBoost)/BoosTexter/ICSIBoost

Tur (2000), Cuendet vd. (2006) ve Tur vd. (2006) tarafından yapılan çalışmalarda gösterildiği üzere, cümle bölütlemede, otomatik konuşma tanıma sisteminden gelen kelimeler dizisi ve komşu kelimeler arasındaki duraksama süreleri önemli ipuçları vermektedir. Cümle bölütleme problemi, verilen bir özellik seti üzerinden sınıfın sonsal olasılık değerinin kestirimine göre herbir kelime sınırının bir sınıf etiketi (s: cümle sınırı olan veya n: cümle sınırı olmayan) ile ilişkilendirildiği bir ikili sınıflandırma problemi olarak ele alınabilir. Verilen bir kelime akışı veya dizisi ($\{w_1, \ ..., \ w_N\}$) için amaç, sınırlar ($\{e_1, \ ..., \ e_N\}, e \in \{s,n\}$) için sınıfların kestirimidir. Burada ($s_i, i = 1,..., N$), w_i ve w_{i+1} arasındaki sınırdır. Genellikle bu, sonsal olasılığın $P(s_i = k \mid o_i), \ k \in \{s,n\}$ kestirimi için sınıflandırıcının eğitilmesi ile yapılır. Burada o_i, ler s_i kelime sınırı için özellik gözlemleridir. İdeal durumda, sınıflandırıcının kararı, en yüksek olasılık değerine $P(s_i = k \mid o_i)$ sahip sınıftır. Bununla birlikte, cümle bölütleme işlevinde cümle sınırı için olasılık $P(s_i = s \mid o_i)$, bir eşik değeri ile karşılaştırılır. Bu eşik değerinin üzerindeki olasılık değerlerinde karar cümle sınırı olduğu yönünde, aksi durumda ise cümle sınırı olmadığı yönünde verilmektedir.

AdaBoost algoritması (Şekil 5.1) Freund ve Schapire (1990), Schapire (1990) ve Schapire (2001) tarafından önerilen ve sınıflandırma ve özellik seçimini aynı anda yapabilen bir algoritmadır. Algoritmanın amacı, eğitim örnekleri üzerinden hesaplanan bir D dağılımına bağlı olarak düşük hatalı bir hipotez oluşturmaktır. AdaBoost algoritması, her bir özellikten zayıf bir sınıflandırıcı oluşturulmasını ve bu zayıf sınıflandırıcılardan kurulu bir komite oluşturulmasını temel alır. Zayıf sınıflandırıcıların karar sınırları, her bir özellik için pozitif ve negatif örneklerin ağırlıklı ortalaması alınarak hesaplanır. Daha sonra hata oranı en düşük olan zayıf sınıflandırıcılar kullanılarak güçlü bir sınıflayıcı oluşturulur. Güçlü sınıflandırıcı içerisinde yer almayan zayıf sınıflandırıcılara ilişkin özellikler elenmiş olur.

Başlangıç Durumunu Hazırlama:
1. Örnek uzayından seçilmiş eğitim seti
$S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, $x_i \in X$ ve $y_i \in Y = -1, +1$(olmak üzere).
2. Başlangıç dağılımı $D_1(i) = \frac{1}{m}$.
Algoritma:
→ $t = 1, \ldots, T$ için aşağıdakileri yinele
D_t dağılımını kullanarak bir zayıf sınıflandırıcı eğit $h_t : X \to \mathbb{R}$
h_t nin ağırlıklarını (α_t) belirle
eğitim seti üzerinden dağılımı yenile:
$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$, ($Z_t$, D_{t+1} bir dağılım olacak biçimde seçilmiş bir normalizasyon faktörüdür)
■ yinelemenin sonu.
En son değer:
$f(x) = \sum_{t=0}^{T} \alpha_t h_t(x)$ ve $H(x) = sign(f(x))$

Şekil 5.1. İkili sınıflandırma (binary classification) işlevi için AdaBoost algoritması

5.2 Yarı Öğreticili Öğrenme, Yarı öğreticili Algoritmalardan Co-taining ve Co-training Stratejilerinin Geliştirilmesi

Yarı öğreticili öğrenmenin amacı istatistiksel modelleri eğitmek için gerekli olan veri miktarını azaltmaktır. Gerçek yaşam uygulamalarının çoğunda verilerin toplanması etiketlenmesinden daha kolaydır. Örneğin, yayın haberlerinden audio işaretlerin kaydedilmesi, yazılması ve bazı etiketler ile etiketlenmesinden daha kolaydır. Yarı öğreticili öğrenme yöntemleri, az miktarda etiketlenmiş veri ve oransal olarak daha yüksek miktarda etiketlenmemiş verinin bulunduğunu varsayar. Bu durumda amaç etiketlenmemiş veriyi kullanarak sistemin başarımını geliştirmektir. Bütün yarı öğreticili öğrenme yöntemlerinde yapılan öncelikle varolan etiketlenmiş veri kullanılarak bir başlangıç sınıflandırıcının eğitilmesidir. Daha sonra temel düşünce bu sınıflandırıcının etiketlenmemiş veriyi kullanarak etiketleme yapması ve etiketlenmiş bu yeni örneklerin sınıflandırıcı başarımını arttırmasıdır. Bu işlemin yinelemeli ve her yinelemede az miktarda etiketlenen örneklerin katılması ile yapılması etiketlenmemiş verinin artımsal bir biçimde kullanılmasını sağlamaktadır. Bu durum, etiketlenmemiş verinin tümünün kullanılması işleminden çok farklıdır ve modellerin davranışını belirgin bir biçimde değiştirmektedir.

Cümle bölütleme işlevi, otomatik konuşma sisteminin çıkışından elde edilen işlenmemiş kelime dizisi biçimindeki kelimelerin cümle sınırı olup olmamasına göre sınıflandırıldığı (etiketlendiği) bir işlevdir. Cümle bölütleme için genellikle istatistiksel yöntemler kullanılmaktadır. Bununla birlikte, bu tür yaklaşımlar hazırlanması pahalı, zaman ve işgücü gerektiren, oldukça belirgin miktarda etiketlenmiş veriye gereksinim duyar. Bu dezavantaj az miktarda etiketlenmiş veri kullanılarak yarı öğreticili öğrenme algoritmaları ile daha yüksek miktardaki veriyi otomatik olarak etiketlemek için yeni metodlar bulunması gerekliliğini ortaya

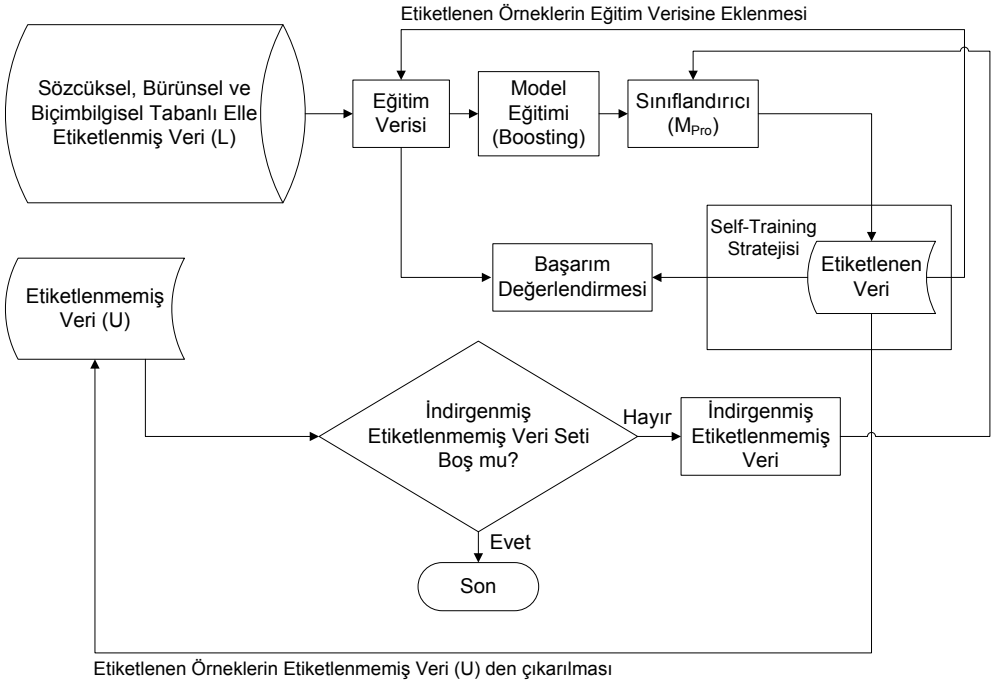
çıkarmıştır. Bu kapsamda Guz vd. (2009) ve Guz vd. (2010) tarafından yapılan çalışmalardaki amaç, biçimbilgisel, sözlüksel ve bürünsel bilginin yarı öğreticili öğrenme algoritmaları ile cümle sınırlarının belirlenmesinde kullanılmasının sağlanmasıdır. Buradaki çalışmanın en önemli amacı, sadece belirli bir miktarda cümle sınırlarının etiketlenmiş olduğu verinin mevcut olduğu durumlarda, verinin geri kalan etiketlenmemiş kısmını etiketleyebilecek etkili ve yeterli bir yarı öğreticili makine öğrenme algoritmasının bulunmasıdır.

5.2.1 Kendi Kendine Eğitme (Self-training) Algoritması

Kendi kendine eğitme en popüler yarı öğreticili öğrenme metodlarından biridir. Kendi kendine eğitmede verilen model verinin etiketlenmemiş kısmının sınıflarını kestirmektedir. Daha sonra otomatik olarak sınıflandırılan örnekler eğitme setine dahil edilmekte ve model yeniden eğitilmektedir. Bu işlem yinelemeli bir biçimde sürdürülmektedir. Yanlış bir biçimde sınıflandırılmış örnekler tarafından getirilen gürültünün ortadan kaldırılması için sadece yüksek güvenilirlik ile sınıflandırılmış örnekler kullanılır. Gerçekleştirdiğimiz ilk yarı öğreticili yaklaşım yinelemeli kendi kendine eğitme yaklaşımıdır. Bunu gerçekleştirmemizin bir başka nedeni ise bu yaklaşımın gerçekleştirdiğimiz diğer yarı öğreticili öğrenme yöntemleri ile yapacağımız performans karşılaştırmalarında ayrı bir baseline olarak kullanılmasıdır. Kendi kendine eğitmede özellik seti tek bir özellikten (bakış) oluşmakta ve sadece tek bir model (bürünsel, sözcüksel veya biçimbilgisel) otomatik olarak kendi kendine örnekleri etiketlemektedir. Kullandığımız kendi kendine eğitme algoritması kod taslağı ve akış şeması sırası ile Şekil 5.2 ve Şekil 5.3 de gösterilmektedir. Burada sınıflandırma gürültüsünü indirgemek için her bir yinelemede en güvenilir örnekler seçilirken bir eşik değeri θ tanımlanmıştır. Deneylerimizde bu eşik değeri eğitme parametrelerin optimize edildiği veri seti (geliştirim seti) kullanılarak optimize edilecektir. Co-training yönteminin performansını değerlendirmek/karşılaştırmak için kendi kendine eğitim yarı öğreticili eğitim algoritması kullanılmıştır. Kendi kendine eğitim için, verilen model verinin etiketlenmemiş kısmı için cümle sınırlarını kestirir. Daha sonra yüksek güvenilirlik skorları ile sınıflandırılmış örnekler eğitim setine ilave edilirler ve model yeniden eğitilir ve tüm işlemler yinelenir. Co-training deneyleri ile uyumlu olunabilmesi için kendi kendine eğitim tüm özelliklerin kullanılması ile değil, bürünsel, sözcüksel ve biçimbilgisel modellerin ayrı ayrı kullanması ile gerçekleştirilmektedir.

Başlangıç Durumunu Hazırlama:
1. Elle etiketlenmiş çok az miktardaki örnek seti (L),
$L = \{(x_1, y_1), \ldots, \ (x_{|L|}, y_{|L|})\}, \quad x_i = (x_{i,Pro}, x_{i,Lex})$.
2. Etiketlenmemiş büyük bir örnek seti (U),
$U = \{(x_1), \ldots, \ (x_{|U|})\}$.
Algoritma:
→ $U \neq \emptyset$ ve geliştirim seti üzerinden hesaplanan hata yakınsamadığı/artmadığı sürece aşağıdakileri yinele (1)
Modelin (M), tüm özellikler ile (L) seti üzerinden eğitilmesi
→ Her bir $x_i \in U$ için aşağıdakileri yinele (2)
Eğer $|f_M(x_i)| > \theta$ ise $U = U - \{x_i\}, \ L = L \cup \{x_i, H_M(x_i)\}$.
■ yinelemenin sonu (2)
■ Yinelemenin sonu (1)

Şekil 5.2. Kendi kendine eğitme (self-training) algoritması

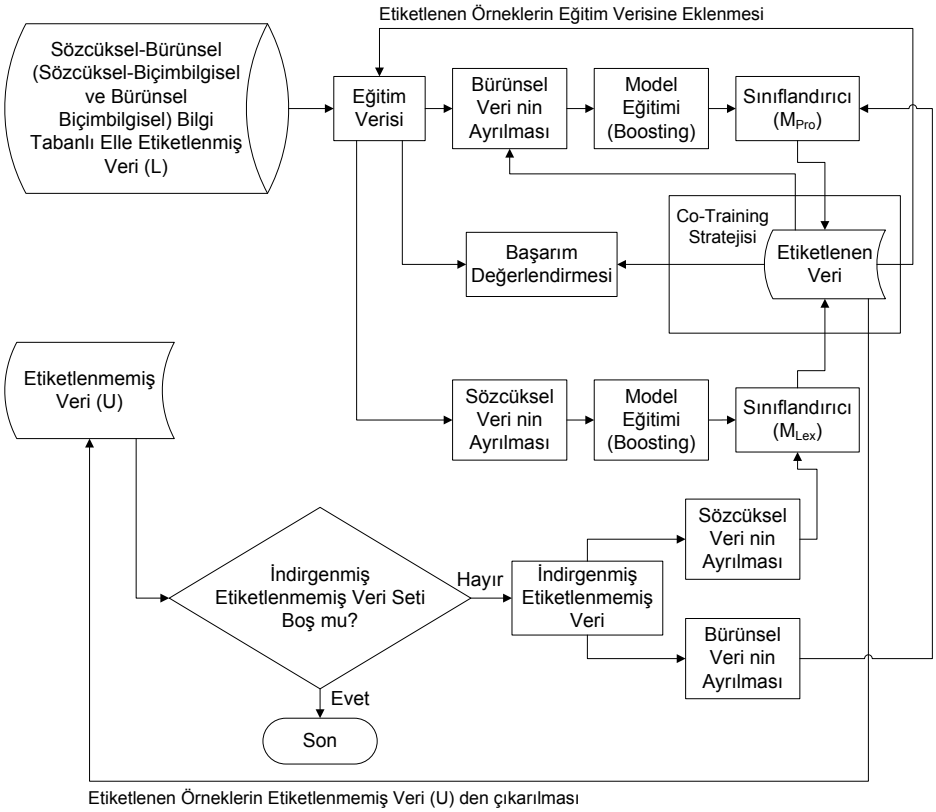


Şekil 5.3. Kendi kendine eğitme (self-training) akış şeması (flow chart)

5.2.2 Co-training Algoritması ve Co-training Stratejileri

Geleneksel tek bakış veya özellik kullanan makine öğrenme konseptinin tersine, çok view lı yaklaşımın özellik seti her biri ayırdedilebilir ve herbiri kendi kendine öğrenebilir yeterlikte iki veya daha fazla farklı bakıştan oluşmaktadır (Hirschberg ve Nakatani,1996; Shriberg, 2000). Bu iki yaklaşımın temel farkı, çok bakış kullanan yöntemde bunlar karşılıklı birbirini eğitirken tek bakışlı yöntemde algoritma kendi kendini eğitir. Co-training oldukça etkili bir yarı öğreticili makine öğrenme algoritmasıdır. Bu algoritma göreceli olarak az miktarda olan etiketlenmiş bir veriden yararlanarak etiketlenmemiş çok miktardaki veriyi azar azar etiketleyen ve bunu

yaparken de çoklu zayıf sınıflandırıcıları kullanan bir algoritmadır. Co-training yaklaşımındaki temel varsayım, sözgelimi probleme özgü olarak kullanılan iki adet zayıf sınıflandırıcıdan birinin, diğer sınıflandırıcının etiketlemeyi başaramadığı örnekleri etiketlemesidir. Böylelikle sınıflandırıcılar karşılıklı olarak birbirlerini eğitmektedirler. Cümle bölütleme probleminin, Co-training yöntemine ilişkin iki temel gerekliliği karşıladığı gerçeği değerlendirildiğinde, bu yöntem için oldukça uygun bir problem olduğu görülecektir. Co-training metodunun uygulanması için sağlanması gereken iki gereklilik veya koşul, veri setinin iki ayrı ve doğal bakış veya özellik setinden oluşmasıdır. Bu projede cümle bölütlemesi için kullanılan olan üç ayrı doğal özellik seti bürünsel, sözcüksel ve biçimbilgisel bilgiyi taşıyan veri setleridir. Co-training her bir bakışta bir hipotez öğrenmeyi baz alan ve etiketlenmemiş örnekler üzerindeki yapılan en güvenilir kestirimlerin eğitme setine yinelemeli olarak eklendiği en etkin çoklu bakış yarı öğreticili makine öğrenmesi yaklaşımlarından biridir. Co-training yaklaşımımız bir çok adımdan oluşmaktadır. Örnek olarak sadece bürünsel ve sözcüksel veriler olması durumunu ele alalım. İlk adımda sadece elle etiketlenmiş az miktardaki veriyi (L) baz alarak sadece bürünsel ve sözcüksel bilgiyi kullanan iki ayrı model (Bürünsel model: M_{Pro} ve Sözcüksel model: M_{Lex}) eğitilmektedir. Daha sonra bu modeller kullanılarak U verisinin etiketlenmemiş kısmına ilişkin cümle sınırlarının kestirilmesi gerçekleştirilmektedir. x_i örnekleri her iki durumda da güvenilirlik değerlerine göre sıraya dizilmektedirler. Bu noktada, her iki taraftan örnek setlerini bulabilmek için farklı örnek seçim metodları uygulanmıştır. Burada sadece bürünsel ve sözcüksel bilginin birlikte kullanıldığı Co-training algoritması verilmiştir. Bunların dışında biçimbilgisel özelliklerin de Co-training de kullanılması ile farklı kombinasyonlarda özellik setleri oluşturularak deneyler tekrarlanmaktadır. Co-training genel akış şeması Şekil 5.4 de, co-training stratejilerine (uzlaşma, uzlaşmama ve self-combined) ilişkin algoritmaların kod taslakları ise sırası ile Şekil (5.5, 5.6, 5.7) de gösterilmektedir.



Şekil 5.4. Co-training (eş öğrenme) akış şeması (flow chart)

5.2.2.1 Uzlaşma/Uyuşma (Agreement) Stratejisi

Bu stratejide, hem bürünsel model hem de sözcüksel modelin her ikisi için de sadece en yüksek güvenilirlik sayılarına sahip örnekler dikkate alınmaktadır. Bu örnekler bu iki ayrı bürünsel ve sözcüksel modellerin eğitim setlerine ilave edilerek bu işlem tekrarlanmaktadır.

Başlangıç Durumunu Hazırlama:
1. Elle etiketlenmiş çok az miktardaki örnek seti (L),
$L = \{(x_1, y_1), \ldots, (x_{|L|}, y_{|L|})\}$, $x_i = (x_{i,Pro}, x_{i,Lex})$.
2. Etiketlenmemiş büyük bir örnek seti (U), $U = \{(x_1), \ldots, (x_{|U|})\}$.
Algoritma:
→ $U \neq \emptyset$ ve geliştirim seti üzerinden hesaplanan hata yakınsamadığı/artmadığı sürece aşağıdakileri yinele (1)
L den iki setin
$L_{Pro} = \{(x_{1,Pro}, y_1), \ldots, (x_{|L|,Pro}, y_{|L|})\}$
$L_{Lex} = \{(x_{1,Lex}, y_1), \ldots, (x_{|L|,Lex}, y_{|L|})\}$
elde edilmesi
L_{Pro} kullanılarak M_{Pro} sınıflandırıcısının (modelinin) eğitilmesi
L_{Lex} kullanılarak M_{Lex} sınıflandırıcısının (modelinin) eğitilmesi
→ Her bir $x_i \in U$ için aşağıdakileri yinele (2)
→ Eğer (3) her iki modelin kestirdiği sınıflar birbirine eşit; $H_{M_{Pro}}(x_{i,Pro}) = H_{M_{Lex}}(x_{i,Lex})$ ve güvenilirlik skorlarının toplamı $|f_{M_{Pro}}(x_{i,Pro})| + |H_{M_{Lex}}(x_{i,Lex})| > \theta$ belirli bir eşik değerinden θ büyük ise
Etiketlenmemiş veri setinden ilgili örnekleri çıkar; $U = U - \{x_i\}$
Etiketlenmiş veri setine ilgili örnekleri kestirilen etiketleri ile birlikte ekle;
$L = L \cup \{(x_i, [H_{M_{Pro}}(x_{i,Pro}) = H_{M_{Lex}}(x_{i,Lex})])\}$
■ Eğer in sonu (3)
■ Yinelemenin sonu (2)
Modelin (M), tüm özellikleri ile (L) seti üzerinden eğitilmesi
■ Yinelemenin sonu (1)

Şekil 5.5. Co-training uzlaşma (agreement) algoritması

5.2.2.2 Uzlaşmama/Uyuşmama (Disagreement) Stratejisi

Bu stratejide, bir modelle en yüksek güvenilirlik skorları ile etiketlenmiş ve fakat diğer model ile en düşük güvenilirlik skorları ile etiketlenmiş örnekler dikkate alınmaktadır. Bu örnekler diğer modelin eğitim setine eklenmektedir. Buradaki amaç, sınıflandırılması güç yeni örnekleri diğer modele katmaktır. Bu işlem, modellerin geliştirim setinde artık bir iyileştirme yapmamasına kadar yinelenebilir. Daha sonra, modellerden biri, modelleri birleştirmek için otomatik veya elle etiketlenmiş örneklerin hem bürünsel hem de sözcüksel özelliklerini kullanarak yeni bir tek model eğitebilir.

Başlangıç Durumunu Hazırlama:
1. Elle etiketlenmiş çok az miktardaki örnek seti (L),
$L = \{(x_1, y_1), \ldots, \ (x_{|L|}, y_{|L|})\}$, $x_i = (x_{i,Pro}, x_{i,Lex})$.
2. Etiketlenmemiş büyük bir örnek seti (U), $U = \{(x_1), \ldots, \ (x_{|U|})\}$.
Algoritma:
L den iki setin
$L_{Pro} = \{(x_{1,Pro}, y_1), \ldots, \ (x_{|L|,Pro}, y_{|L|})\}$
$L_{Lex} = \{(x_{1,Lex}, y_1), \ldots, \ (x_{|L|,Lex}, y_{|L|})\}$
elde edilmesi
Diğer iki setin $U_{Pro} = U_{Lex} = U$ elde edilmesi
→ $U_{Pro} \neq \emptyset$ ve $U_{Lex} \neq \emptyset$ ve geliştirim seti üzerinden hesaplanan hata yakınsamadığı/artmadığı sürece aşağıdakileri yinele (1)
L_{Pro} kullanılarak M_{Pro} sınıflandırıcısının (modelinin) eğitilmesi
L_{Lex} kullanılarak M_{Lex} sınıflandırıcısının (modelinin) eğitilmesi
→ Her bir $x_i \in U_{Lex}$ için aşağıdakileri yinele (2)
→ Eğer (3) güvenilirlik skorlarının farkı $|f_{M_{Pro}}(x_{i,Pro})| - |f_{M_{Lex}}(x_{i,Lex})| > \theta_1$ belirli bir eşik değerinden θ_1 büyük ise
$U_{Lex} = U_{Lex} - \{x_i\}$
$L_{Lex} = L_{Lex} \cup \{(x_{i,Lex}, H_{M_{Pro}}(x_{i,Pro}))\}$
$L = L \cup \{(x_i, H_{M_{Pro}}(x_{i,Pro}))\}$
■ Eğer in sonu (3)
■ Yinelemenin sonu (2)
→ Her bir $x_i \in U_{Pro}$ için aşağıdakileri yinele (4)
→ Eğer (5) güvenilirlik skorlarının farkı $|f_{M_{Lex}}(x_{i,Lex})| - |f_{M_{Pro}}(x_{i,Pro})| > \theta_2$ belirli bir eşik değerinden θ_2 büyük ise
$U_{Pro} = U_{Pro} - \{x_i\}$
$L_{Pro} = L_{Pro} \cup \{(x_{i,Pro}, H_{M_{Lex}}(x_{i,Lex}))\}$
$L = L \cup \{(x_i, H_{M_{Lex}}(x_{i,Lex}))\}$
■ Eğer in sonu (5)
■ Yinelemenin sonu (4)
Modelin (M), tüm özellikleri ile (L) seti üzerinden eğitilmesi
■ Yinelemenin sonu (1)

Şekil 5.6. Co-training uzlaşmama (disagreement) algoritması

5.2.2.3 Self-Combined Stratejisi

Bu yaklaşımda, kendi kendine eğitim stratejisi kullanılarak bürünsel ve sözcüksel modellerin kendilerinin herbir iterasyonda aldığı güvenilirlik skorlarının en yükseğine sahip olan örnekler dikkate alınmaktadır.

Başlangıç Durumunu Hazırlama:
1. Elle etiketlenmiş çok az miktardaki örnek seti (L),
$L = \{(x_1, y_1), \ldots, \ (x_{|L|}, y_{|L|})\}$, $x_i = (x_{i,Pro}, x_{i,Lex})$.
2. Etiketlenmemiş büyük bir örnek seti (U), $U = \{(x_1), \ldots, \ (x_{|U|})\}$.
Algoritma:
İki setin $U_{Pro} = U_{Lex} = U$ elde edilmesi
→ $U_{Pro} \neq \emptyset$ ve $U_{Lex} \neq \emptyset$ ve geliştirim seti üzerinden hesaplanan hata yakınsamadığı/artmadığı sürece aşağıdakileri yinele (1)
L den iki setin
$L_{Pro} = \{(x_{1,Pro}, y_1), \ldots, \ (x_{|L|,Pro}, y_{|L|})\}$
$L_{Lex} = \{(x_{1,Lex}, y_1), \ldots, \ (x_{|L|,Lex}, y_{|L|})\}$
elde edilmesi
L_{Pro} kullanılarak M_{Pro} sınıflandırıcısının (modelinin) eğitilmesi
L_{Lex} kullanılarak M_{Lex} sınıflandırıcısının (modelinin) eğitilmesi
→ Her bir $x_i \in U_{Pro}$ için aşağıdakileri yinele (2)
→ Eğer (3) $|f_{M_{Pro}}(x_{i,Pro})| > \theta_1$ ve $H_{M_{Pro}}(x_{i,Pro}) = H_{M_{Lex}}(x_{i,Lex})$ ise
$U_{Pro} = U_{Pro} - \{x_i\}$
$U_{Lex} = U_{Lex} - \{x_i\}$
$L_{Pro} = L_{Pro} \cup \{(x_{i,Pro}, H_{M_{Pro}}(x_{i,Pro}))\}$
$L = L \cup \{(x_i, H_{M_{Pro}}(x_{i,Pro}))\}$
■ Eğer in sonu (3)
■ Yinelemenin sonu (2)
→ Her bir $x_i \in U_{Lex}$ için aşağıdakileri yinele (4)
→ Eğer (5) $|f_{M_{Lex}}(x_{i,Lex})| > \theta_2$ ve $H_{M_{Lex}}(x_{i,Lex}) = H_{M_{Pro}}(x_{i,Pro})$ ise
$U_{Lex} = U_{Lex} - \{x_i\}$
$U_{Pro} = U_{Pro} - \{x_i\}$
$L_{Lex} = L_{Lex} \cup \{(x_{i,Lex}, H_{M_{Lex}}(x_{i,Lex}))\}$
$L = L \cup \{(x_i, H_{M_{Lex}}(x_{i,Lex}))\}$
■ Eğer in sonu (5)
■ Yinelemenin sonu (4)
Modelin (M), tüm özellikleri ile (L) seti üzerinden eğitilmesi
■ Yinelemenin sonu (1)

Şekil 5.7. Co-training self-combined algoritması

6. DENEYLER VE SONUÇLAR

6.1 Farklı Bürünsel Özellik Setlerinin Çıkarılması ve Cümle Bölütleme Performanslarının Karşılaştırılması:

Tablo 6.1 de verilen herbir konuşmacı için tek, ikili ve üçlü olmak üzere 9 farklı bürünsel özellik seti oluşturulmuştur. Tablo 6.1 ve Tablo 6.2 de sırası ile bu konuşmacılara ilişkin veri büyüklükleri ve kullanılan bürünsel özellik setleri yeralmaktadır. Burada M1; süre, perde ve enerji özelliklerinin etkin olanlarından oluşturulmuş özel bir bürünsel özellik setidir.

Tablo 6.1. Konuşmacı bazlı konuşma veri setleri

	Veri Büyüklüğü (W: Kelimeler, S: Cümleler, K=10^3)					
Konuşmacı ID	Eğitim		Geliştirim		Test	
	(W)	(S)	(W)	(S)	(W)	(S)
Konuşmacı 1	10K	660	4K	282	6K	419
Konuşmacı 2	10K	720	4K	283	6K	445

Konuşmacılara ilişkin 0.5K, 1K, 2K, 4K, 6K, 8K ve 10K büyüklüklerinde eğitim setleri oluşturulmuştur. Herbir konuşmacı için 10-kez çapraz doğrulama metodu kullanılmıştır. Herbir bürünsel set için eğitim modelleri oluşturulmuş ve geliştirim seti üzerinde parametreler optimize edildikten sonra test set üzerinde F-measure ve NIST değerleri ölçülmüştür.

Tablo 6.2. Bürünsel özelliklerin içerik ve büyüklükleri

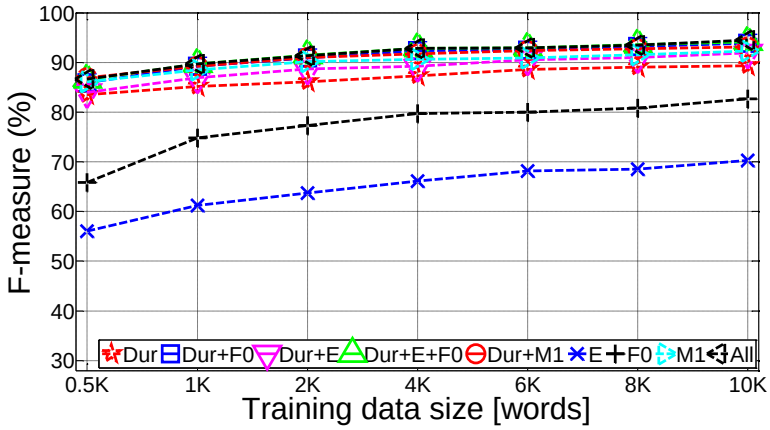
Özellik Kodu	İçerik	Büyüklük (özellik sayısı)
Dur+E+F_0 (Süre+En+ F_0)	Süre, Durak, Enerji ve Perde özellikleri	222
All (Tüm)	Tüm bürünsel özellikler	222
Dur+M1 (Süre+M1)	Süre, Durak ve M1 özellikleri	94
Dur+F_0 (Süre+ F_0)	Süre, Durak ve Perde özellikleri	142
M1 (M1)	M1 özellikleri	33
Dur+E (Süre+En)	Süre, Durak ve Enerji özellikleri	146
Dur (Süre)	Süre ve Durak özellikleri	66
F_0 (F_0)	Perde özellikleri	76
E (En)	Enerji özellikleri	80

Konuşmacılara ilişkin sonuçlar Tablo 6.3 ve Şekil 6.1, Şekil 6.2, Şekil 6.3 ve Şekil 6.4 de verilmektedir. Konuşmacı bazlı deneylerde süre, perde ve enerji tabanlı bürünsel özellik setlerinin cümle bölütleme başarımları karşılaştırıldığında, başarım sıralamasının büyükten küçüğe doğru; süre tabanlı, perde ve enerji tabanlı özellik setleri olarak sıralandığı görülmektedir. Bu sonuç özellikle süre tabanlı ve duraklama özelliklerinin cümle sınırlarının kestiriminde önemli ipuçları olduklarını ortaya koymaktadır. Ayrıca sadece 94 adet bürünsel özellikten oluşan Dur+M1 (Süre+M1) bürünsel özelliklerinin birlikte kullanılması ile 222 adet

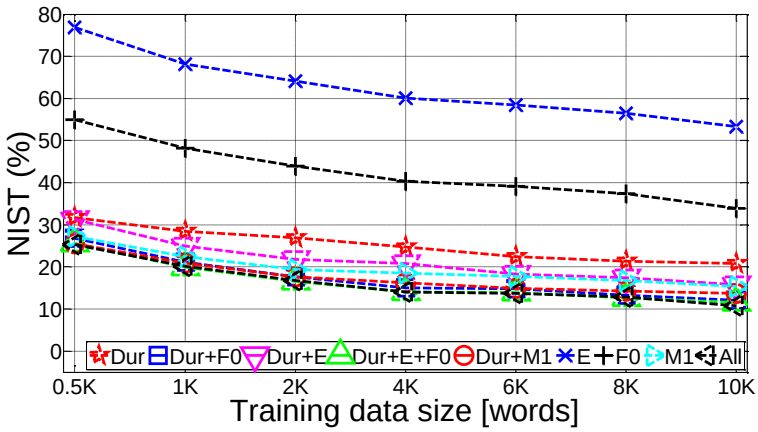
olan tüm bürünsel özelliklerin kullanılması durumundaki başarıma oldukça yakın sonuçlar elde edilmiştir. Bu durumda çok daha az bürünsel özellik kullanılması ile daha hızlı eğitim süreçleri gerçekleştirilmektedir. Diğer taraftan Dur+M1 ile yüksek başarımı beklenen Dur+F0 (142 özellik) dan daha iyi sonuçlar elde edilmiştir. Ayrıca oluşturduğumuz ve sadece 33 özellikten meydana gelen M1 bürünsel özellik setinin kullanılması ile %90 düzeylerinde F-measure değerleri elde edilmiştir (Dalva vd., 2014a; Dalva vd., 2014b).

Tablo 6.3. Birinci ve ikinci konuşmacıların ortalama F-measure ve NIST hata oranları

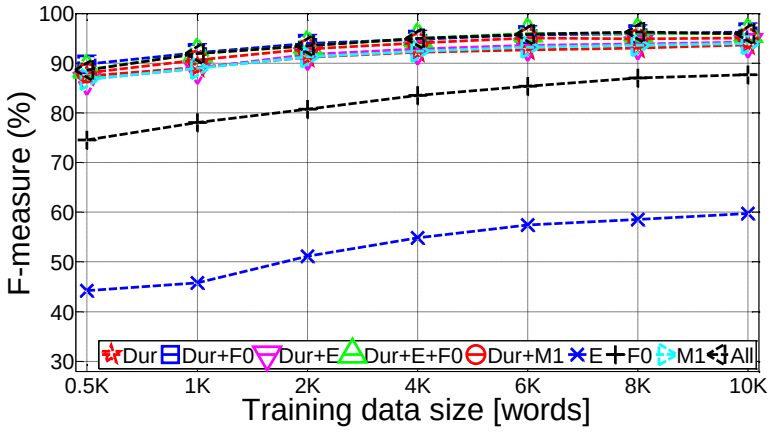
	Eğitim Seti: Konuşmacı 1 Test Seti: Konuşmacı 1		Eğitim Seti: Konuşmacı 2 Test Seti: Konuşmacı 2	
Özellik Kodu	F-measure (%)	NIST(%)	F-measure (%)	NIST(%)
Dur+E+F_0	91.63	16.24	94.06	11.55
All	91.59	16.31	93.84	11.94
Dur+M1	91.13	17.25	93.82	11.98
Dur+ F_0	91.01	17.58	92.93	13.71
M1	90.02	19.60	91.69	16.07
Dur+E	88.89	21.49	91.43	16.66
Dur	86.99	25.19	91.32	16.88
F_0	77.31	42.52	82.41	33.26
E	64.86	62.51	53.09	80.21



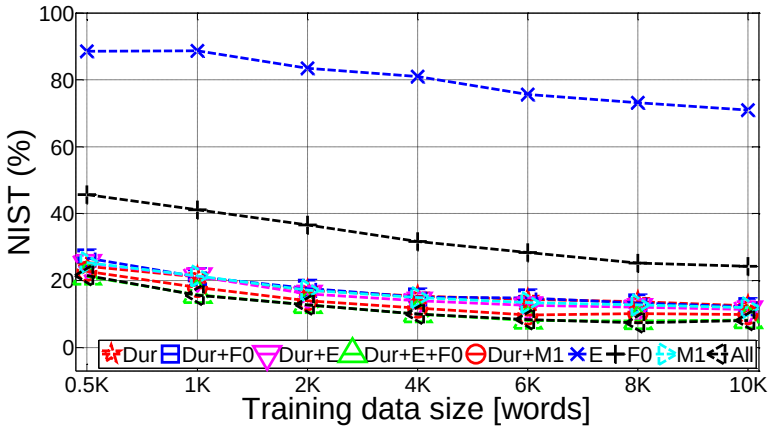
Şekil 6.1. Konuşmacıya ilişkin F-measure skorları



Şekil 6.2. Konuşmacıya ilişkin NIST hata oranları



Şekil 6.3. Konuşmacıya ilişkin F-measure skorları



Şekil 6.4. Konuşmacıya ilişkin NIST hata oranları

6.2 Co-Taining ile Cümle Bölütleme Deneylerine İlişkin Veri Seti Profili

Yapılan deneylere ilişkin kullanılan eğitim seti, geliştirim seti ve test set in yeraldığı veri profili Tablo 6.4 ve Tablo 6.5 de verilmektedir. Deneylerde kullanılan geliştirim ve test kümeleri hem Praat hem de Algemy kullanılarak elde edilen verilerden oluşmaktadır.

Tablo 6.4. Veri seti profili

Training Set	Kelime Sınırı Sayısı	Cümle Sınırı Sayısı	Kelime Sayısı
Praat	57332	4043	61375
Algemy	0	0	0
Toplam	57332	4043	61375
Development Set	Kelime Sınırı Sayısı	Cümle Sınırı Sayısı	Kelime Sayısı
Praat	11549	816	12365
Algemy	8546	622	9168
Toplam	20095	1438	21533
Test Set	Kelime Sınırı Sayısı	Cümle Sınırı Sayısı	Kelime Sayısı
Praat	11572	808	12380
Algemy	8578	592	9170
Toplam	20150	1400	21550

Tablo 6.5. Praat tabanlı purdue prosodic feature extraction tool kullanılarak oluşturulan verinin profili

Set	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Training Set	57332	4043	61375
Development Set	11549	816	12365
Test Set	11572	808	12380

Eğitim Seti: Eğitim Seti'nde yeralan kayıtlara ilişkin konuşmacı bazlı kelime ve cümle sayıları ile birlikte ortam bilgisi Tablo 6.6 da kayıt ortamı bazlı kelime ve cümle sayılarını gösteren bilgiler Tablo 6.7 de yeralmaktadır.

Tablo 6.6. Eğitim seti konuşmacı bazlı kelime ve cümle sayıları ile ortam bilgisi

Konuşmacı	Cinsiyeti	Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Alparslan Esmer	E	Stüdyo	9541	700	10241
Alparslan Esmer	E	Gürültü	2588	222	2810
Arzu Çakır	K	Telefon	2987	175	3162
Aydan Kızıldağlı	E	Stüdyo	600	37	637
Barış ornallı	E	Stüdyo	1554	117	1671
Cem Dalaman	E	Stüdyo	3484	181	3665
Cem Dalaman	E	Gürültü	316	16	332
Değer Akal	K	Telefon	2896	170	3066
Değer Akal	K	Gürültü	706	46	752
Devrim Çubukçu	E	Stüdyo	3912	292	4204
Devrim Çubukçu	E	Gürültü	624	61	685
Elif Özmenek	K	Stüdyo	2584	145	2729
Elif Ural	K	Telefon	1071	71	1142
Güven Özalp	E	Stüdyo	2542	159	2701
Hale Ebiri	K	Stüdyo	15686	1172	16858
Hale Ebiri	K	Gürültü	2557	218	2775
Hülya Polat	K	Stüdyo	724	58	782
Mevlüt Katık	E	Telefon	1363	68	1431
Özge Övün	K	Stüdyo	1123	81	1204
Özge Övün	K	Gürültü	474	54	528
Toplam			57332	4043	61375

Tablo 6.7. Eğitim seti kayıt ortamı bazlı kelime ve cümle sayıları

Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Stüdyo	41750	2942	44692
Gürültü	7265	617	7882
Telefon	8317	484	8801
Toplam	57332	4043	61375

Eğitim seti, tüm deneylerin başlangıcında kayıt tarihlerine göre üç farklı şekilde sıralanmış (3-kez) ve deney başlangıçlarında ilk 1K (1000), 3K (3000), 6K (6000) adet kelime kullanılmıştır. Başlangıç eğitim setlerine ilişkin kelime ve cümle sayıları Tablo 6.8 de verilmektedir. Eğitim seti, kayıt tarihi bazında sıralandığı için kayıt ortamı bazlı kelime ve cümle sayıları verinin tamamında ve Tablo 6.8 de yer alan altkümelerde benzerlik göstermektedir.

Tablo 6.8. Başlangıç eğitim seti verilerine ilişkin bilgiler

Sıralama No	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Sıralama 1	934	66	1000
	2801	199	3000
	5605	395	6000
Sıralama 2	933	67	1000
	2800	200	3000
	5610	390	6000
Sıralama 3	931	69	1000
	2800	200	3000
	5584	416	6000

Geliştirim Seti: Geliştirim Seti'nde yeralan kayıtlara ilişkin konuşmacı bazlı kelime ve cümle sayıları ile birlikte ortam bilgisi Tablo 6.9 da kayıt ortamı bazlı kelime ve cümle sayılarını gösteren bilgiler Tablo 6.10 da yeralmaktadır. Burada gösterilen veri seti, deneylerde kullanılan geliştirim setinin bir alt kümesidir.

Tablo 6.9. Geliştirim seti konuşmacı bazlı kelime ve cümle sayıları ile ortam bilgisi

Konuşmacı	Cinsiyeti	Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Alparslan Esmer	E	Stüdyo	2021	147	2168
Alparslan Esmer	E	Gürültü	674	64	738
Arzu Çakır	K	Telefon	279	20	299
Aydan Kızıldağlı	E	Stüdyo	91	7	98
Barış Ornallı	E	Stüdyo	178	11	189
Cem Dalaman	E	Stüdyo	609	30	639
Cem Dalaman	E	Gürültü	1283	71	1354
Değer Akal	K	Telefon	952	57	1009
DevrimCubukcu	E	Stüdyo	921	71	992
DevrimCubukcu	E	Gürültü	104	9	113
Elif Özmenek	K	Stüdyo	396	26	422
Elif Ural	K	Telefon	248	14	262
Güven Özalp	E	Stüdyo	475	31	506
Hale Ebiri	K	Stüdyo	1857	146	2003
Hale Ebiri	K	Gürültü	670	65	735
Mevlut Katık	E	Telefon	578	30	608
Özge Övün	K	Stüdyo	108	8	116
Özge Övün	K	Gürültü	105	9	114
Toplam			11549	816	12365

Tablo 6.10. Geliştirim seti kayıt ortamı bazlı kelime ve cümle sayıları

Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Stüdyo	6656	477	7133
Gürültü	2836	218	3054
Telefon	2057	121	2178
Toplam	11549	816	12365

Test Set: Test Seti'nde yeralan kayıtlara ilişkin konuşmacı bazlı kelime ve cümle sayıları ile birlikte ortam bilgisi Tablo 6.11 de kayıt ortamı bazlı kelime ve cümle sayılarını gösteren

bilgiler Tablo 6.12 de yeralmaktadır. Burada gösterilen veri seti, deneylerde kullanılan test setinin bir alt kümesidir.

Tablo 6.11. Test seti konuşmacı bazlı kelime ve cümle sayıları ile ortam bilgisi

Konuşmacı	Cinsiyeti	Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Alparslan Esmer	E	Stüdyo	1816	135	1951
Alparslan Esmer	E	Gürültü	343	33	376
Arzu Çakır	K	Telefon	430	22	452
Aydan Kızıldağlı	E	Stüdyo	186	18	204
Barış Ornallı	E	Stüdyo	416	33	449
Cem Dalaman	E	Stüdyo	962	41	1003
Cem Dalaman	E	Gürültü	293	18	311
Değer Akal	K	Telefon	1312	78	1390
Değer Akal	K	Gürültü	254	16	270
Devrim Çubukçu	E	Stüdyo	487	43	530
Devrim Çubukçu	E	Gürültü	111	12	123
Elif Özmenek	K	Stüdyo	151	10	161
Elif Ural	K	Telefon	227	14	241
Güven Özalp	E	Stüdyo	594	38	632
Hale Ebiri	K	Stüdyo	2414	179	2593
Hale Ebiri	K	Gürültü	525	44	569
Hülya Polat	K	Stüdyo	127	9	136
Mevlüt Katık	E	Telefon	276	13	289
Özge Övün	K	Stüdyo	548	39	587
Özge Övün	K	Gürültü	100	13	113
Toplam			11572	808	12380

Tablo 6.12. Test seti kayıt ortamı bazlı kelime ve cümle sayıları

Ortam	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Stüdyo	7701	545	8246
Gürültü	1626	136	1762
Telefon	2245	127	2372
Toplam	11572	808	12380

6.3 Kullanılan Sözcüksel, Bürünsel ve Biçimbilgisel Özellikler

Deneylerin tümünde kullanılan sözcüksel, bürünsel ve biçimbilgisel özellikler aşağıda verilmektedir. Bu bağlamda 34 adet bürünsel özellik, 6 adet sözcüksel özellik ve 10 adet biçimbilgisel özellik kullanılmıştır.

Sözcüksel Özellikler: Deneylerde toplam 6 adet sözcüksel özellik kullanılmıştır.

```
LEX_PREVIOUS: text.
LEX_CURRENT: text.
LEX_NEXT: text.
LEX_CURRENT_NEXT: text.
LEX_PREVIOUS_CURRENT: text.
LEX_TRIAGRAM: text.
```

Bürünsel Özellikler: Deneylerde toplam 34 adet bürünsel özellik kullanılmıştır.

```
PAUSE_DUR: continuous.
PATTERN_BOUNDARY$: text.
SLOPE_DIFF: continuous.
```

```
ENERGY_PATTERN_BOUNDARY: text.
ENERGY_SLOPE_DIFF: continuous.
F0K_WORD_DIFF_HIHI_N: continuous.
F0K_WORD_DIFF_HILO_N: continuous.
F0K_WORD_DIFF_LOLO_N: continuous.
F0K_WORD_DIFF_LOHI_N: continuous.
F0K_WORD_DIFF_MNMN_N: continuous.
F0K_WIN_DIFF_HIHI_N: continuous.
F0K_WIN_DIFF_HILO_N: continuous.
F0K_WIN_DIFF_LOLO_N: continuous.
F0K_WIN_DIFF_LOHI_N: continuous.
F0K_WORD_DIFF_BEGBEG: continuous.
F0K_WORD_DIFF_ENDBEG: continuous.
F0K_INWORD_DIFF: continuous.
LAST_SLOPE: continuous.
LAST_SLOPE_N: continuous.
ENERGY_WORD_DIFF_HIHI_N: continuous.
ENERGY_WORD_DIFF_HILO_N: continuous.
ENERGY_WORD_DIFF_LOLO_N: continuous.
ENERGY_WORD_DIFF_LOHI_N: continuous.
ENERGY_WORD_DIFF_MNMN_N: continuous.
ENERGY_WIN_DIFF_HIHI_N: continuous.
ENERGY_WIN_DIFF_HILO_N: continuous.
ENERGY_WIN_DIFF_LOLO_N: continuous.
ENERGY_WIN_DIFF_LOHI_N: continuous.
ENERGY_WORD_DIFF_BEGBEG: continuous.
ENERGY_WORD_DIFF_ENDBEG: continuous.
ENERGY_INWORD_DIFF: continuous.
ENERGY_LAST_SLOPE: continuous.
ENERGY_LAST_SLOPE_N: continuous.
PAUSE_DUR_PREV: continuous.
```

Biçimbilgisel Özellikler: Deneylerde toplam 10 adet biçimbilgisel özellik kullanılmıştır.

```
lastMarkerA3sg: 1 ,0.
lastMarkerNom: 1, 0.
lastIGhasVerb: 1, 0.
lastPOS: Adj, Adverb, Conj, Det, Dup, Interj, Noun, Num, Postp, Pron, Ques, Verb.
PrevLast3: text.
CurrentLasr3: text.
NextLast3: text.
PrevCurrentLast3: text.
CurrentNextLast3: text.
PrevCurrentNextLast3: text.
```

6.4 Deneysel Başarım Değerlendirme Ölçütleri

Deneylerimizde kullandığımız başarım değerlendirme ölçütleri (evaluation metrics) literatürde Cümle Bölütleme başarım analizlerinde genel olarak kullanılan F-measure ve NIST hata oranı dır.

NIST hata oranı; yanlış karar verilen cümle sınırı kararlarının sayısının referans cümle sınırı sayısına oranıdır.

$$NIST = \frac{f_n + f_p}{t_p + f_n},$$

F-measure ise kesinlik (precision) ve geri getirme (recall) nin harmonik ortalamasıdır.

$$F - measure = \frac{2 \times precision \times recall}{precision + recall},$$

$$precision = \frac{t_p}{t_p + f_p}, \quad recall = \frac{t_p}{t_p + f_n}$$

Burada f_n : yanlış negatif, f_p : yanlış pozitif, ve t_p : gerçek pozitif tir.

6.5 Algoritmalar

6.5.1 Kendi Kendine Eğitme Algoritması

Girişte bulunan veriler:

Eğitim seti: Cümleler bozulmayacak şekilde, örnekler üç farklı sıralama ile bulunmaktadır.

Geliştirim seti: Algoritma akışı boyunca sabit kalacaktır.

Test seti: Algoritma akışı boyunca sabit kalacaktır.

Aşağıdaki algoritma, Eğitim setinin üç farklı sıralama biçimi için ayrı ayrı uygulanacaktır.

Aşama 1: Eğitim kümesinin, etiketlenmiş (L) küme ve etiketlenmemiş (U) küme olmak üzere iki kümeye bölünmesidir.

Eğitim kümesi, etiketlenmiş (L) ve etiketlenmemiş (U) olmak üzere iki ayrı kümeye ayrılır. L kümesinin içerisinde başlangıç olarak 1000, 3000 veya 6000 kelime, onlara karşı gelen özellikler ve orjinal etiketleri ile bulunur. U kümesi ise geriye kalan kelimeler, onlara karşı gelen özellikler ve orjinal etiketleri ile bulunur.

Aşama 2: Başlangıç (Baseline) modelinin eğitimidir.

Not: Algoritma boyunca tüm model eğitimleri, geliştirim kümesi üzerindeki hatayı minimize eden icsiboost (Favre vd., 2007) iterasyonu sayısı ile yapılmaktadır. Maksimum icsiboost iterasyonu limiti 2000 iterasyon olarak belirlenmiştir.

Aşama 3: Kendi kendine eğitim stratejisinin başlangıcıdır. Aşağıdaki adımlar 25 kere tekrarlanacaktır. Her tekrarda, miktarı (100, 250, 500, 1000, 1500) kadar en iyi örnek U kümesinden hipotez edilen etiketleri ile beraber L kümesine taşınır ve U kümesinden kaldırılır. Bu taşıma işlemi aşağıdaki kurallara göre yapılır.

Aşama 3a: En son eğitilen güncel model ile, U kümesinin hipotez etiketleri ve boosting skorları elde edilir.

Not: Başlangıç olarak baseline model kullanılır. Daha sonraki tekrarlarda Aşama 3d'de eğitilen modeller kullanılacaktır.

Aşama 3b: U kümesinin içerisinde yer alan örnekler ve bu örneklere karşı düşen hipotez etiketleri, bu örneklere karşı düşen mutlak boosting skorlarına göre azalan şekilde sıralanır.

Aşama 3c: Mutlak boosting skorları en yüksek olan ilk artış miktarı (Ör: 100 adet) kadar örnek, özellikler ve güncel modelin hipotez ettiği etiketleri ile L kümesine eklenir ve bu örnekler U kümesinden çıkartılır. Bu aşama sonunda eğer artış miktarı 100 ise, L kümesinin yeni eleman sayısı L+100, U kümesinin yeni eleman sayısı U-100 olmuştur. U kümesinde kalan elemanlar orjinal etiketleri ile kalırlar.

Aşama 3d: Güncel L kümesi kullanarak model eğitimi Aşama 2'de açıklanan biçimde yapılır. Böylece model güncellenir.

Aşama 3e: Aşama 3d'de eğitilen modelin geliştirim seti ve test seti üzerinden F-measure başarımları ve NIST hata oranları güncel L kümesinin eleman sayısı ile beraber kaydedilir. 25 tekrar henüz bitmediyse Aşama 3a'ya dönülür.

6.5.2 Kendi Kendine Eğitme Algoritması Kod Taslağı:

- Lex, Pros, Morp, Lex+Morp, Lex+Pros, Pros+Morp özellik setleri ile,
- Başlangıç olarak 1000, 3000 ve 6000 kelimeden oluşan L kümesi ile,
- 25 tekrarla,
- Her tekrarda 100, 250, 500, 1000, 1500 kelimelik artış miktarları ile çalıştırılmıştır.

Başlangıç:

1. El ile etiketlenmiş küçük bir ***L*** (etiketlenmiş veri) kümesi

   $L = \{(x_1, y_1), \dots, (x_{|L|}, y_{|L|})\}$

   $x_i = (x_{i,features})$ ve $y_i = Y \in \{s, n\}$

2. Etiketlenmemiş büyük ***U*** (etiketlenmemiş veri) kümesi

   $U = \{(x_1, \dots, x_{|U|})\}$

Algoritma:

while $\boldsymbol{U \neq \emptyset}$
geliştirim seti üzerinde hata oranı artmadığı veya sabit kalmadığı sürece
Do
L verisi kullanılarak ***M*** modeli eğitilir
for $i = 1{:}|\boldsymbol{U}|$ (Tüm ***U*** elemanları için)
if $|f_M(x_i)| > \theta$
$\boldsymbol{U = U - \{x_i\}}$
$\boldsymbol{L = L \cup \{x_i, H_M(x_i)\}}$
end if
end for
end while

6.5.3 Co-Training Algoritmaları

Co-Training algoritmalarının, Kendi Kendine Eğitme algoritmasından en önemli farkı bu algoritmaların tek bakışlı değil, iki bakışlı algoritmalar olmasıdır. Bu çalışmada üç farklı Co-Training stratejisi kullanılmıştır. Başka bir deyişle, aynı örneklerin iki farklı modelden gelen hipotez etiketlerinin ve boosting skorlarının kullanımı ile çalışır. (Örneğin: Bakış1: Bürünsel

model, Bakış2: Sözcüksel model.) Bu stratejiler; uzlaşma, uzlaşmama ve self-combined stratejisi adları ile anılacaktır. Her üç stratejinin de başlangıç aşaması ortaktır. Co-Training algoritmaları eğitim kümesinin üç farklı dizilimi için ayrı ayrı uygulanmıştır.

6.5.3.1 Co-Training Algoritmaları'nın Başlangıç Aşaması

Girişte bulunan veriler kullanılacak olan her iki bakışa ait özellikleri de (bakış1, bakış2) içerir.
Eğitim seti: Cümleler bozulmayacak şekilde, örnekler üç farklı sıralama ile bulunmaktadır.
Geliştirim seti: Algoritma akışı boyunca örnekler sabit kalacaktır.
Test seti: Algoritma akışı boyunca örnekler sabit kalacaktır.
Başlangıç: Geliştirim seti ve Test seti, bakış1 ve bakış2 özelliklerinden oluşacak şekilde ayrıştırılır. Böylece, Dev_{View1}, Dev_{View2}, $Dev_{View1,View2}$, $Test_{View1}$, $Test_{View2}$ ve $Test_{View1,View2}$ elde edilmiştir. Algoritma akışı boyunca bu kümeler üzerinde herhangi bir değişiklik yapılmayacaktır.
Eğitim kümesinin etiketlenmiş (L) küme ve etiketlenmemiş (U) küme olmak üzere iki kümeye bölünmesidir.
Eğitim kümesi, etiketlenmiş (L) ve etiketlenmemiş (U) olmak üzere iki ayrı kümeye ayrılır. L kümesinin içerisinde başlangıç olarak 1000, 3000 veya 6000 kelime, onlara karşı gelen özellikleri ve orjinal etiketleri ile bulunur. U kümesi ise geriye kalan kelimeler, onlara karşı gelen özellikleri ve orjinal etiketleri ile bulunur.

Başlangıç:

1. El ile etiketlenmiş küçük bir ***L*** kümesi
   $L = \{(x_1, y_1), \dots, (x_{|L|}, y_{|L|})\}$
   $x_i = (x_{i,view1}, x_{i,view2})\ ve\ y_i = Y \in \{s, n\}$
2. Etiketlenmemiş büyük ***U*** kümesi
   $U = \{(x_1, \dots, x_{|U|})\}$

Co-Training Algoritmaları:

- Bakış1=Lex, Bakış2=Morp
- Bakış1=Pros, Bakış2=Lex
- Bakış1=Pros, Bakış2=Morp
- Başlangıç olarak 1000, 3000 ve 6000 kelimeden oluşan (L) kümesiyle
- 25 tekrarla
- Her tekrarda 100, 250, 500, 1000, 1500 kelimelik artış miktarları ile çalıştırılmıştır.

6.5.3.2 Co-Training, Uzlaşma Algoritması

Aşama 1: L ve U kümeleri, bakış1 ve bakış2 özelliklerinden oluşacak şekilde ayrıştırılır. L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilmiş olur.

Aşama 2: L_{View1} verisi kullanılarak M_{View1} modeli ve L_{View2} verisi kullanılarak M_{View2} modeli, $L_{View1,View2}$ verisi kullanılarak $M_{View1,View2}$ modeli eğitilir. M_{View1} modelinin performans ölçümleri $Test_{View1}$ kümesi kullanılarak, M_{View2} modelinin performans ölçümleri $Test_{View2}$ kümesi kullanılarak ve $M_{View1,View2}$ modelinin performans ölçümleri $Test_{View1,View2}$ kümesi kullanılarak yapılır ve bu skorlar kaydedilir. L ve U verilerinin başlangıç durumları için bu skorlar baseline modellerin performans sonuçlarını, yinelemeli bir biçimde değişecek olan L ve U verilerinin her bir yinelemedeki skorları ise güncel modellerin performans sonuçlarını ifade edecektir.

Not: Algoritma boyunca tüm model eğitimleri, ilgili geliştirim seti üzerindeki hatayı minimize eden icsiboost yineleme sayısı ile yapılmaktadır. Maksimum icsiboost yineleme limiti 2000 yineleme olarak belirlenmiştir.

Aşama 3: M_{View1} modeli kullanılarak U_{View1} kümesinin içerisindeki örneklerin hipotez etiketleri ve bu hipotezlere karşı düşen boosting skorları elde edilir. Aynı işlem U_{View2} kümesi için M_{View2} modeli kullanılarak tekrarlanır.

Aşama 4: U kümesinin içerisindeki örnekler, bu örneklere karşı düşen M_{View1} modelinin hipotez ettiği etiket, M_{View1} modelinin mutlak boosting skoru, M_{View2} modelinin hipotez ettiği etiket, M_{View2} modelinin mutlak boosting skoru ve iki modelin mutlak boosting skorları toplamı (uzlaşma skoru) yan yana gelecek şekilde ve uzlaşma skorları azalan sırada olacak şekilde dizilir. Uzlaşma skoru en yüksek olan ilk artış miktarı kadar örnek (100, 250, 500, 1000, 1500) eğer her iki model de (M_{View1} ve M_{View2}) aynı hipotez kararını ("s" veya "n") verdiyse (uzlaşma durumu), U kümesinden kaldırılır ve hipotez edilen etiketi ile L kümesine eklenir.

Aşama 5: Güncel L ve U kümelerinden, güncel L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilir ve Aşama 2 tekrar edilerek güncel modellerin performans skorları elde edilir.

3., 4. ve 5. Aşamalar 25 kere tekrar edilir.

Co-Training, Uzlaşma Algoritması Kod Taslağı:

```
Algoritma:
while U ≠ ∅ ve geliştirim seti üzerinden hata oranı artmadığı veya sabit kalmadığı
sürece
Do
L verisi kullanılarak iki ayrı küme elde edilir
L_view1 = {(x_1,view1, y_1), … , (x_|L|,view1, y_|L|)}
L_view2 = {(x_1,view2, y_1), … , (x_|L|,view2, y_|L|)}
L_view1 kullanılarak M_view1 elde edilir.
L_view2 kullanılarak M_view2 elde edilir.
```

$\boldsymbol{U}$ verisi kullanılarak iki ayrı küme elde edilir
$U_{view1} = \{(x_{1,view1}, y_1), \dots, (x_{|U|,view1}, y_{|U|})\}$
$U_{view2} = \{(x_{1,view2}, y_1), \dots, (x_{|U|,view2}, y_{|U|})\}$
for $i = 1{:}|\boldsymbol{U}|$ (Tüm etiketlenmemiş veri kümesi için)
if $H_{M_{view1}}(x_{i,view1}) = H_{M_{view2}}(x_{i,view2})$ **ve** $|f_{M_{view1}}(x_{i,view1})| + |f_{M_{view2}}(x_{i,view2})| > \theta$
$\boldsymbol{U} = \boldsymbol{U} - \{x_i\}$
$\boldsymbol{L} = \boldsymbol{L} \cup \left\{\left(x_i, H_{M_{view1}}(x_{i,view1})\right)\right\}$
end if
end for
end while

6.5.3.3 Co-Training, Uzlaşmama Algoritması

Aşama 1: L ve U kümeleri, bakış1 ve bakış2 özelliklerinden oluşacak şekilde ayrıştırılır. L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilmiş olur.

Aşama 2: L_{View1} verisi kullanılarak M_{View1} modeli ve L_{View2} verisi kullanılarak M_{View2} modeli, $L_{View1,View2}$ verisi kullanılarak $M_{View1,View2}$ modeli eğitilir. M_{View1} modelinin performans ölçümleri $Test_{View1}$ kümesi kullanılarak, M_{View2} modelinin performans ölçümleri $Test_{View2}$ kümesi kullanılarak ve $M_{View1,View2}$ modelinin performans ölçümleri $Test_{View1,View2}$ kümesi kullanılarak yapılır ve bu skorlar kaydedilir. L ve U verilerinin başlangıç durumları için bu skorlar baseline modellerin performans sonuçlarını, yinelemeli bir biçimde değişecek olan L ve U verilerinin her bir yinelemedeki skorları ise güncel modellerin performans sonuçlarını ifade edecektir.

Not: Algoritma boyunca tüm model eğitimleri, ilgili geliştirim set üzerindeki hatayı minimize eden icsiboost yineleme sayısı ile yapılmaktadır. Maksimum icsiboost yineleme limiti 2000 yineleme olarak belirlenmiştir.

Aşama 3: M_{View1} modeli kullanılarak U_{View1} kümesinin içerisindeki örneklerin hipotez etiketleri ve bu hipotezlere karşı düşen boosting skorları elde edilir. Aynı işlem U_{View2} kümesi için M_{View2} modeli kullanılarak tekrarlanır.

Aşama 4a: U kümesinin içerisindeki örnekler, bu örneklere karşı düşen M_{View1} modelinin hipotez ettiği etiket, M_{View1} modelinin mutlak boosting skoru, M_{View2} modelinin hipotez ettiği etiket, M_{View2} modelinin mutlak boosting skoru ve iki modelin mutlak boosting skorlarının mutlak farkı (uzlaşmama skoru) yan yana gelecek şekilde ve uzlaşmama skorları azalan sırada olacak şekilde dizilir.

Aşama 4b: Her bir örnek için, mutlak boosting skoru diğerine göre daha yüksek olan modelin hipotezi, uzlaşmama stratejisinin hipotezi olarak kabul edilir.

Aşama 4c: Uzlaşmama skoru en yüksek olan ilk artış miktarı kadar örnek (100, 250, 500, 1000, 1500), U kümesinden kaldırılır ve uzlaşmama stratejisi tarafından hipotez edilen etiketi ile L kümesine eklenir.

Aşama 5: Güncel L ve U kümelerinden, güncel L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilir ve Aşama 2 tekrar edilerek güncel modellerin performans skorları elde edilir.

3., 4. ve 5. Aşamalar 25 kere tekrar edilir.

Co-Training, Uzlaşmama Algoritması Kod Taslağı:

```
Algoritma:
while U ≠ ∅ ve geliştirim seti üzerinden hata oranı artmadığı veya sabit kalmadığı
sürece
Do
L verisi kullanılarak iki ayrı küme elde edilir
L_view1 = {(x_1,view1, y_1), ..., (x_|L|,view1, y_|L|)}
L_view2 = {(x_1,view2, y_1), ..., (x_|L|,view2, y_|L|)}
L_view1 kullanılarak M_view1 elde edilir.
L_view2 kullanılarak M_view2 elde edilir.
U verisi kullanılarak iki ayrı küme elde edilir
U_view1 = {(x_1,view1, y_1), ..., (x_|U|,view1, y_|U|)}
U_view2 = {(x_1,view2, y_1), ..., (x_|U|,view2, y_|U|)}
        for i = 1:|U|  (Tüm etiketlenmemiş veri kümesi için)
        if ||f_M_view1(x_i,view1)| − |f_M_view2(x_i,view2)|| > θ
        U = U − {x_i}
                if  |f_M_view1(x_i,view1)| > |f_M_view2(x_i,view2)|
                        L = L ∪ {(x_i, H_M_view1(x_i,view1))}
                else
                        L = L ∪ {(x_i, H_M_view2(x_i,view2))}
                end if
        end if
        end for
end while
```

6.5.3.4 Co-Training, Self-Combined Algoritması

Aşama 1: L ve U kümeleri, bakış1 ve bakış2 özelliklerinden oluşacak şekilde ayrıştırılır. L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilmiş olur.

Aşama 2: L_{View1} verisi kullanılarak M_{View1} modeli ve L_{View2} verisi kullanılarak M_{View2} modeli, $L_{View1,View2}$ verisi kullanılarak $M_{View1,View2}$ modeli eğitilir. M_{View1} modelinin performans ölçümleri $Test_{View1}$ kümesi kullanılarak, M_{View2} modelinin performans ölçümleri $Test_{View2}$ kümesi kullanılarak ve $M_{View1,View2}$ modelinin performans ölçümleri $Test_{View1,View2}$ kümesi kullanılarak yapılır ve bu skorlar kaydedilir. L ve U verilerinin başlangıç durumları için bu skorlar baseline modellerin performans sonuçlarını, yinelemeli bir biçimde değişecek olan L ve U verilerinin her bir yinelemedeki skorları ise güncel modellerin performans sonuçlarını ifade edecektir.

Not: Algoritma boyunca tüm model eğitimleri, ilgili geliştirim seti üzerindeki hatayı minimize eden icsiboost yineleme sayısı ile yapılmaktadır. Maksimum icsiboost yineleme limiti 2000 yineleme olarak belirlenmiştir.

Aşama 3: M_{View1} modeli kullanılarak U_{View1} kümesinin içerisindeki örneklerin hipotez etiketleri ve bu hipotezlere karşı düşen boosting skorları elde edilir. Aynı işlem U_{View2} kümesi için M_{View2} modeli kullanılarak tekrarlanır.

Aşama 4a: U_{View1} kümesinin içerisindeki örnekler, bu örneklere karşı düşen M_{View1} modelinin boosting skorları ve M_{View1} modelinin hipotezleri yan yana gelecek şekilde, M_{View1} modelinin mutlak boosting skorları azalacak şekilde dizilir. M_{View1} modeline göre mutlak skoru en yüksek olan ilk artış miktarı (100, 250, 500, 1000, 1500) kadar örnek eğer her iki model de (M_{View1} ve M_{View2}) aynı hipotez kararını ("s" veya "n") verdiyse (uzlaşma durumu), U kümesinden kaldırılır ve hipotez edilen etiketi ile L kümesine eklenir.

Aşama 4b: U_{View2} kümesinin içerisindeki örnekler, bu örneklere karşı düşen M_{View2} modelinin boosting skorları ve M_{View2} modelinin hipotezleri yan yana gelecek şekilde, M_{View2} modelinin mutlak boosting skorları azalacak şekilde dizilir. M_{View2} modeline göre mutlak skoru en yüksek olan ilk artış miktarı (100, 250, 500, 1000, 1500) kadar örnek eğer her iki model de (M_{View1} ve M_{View2}) aynı hipotez kararını ("s" veya "n") verdiyse (uzlaşma durumu), U kümesinden kaldırılır ve hipotez edilen etiketi ile L kümesine eklenir.

Aşama 5: Güncel L ve U kümelerinden, güncel L_{View1}, L_{View2}, $L_{View1,View2}$, U_{View1}, U_{View2} ve $U_{View1,View2}$ kümeleri elde edilir ve Aşama 2 tekrar edilerek güncel modellerin performans skorları elde edilir.

3., 4. ve 5. Aşamalar 25 kere tekrar edilir.

Co-Training, Self-Combined Algoritması Kod Taslağı:

Algoritma:
while $\boldsymbol{U} \neq \emptyset$ ve geliştirim seti üzerinden hata oranı artmadığı veya sabit kalmadığı sürece
Do
$\boldsymbol{L}$ verisi kullanılarak iki ayrı küme elde edilir
$L_{view1} = \{(x_{1,view1}, y_1), \dots, (x_{|L|,view1}, y_{|L|})\}$
$L_{view2} = \{(x_{1,view2}, y_1), \dots, (x_{|L|,view2}, y_{|L|})\}$
L_{view1} kullanılarak M_{view1} elde edilir.
L_{view2} kullanılarak M_{view2} elde edilir.
$\boldsymbol{U}$ verisi kullanılarak iki ayrı küme elde edilir
$U_{view1} = \{(x_{1,view1}, y_1), \dots, (x_{|U|,view1}, y_{|U|})\}$
$U_{view2} = \{(x_{1,view2}, y_1), \dots, (x_{|U|,view2}, y_{|U|})\}$
for $i = 1:|U_{view1}|$ (Tüm $x_i \epsilon U_{view1}$ elemanları için)
if $|f_{M_{view1}}(x_{i,view1})| > \theta_1$ **ve** $H_{M_{view1}}(x_{i,view1}) = H_{M_{view2}}(x_{i,view2})$
$\boldsymbol{U} = \boldsymbol{U} - \{x_i\}$
$\boldsymbol{L} = \boldsymbol{L} \cup \{(x_i, H_{M_{view1}}(x_{i,view1}))\}$
end if
end for
for $i = 1:|U_{view2}|$ (Tüm $x_i \epsilon U_{view2}$ elemanları için)
if $|f_{M_{view2}}(x_{i,view2})| > \theta_2$ **ve** $H_{M_{view1}}(x_{i,view1}) = H_{M_{view2}}(x_{i,view2})$
$\boldsymbol{U} = \boldsymbol{U} - \{x_i\}$
$\boldsymbol{L} = \boldsymbol{L} \cup \{(x_i, H_{M_{view2}}(x_{i,view2}))\}$

```
        end if
        end for
end while
```

6.6 Eğitim Setinin 3 Farklı Dizilimi (3-kez) Kullanılarak Gerçekleştirilen Deneylerin Ortalama Sonuçları ve Değerlendirme

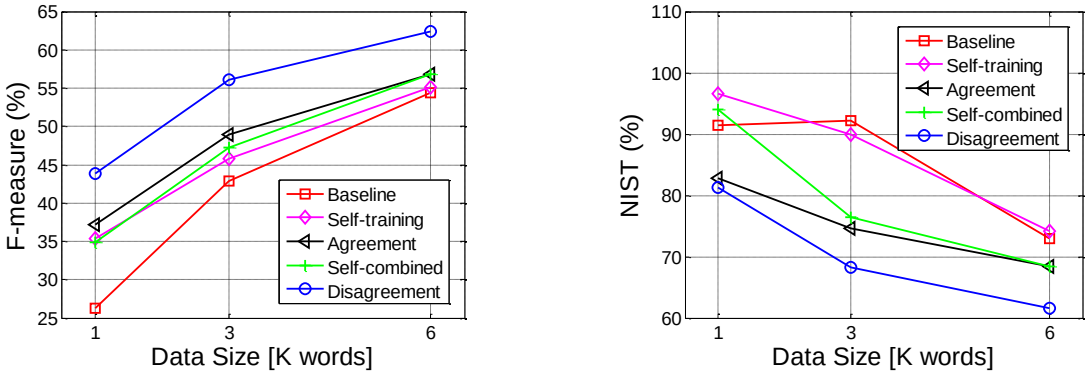
Aşağıdaki tabloda başlangıçta kullanılan eğitim kümelerinin (L) kelime ve cümle sınır sayıları belirtilmiştir. Algoritmalar eğitim kümesinin üç farklı dizilimi için ayrı ayrı çalıştırılmıştır.

Tablo 6.13. 3 Farklı dizilim için eğitim setlerinin kelime ve cümle sınır sayıları

	Kelime Sınırı	Cümle Sınırı	Toplam Kelime
Dizilim 1	934	66	1000
	2801	199	3000
	5605	395	6000
Dizilim 2	933	67	1000
	2800	200	3000
	5610	390	6000
Dizilim 3	931	69	1000
	2800	200	3000
	5584	416	6000

Sözcüksel ve Biçimbilgisel özellikler ile Baseline, Kendi Kendine Eğitme, Co-Training (Uzlaşma, Uzlaşmama ve Self-combined) deneylerinin maksimum performansa (max F-measure) göre üç dizilimin ortalama sonuçları:

Bu deneyde sözcüksel (bakış1) ve biçimbilgisel (bakış2) özellikler kullanılmış ve sadece sözcüksel özellikleri kullanan sınıflandırıcının (sözcüksel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



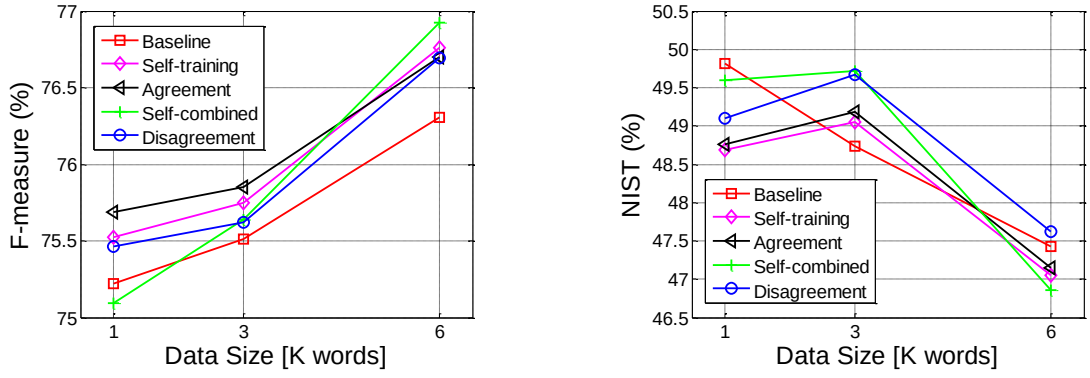
(a) (b)

Şekil 6.5. Sadece sözcüksel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.14. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	26,31	91,38	0	0	0
Kendi Kendine Eğitme	35,41	96,62	1166,67	21,33	26333,33
Uzlaşma	37,17	82,86	1500,00	22,00	34000,00
Self-Combined	34,85	94,07	1333,33	15,67	14591,00
Uzlaşmama	**43,81**	**81,21**	1333,33	18,33	26833,33
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	42,90	92,14	0	0	0
Kendi Kendine Eğitme	45,78	89,88	1000,00	17,33	20333,33
Uzlaşma	48,95	74,55	1000,00	24,33	27333,33
Self-Combined	47,29	76,50	1333,33	21,00	27129,00
Uzlaşmama	**56,12**	**68,26**	1500,00	23,67	38500,00
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	54,41	73,02	0	0	0
Kendi Kendine Eğitme	55,09	74,09	1166,67	17,33	25166,67
Uzlaşma	56,83	68,40	1333,33	21,33	34166,67
Self-Combined	56,83	68,38	1500,00	22,33	35647,33
Uzlaşmama	**62,45**	**61,52**	1500,00	21,00	37500,00

Bu deneyde sözcüksel (bakış1) ve biçimbilgisel (bakış2) özellikler kullanılmış ve sadece biçimbilgisel özellikleri kullanan sınıflandırıcının (biçimbilgisel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



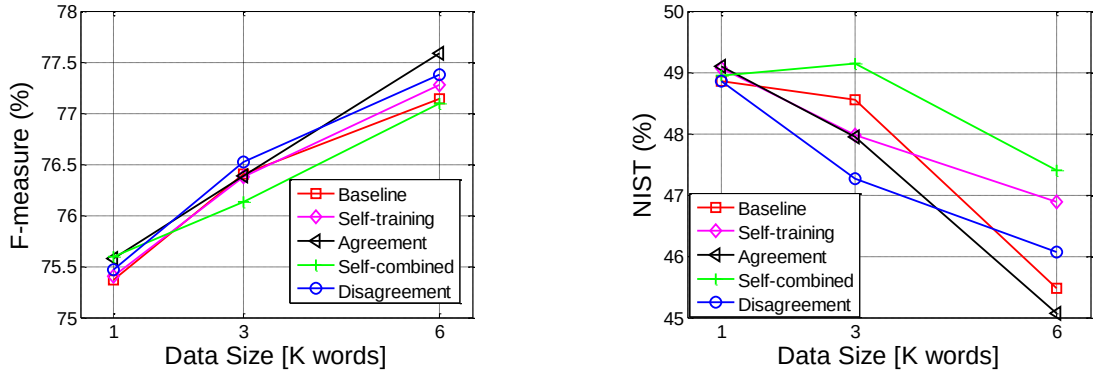
(a) (b)

Şekil 6.6. Sadece biçimbilgisel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.15. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,22	49,81	0	0	0
Kendi Kendine Eğitme	75,52	**48,69**	1033,33	18,00	22866,67
Uzlaşma	**75,69**	48,76	1000,00	19,67	21666,67
Self-Combined	75,10	49,60	233,33	16,33	4358,00
Uzlaşmama	75,47	49,10	1083,33	19,67	21333,33
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,51	**48,74**	0	0	0
Kendi Kendine Eğitme	75,75	49,05	533,33	13,00	13000,00
Uzlaşma	**75,85**	49,19	1033,33	8,67	14133,33
Self-Combined	75,64	49,71	450,00	10,67	5167,67
Uzlaşmama	75,62	49,67	333,33	4,67	5083,33
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	76,30	47,43	0	0	0
Kendi Kendine Eğitme	76,76	47,05	700,00	13,00	19700,00
Uzlaşma	76,70	47,14	1033,33	10,67	19666,67
Self-Combined	**76,92**	**46,86**	616,67	6,33	8419,67
Uzlaşmama	76,69	47,62	283,33	16,00	11500,00

Bu deneyde sözcüksel (bakış1) ve biçimbilgisel (bakış2) özellikler kullanılmış ve hem sözcüksel hem de biçimbilgisel özellikleri kullanan sınıflandırıcıların (sözcüksel+biçimbilgisel model) ürettiği kararlar ile oluşturulan etiketler (s veya n) baz alınmıştır.



(a) (b)

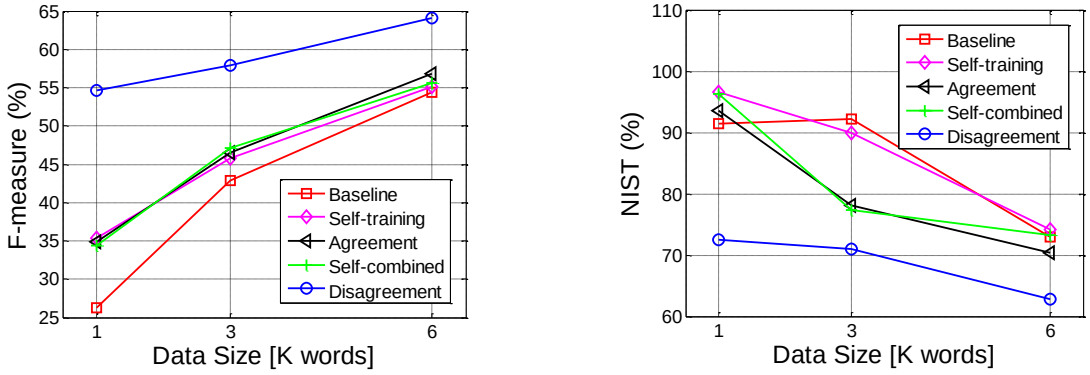
Şekil 6.7. Sözcüksel ve biçimbilgisel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.16. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,37	48,86	0	0	0
Kendi Kendine Eğitme	75,40	49,07	533,33	14,33	8233,33
Uzlaşma	75,58	49,09	1033,33	18,67	17800,00
Self-Combined	**75,59**	48,95	700,00	14,67	4884,00
Uzlaşmama	75,47	**48,86**	833,33	10,00	6666,67
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	76,41	48,55	0	0	0
Kendi Kendine Eğitme	76,38	47,98	866,67	8,33	11933,33
Uzlaşma	76,39	47,95	1166,67	9,67	13666,67
Self-Combined	76,13	49,14	233,33	13,00	7015,00
Uzlaşmama	**76,53**	**47,26**	916,67	10,67	10083,33
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	77,14	45,48	0	0	0
Kendi Kendine Eğitme	77,27	46,88	1333,33	12,00	21666,67
Uzlaşma	**77,59**	**45,07**	1333,33	16,67	27666,67
Self-Combined	77,10	47,40	700,00	19,33	15938,67
Uzlaşmama	77,37	46,07	1000,00	17,67	26000,00

Sözcüksel ve Bürünsel özelliklerin kullanıldığı Baseline, Kendi Kendine Eğitme, Co-Training (Uzlaşma, Uzlaşmama ve Self-combined) deneylerinin maksimum performansa (max F-measure) göre üç dizilimin ortalama sonuçları:

Bu deneyde sözcüksel (bakış1) ve bürünsel (bakış2) özellikler kullanılmış ve sadece sözcüksel özellikleri kullanan sınıflandırıcının (sözcüksel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



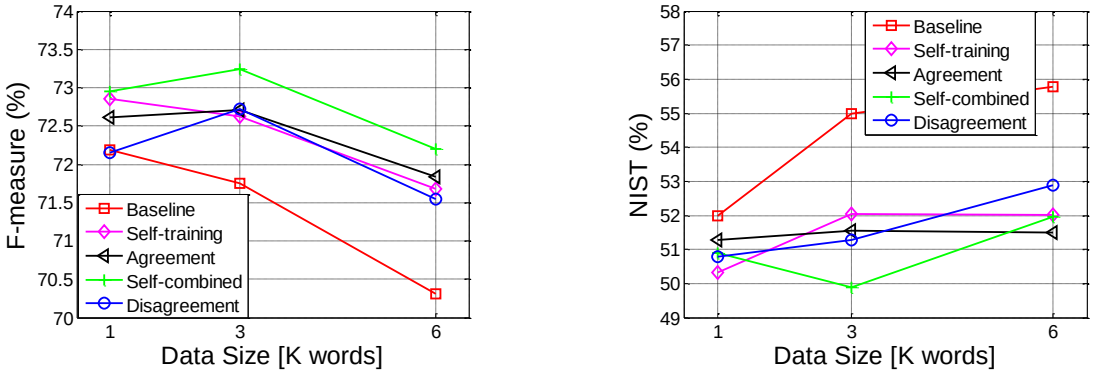
(a) (b)

Şekil 6.8. Sadece sözcüksel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.17. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	26,31	91,38	0	0	0
Kendi Kendine Eğitme	35,41	96,62	1166,67	21,33	26333,33
Uzlaşma	34,83	93,57	1333,33	12,00	16666,67
Self-Combined	34,41	96,29	1000,00	15,33	15003,33
Uzlaşmama	**54,60**	**72,52**	1333,33	25,00	34333,33
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	42,90	92,14	0	0	0
Kendi Kendine Eğitme	45,78	89,88	1000,00	17,33	20333,33
Uzlaşma	46,52	78,12	1333,33	22,33	32666,67
Self-Combined	47,12	77,36	1500,00	20,33	31383,67
Uzlaşmama	**57,92**	**70,98**	1500,00	24,33	39500,00
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	54,41	73,02	0	0	0
Kendi Kendine Eğitme	55,09	74,09	1166,67	17,33	25166,67
Uzlaşma	56,86	70,36	1166,67	24,00	33666,67
Self-Combined	55,63	73,31	1083,33	22,33	27898,00
Uzlaşmama	**64,08**	**62,86**	1500,00	24,00	42000,00

Bu deneyde sözcüksel (bakış1) ve bürünsel (bakış2) özellikler kullanılmış ve sadece bürünsel özellikleri kullanan sınıflandırıcının (bürünsel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



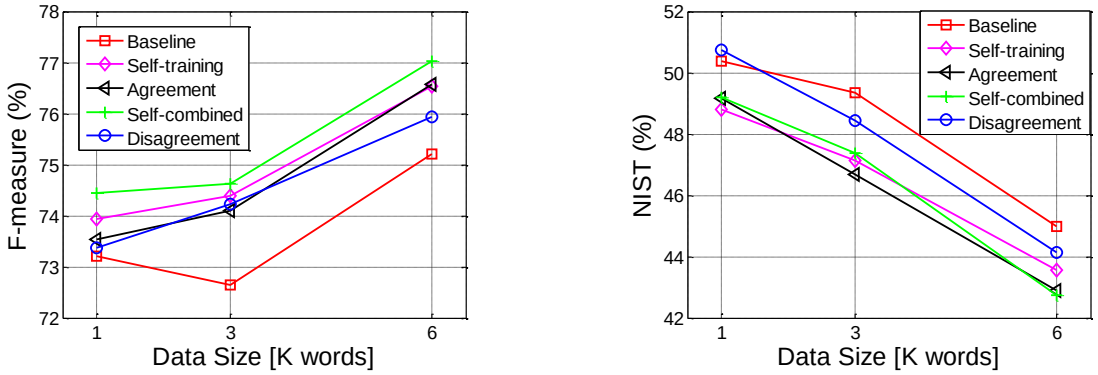
(a) (b)

Şekil 6.9. Sadece bürünsel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.18. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	72,19	52,00	0	0	0
Kendi Kendine Eğitme	72,85	**50,33**	616,67	13,33	13450,00
Uzlaşma	72,61	51,28	150,00	8,33	2083,33
Self-Combined	**72,96**	50,90	200,00	14,33	3628,00
Uzlaşmama	72,16	50,78	400,00	9,00	4600,00
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	71,75	55,00	0	0	0
Kendi Kendine Eğitme	72,63	52,05	700,00	8,67	8966,67
Uzlaşma	72,70	51,55	1166,67	8,67	11833,33
Self-Combined	**73,24**	**49,88**	833,33	8,33	9131,00
Uzlaşmama	72,72	51,29	616,67	6,67	4583,33
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	70,31	55,79	0	0	0
Kendi Kendine Eğitme	71,68	52,02	700,00	13,00	12100,00
Uzlaşma	71,84	**51,50**	1166,67	3,67	10166,67
Self-Combined	**72,20**	51,95	1083,33	10,33	16160,33
Uzlaşmama	71,54	52,88	1033,33	14,33	16766,67

Bu deneyde sözcüksel (bakış1) ve bürünsel (bakış2) özellikler kullanılmış ve hem bürünsel hem de sözcüksel özellikleri kullanan sınıflandırıcıların (bürünsel+sözcüksel model) ürettiği kararlar ile oluşturulan etiketler (s veya n) baz alınmıştır.



(a) (b)

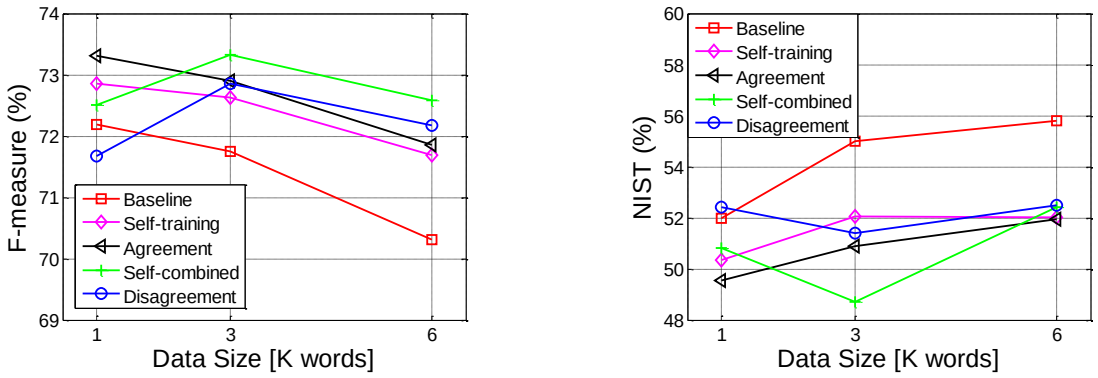
Şekil 6.10. Sözcüksel ve bürünsel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.19. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	73,20	50,38	0	0	0
Kendi Kendine Eğitme	73,94	**48,81**	366,67	8,33	3833,33
Uzlaşma	73,54	49,17	1166,67	9,67	14166,67
Self-Combined	**74,44**	49,19	500,00	13,00	7358,67
Uzlaşmama	73,38	50,74	1166,67	7,33	9000,00
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	72,65	49,33	0	0	0
Kendi Kendine Eğitme	74,40	47,14	700,00	9,33	12033,33
Uzlaşma	74,10	**46,69**	833,33	6,00	7500,00
Self-Combined	**74,62**	47,38	700,00	14,00	14241,33
Uzlaşmama	74,23	48,45	450,00	4,67	6366,67
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,21	45,00	0	0	0
Kendi Kendine Eğitme	76,54	43,57	450,00	13,33	15233,33
Uzlaşma	76,58	42,91	1033,33	1,67	8033,33
Self-Combined	**77,03**	**42,74**	916,67	10,67	17564,00
Uzlaşmama	75,93	44,14	750,00	14,33	20416,67

Bürünsel ve Biçimbilgisel özelliklerin kullanıldığı Baseline, Kendi Kendine Eğitme, Co-Training (Uzlaşma, Uzlaşmama ve Self-combined) deneylerinin maksimum performansa (max F-measure) göre üç dizilimin ortalama sonuçları:

Bu deneyde bürünsel (view 1) ve biçimbilgisel (view 2) özellikler kullanılmış ve sadece bürünsel özellikleri kullanan sınıflandırıcının (bürünsel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



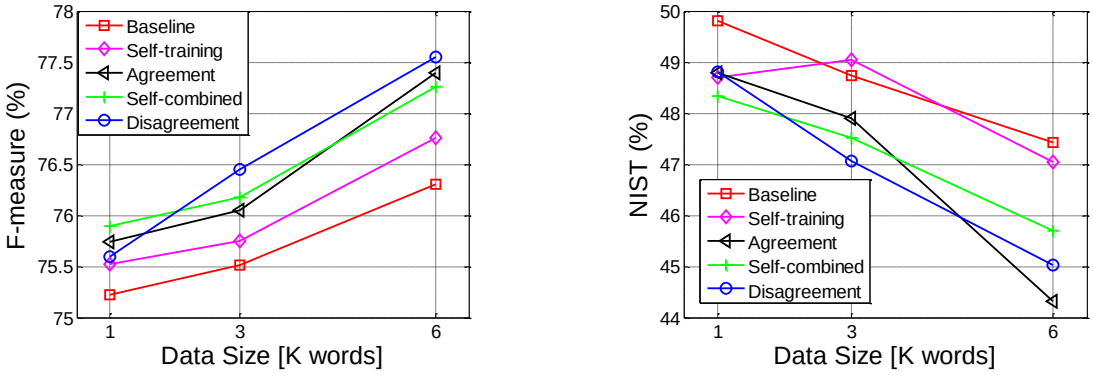
(a) (b)

Şekil 6.11. Sadece bürünsel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.20. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	72,19	52,00	0	0	0
Kendi Kendine Eğitme	72,85	50,33	616,67	13,33	13450,00
Uzlaşma	**73,32**	**49,55**	1166,67	13,67	20166,67
Self-Combined	72,51	50,83	583,33	16,33	9943,33
Uzlaşmama	71,67	52,41	583,33	4,67	3583,33
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	71,75	55,00	0	0	0
Kendi Kendine Eğitme	72,63	52,05	700,00	8,67	8966,67
Uzlaşma	72,90	50,91	1083,33	8,33	10500,00
Self-Combined	**73,33**	**48,71**	700,00	15,33	15859,33
Uzlaşmama	72,85	51,41	283,33	16,33	8416,67
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	70,31	55,79	0	0	0
Kendi Kendine Eğitme	71,68	52,02	700,00	13,00	12100,00
Uzlaşma	71,86	**51,95**	866,67	9,00	10133,33
Self-Combined	**72,59**	52,43	1166,67	16,00	21312,67
Uzlaşmama	72,18	52,48	916,67	9,33	17416,67

Bu deneyde bürünsel (bakış1) ve biçimbilgisel (bakış2) özellikler kullanılmış ve sadece biçimbilgisel özellikleri kullanan sınıflandırıcının (biçimbilgisel model) ürettiği karar ile oluşturulan etiketler (s veya n) baz alınmıştır.



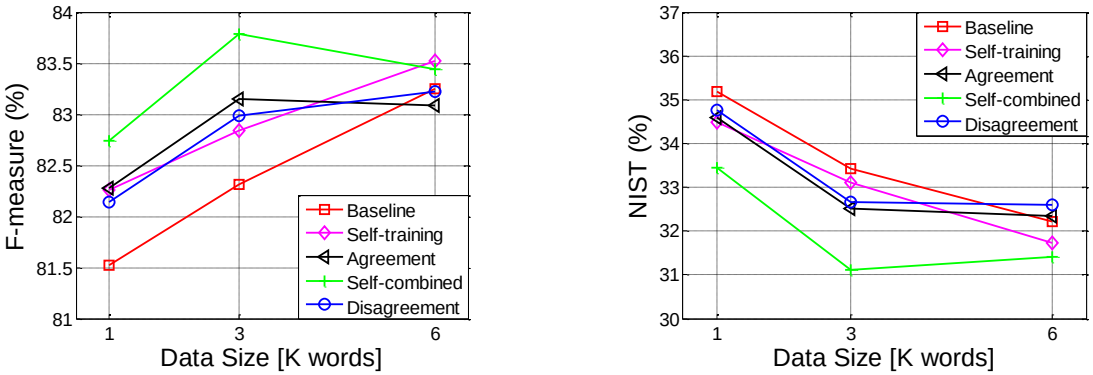
(a) (b)

Şekil 6.12. Sadece biçimbilgisel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.21. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,22	49,81	0	0	0
Kendi Kendine Eğitme	75,52	48,69	1033,33	18,00	22866,67
Uzlaşma	75,74	48,79	533,33	18,00	11500,00
Self-Combined	**75,90**	**48,34**	700,00	23,33	16838,00
Uzlaşmama	75,59	48,81	450,00	8,00	7050,00
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	75,51	48,74	0	0	0
Kendi Kendine Eğitme	75,75	49,05	533,33	13,00	13000,00
Uzlaşma	76,05	47,90	866,67	11,00	14533,33
Self-Combined	76,18	47,52	1083,33	18,33	21906,00
Uzlaşmama	**76,45**	**47,07**	1083,33	13,33	19250,00
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	76,30	47,43	0	0	0
Kendi Kendine Eğitme	76,76	47,05	700,00	13,00	19700,00
Uzlaşma	77,39	44,31	1166,67	20,67	30833,33
Self-Combined	77,26	45,69	1083,33	17,00	25915,33
Uzlaşmama	**77,55**	**45,02**	1000,00	22,00	28000,00

Bu deneyde bürünsel (bakış1) ve biçimbilgisel (bakış2) özellikler kullanılmış ve hem bürünsel hem de biçimbilgisel özellikleri kullanan sınıflandırıcıların (bürünsel+biçimbilgisel) ürettiği kararlar ile oluşturulan etiketler (s veya n) baz alınmıştır.



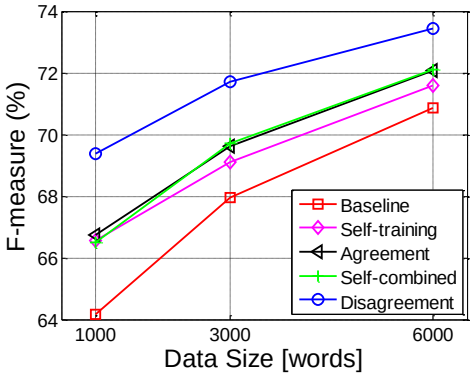
(a) (b)

Şekil 6.13. Bürünsel ve biçimbilgisel özelliklerin kullanıldığı farklı öğrenme stratejilerinin (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

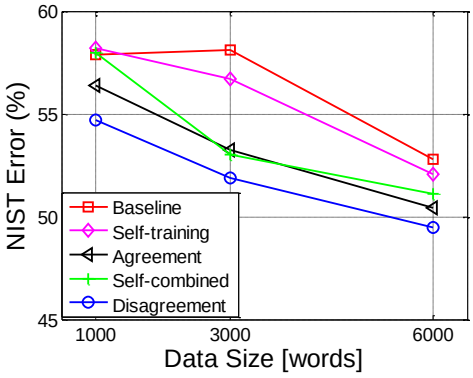
Tablo 6.22. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejileri üzerinde yineleme işleminin etkisi (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	81,52	35,19	0	0	0
Kendi Kendine Eğitme	82,26	34,48	450,00	9,67	4016,67
Uzlaşma	82,27	34,60	750,00	16,33	13583,33
Self-Combined	**82,74**	**33,45**	833,33	10,00	8497,33
Uzlaşmama	82,14	34,76	283,33	8,67	3333,33
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	82,31	33,43	0	0	0
Kendi Kendine Eğitme	82,84	33,12	1000,00	11,00	11666,67
Uzlaşma	83,15	32,52	700,00	7,33	9733,33
Self-Combined	**83,78**	**31,12**	1166,67	12,67	16342,00
Uzlaşmama	82,99	32,67	250,00	15,67	6916,67
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Opt. Artış	Opt. Yineleme	Opt. Veri Büyüklüğü
Baseline	83,25	32,21	0	0	0
Kendi Kendine Eğitme	**83,53**	31,74	450,00	17,67	13216,67
Uzlaşma	83,09	32,33	700,00	13,33	14466,67
Self-Combined	83,44	**31,41**	833,33	15,33	19884,33
Uzlaşmama	83,22	32,60	533,33	11,00	9566,67

Strateji bazında özelliklerden bağımsız ortalama sonuçlar:



(a)



(b)

Şekil 6.14. Farklı öğrenme stratejilerin ortalama (a) F-measure ve (b) NIST hata oranları (max F-measure' a göre)

Tablo 6.23. Sadece 1K, 3K ve 6K elle etiketlenmiş veri bulunması durumunda farklı öğrenme stratejilerinin ortalama sonuçları (max F-measure' a göre)

Elle Etiketenmiş Veri = 1K	F(%)	NIST(%)	Ortalama Artış Değeri	Ortalama Yineleme Değeri	Ortalama Eklenen Örnek Sayısı
Baseline	64.17	57.86	0	0	0
Kendi Kendine Eğitme	66.57	58.18	775,9267	15,2944	14709,2589
Uzlaşma	66.75	56.40	959,2589	15,3711	15848,1489
Self-Combined	66.50	57.95	675,9244	15,4433	9355,7400
Uzlaşmama	**69.36**	**54.68**	829,6278	12,2967	11970,3689
Elle Etiketenmiş Veri = 3K	F(%)	NIST(%)	Ortalama Artış Değeri	Ortalama Yineleme Değeri	Ortalama Eklenen Örnek Sayısı
Baseline	67.96	58.11	0	0	0
Kendi Kendine Eğitme	69.11	56.68	781,4811	11,8511	10359,2589
Uzlaşma	69.62	53.26	1020,3700	11,8144	12766,6656
Self-Combined	69.70	53.03	888,8878	14,8511	13463,8889
Uzlaşmama	**71.71**	**51.89**	770,3700	13,3344	12411,1111
Elle Etiketenmiş Veri = 6K	F(%)	NIST(%)	Ortalama Artış Değeri	Ortalama Yineleme Değeri	Ortalama Eklenen Örnek Sayısı
Baseline	70.84	52.79	0	0	0
Kendi Kendine Eğitme	71.60	52.05	818,5189	14,4067	18227,7789
Uzlaşma	72.02	50.44	1088,8889	13,4456	20977,7789
Self-Combined	72.11	51.13	998,1478	15,5167	20971,1478
Uzlaşmama	**73.44**	**49.46**	946,2956	16,6289	23240,7422

Tablo 6.24. Öğrenme stratejilerinin ortalama sonuçları (max F-measure' a göre)

Öğrenme Stratejisi	F-measure (%)	NIST (%)
Baseline	67.66	56.26
Kendi Kendine Eğitme	69.09	55.64
Uzlaşma	69.48	53.37
Self-Combined	69.43	54.04
Uzlaşmama	**71.50**	**52.01**

Sözcüksel ve Biçimbilgisel özelliklerin kullanıldığı deneylerde, sadece sözcüksel model kullanılarak edilen sonuçlar Şekil 6.5 ve Tablo 6.14 de verilmektedir. Sadece biçimbilgisel modelin kullanılması ile elde edilen sonuçlar Şekil 6.6 ve Tablo 6.15 de verilmektedir. Hem sözcüksel hem de biçimbilgisel modellerin birlikte kullanılması ile elde edilen sonuçlar Şekil 6.7 ve Tablo 6.16 da verilmektedir. Oldukça az miktarda (1K) elle etiketlenmiş örnek bulunması durumunda sadece sözcüksel modelin kullanılması durumunda co-training uzlaşmama stratejisi ile cümle bölütleme performansının %26.31 den %43.81 e yükseldiği ve %66.51 düzeyinde bir geliştirim elde edildiği görülmüştür. Sadece biçimbilgisel modelin sözkonusu olduğu durumda yine co-training stratejilerimizden biri olan uzlaşma stratejisi ile cümle bölütleme performansının %75.22 den %75.69 a yükseldiği ve %0,62 düzeyinde bir geliştirim sağlandığı görülmüştür. Sözcüksel ve biçimbilgisel modellerin birlikte kullanıldığı durumda co-training stratejilerimizden biri olan self-combined stratejisi ile cümle bölütleme performansının %75.37 den %75.59 a yükseldiği ve %0.29 oranında geliştirim sağladığı görülmüştür.

Sözcüksel ve Bürünsel özelliklerin kullanıldığı deneylerde, sadece sözcüksel model kullanılarak edilen sonuçlar Şekil 6.8 ve Tablo 6.17 de verilmektedir. Sadece bürünsel modelin kullanılması ile elde edilen sonuçlar Şekil 6.9 ve Tablo 6.18 de verilmektedir. Hem sözcüksel hem de bürünsel modellerin birlikte kullanılması ile elde edilen sonuçlar Şekil 6.10 ve Tablo 6.19 da verilmektedir. Oldukça az miktarda (1K) elle etiketlenmiş örnek bulunması durumunda sadece sözcüksel modelin kullanılması durumunda co-training uzlaşmama stratejisi ile cümle bölütleme performansının %26.31 den %54.60 a yükseldiği ve %107.53 düzeyinde bir geliştirim elde edildiği görülmüştür. Sadece bürünsel modelin sözkonusu olduğu durumda yine co-training stratejilerimizden biri olan self-combined stratejisi ile cümle bölütleme performansının %72.19 dan %72.96 ya yükseldiği ve %1.07 düzeyinde bir geliştirim sağlandığı görülmüştür. Sözcüksel ve bürünsel modellerin birlikte kullanıldığı durumda co-training stratejilerimizden biri olan self-combined stratejisi ile cümle bölütleme performansının %73.20 den %74.44 e yükseldiği ve %1.69 oranında geliştirim sağladığı görülmüştür.

Bürünsel ve Biçimbilgisel özelliklerin kullanıldığı deneylerde, sadece bürünsel model kullanılarak edilen sonuçlar Şekil 6.11 ve Tablo 6.20 de verilmektedir. Sadece biçimbilgisel modelin kullanılması ile elde edilen sonuçlar Şekil 6.12 ve Tablo 6.21 de verilmektedir. Hem bürünsel hem de biçimbilgisel modellerin birlikte kullanılması ile elde edilen sonuçlar Şekil 6.13 ve Tablo 6.22 de verilmektedir. Oldukça az miktarda (1K) elle etiketlenmiş örnek bulunması durumunda sadece bürünsel modelin kullanılması durumunda co-training uzlaşma

stratejisi ile cümle bölütleme performansının %72.19 dan %73.32 ye yükseldiği ve %1.57 düzeyinde bir geliştirim elde edildiği görülmüştür. Sadece biçimbilgisel modelin sözkonusu olduğu durumda yine co-training stratejilerimizden biri olan self-combined stratejisi ile cümle bölütleme performansının %75.22 den %75.90 a yükseldiği ve %0,90 düzeyinde bir geliştirim sağlandığı görülmüştür. Bürünsel ve biçimbilgisel modellerin birlikte kullanıldığı durumda co-training stratejilerimizden biri olan self-combined stratejisi ile cümle bölütleme performansının %81.52 den %82.74 e yükseldiği ve %1.5 oranında geliştirim sağladığı görülmüştür.

3 farklı eğitim seti diziliminden ve yukarıda belirtilen 9 adet farklı model kombinasyonundan elde edilen tüm sonuçlar strateji bazında değerlendirilmiş ve ortalama performans değerleri Şekil 6.14, Tablo 6.23 ve Tablo 6.24 de özetlenmiştir. Elde edilen sonuçlar değerlendirildiğinde co-training stratejilerinin tümünün baseline ve kendi kendine eğitme performanslarından daha iyi olduğu görülmektedir. Örneğin sadece 1K elle etiketlenmiş örnek mevcut olduğunda ortalama cümle bölütleme performansının en yüksek düzeyde co-training uzlaşmama stratejisi ile gerçekleştirildiği ve performansın baseline a göre %64.17 den %69.36 ya yükseldiği, başka bir deyiş ile %8.09 oranında bir geliştirim sağlandığı görülmüştür. 1K, 3K ve 6K elle etiketlenmiş veri miktarları için elde edilen sonuçların ortalaması incelendiğinde en yüksek cümle bölütleme performansının co-training uzlaşmama stratejisi ile gerçekleştirildiği ve baseline a göre başarımın %67.66 dan %71.50 ye yükseldiği, başka bir ifade ile %5.68 oranında bir iyileştirme elde edildiği görülmüştür.

Beklendiği üzere tüm elle etiketlenmiş başlangıç veri büyüklüklerinde (1K, 3K ve 6K) etiketlenmiş veri sayısı arttıkça cümle bölütleme performansları birbirine yaklaşmaktadır. Tüm modellerde görüldüğü üzere co-training stratejileri özellikle çok az sayıda elle etiketlenmiş veri olması durumunda baseline performansını ciddi bir biçimde geliştirmektedir.

Sözcüksel modellerin daha yüksek performans göstermelerinin bir nedeni de konuşma verilerinin toplantı verileri yerine haber kaynaklı veriler olmasıdır. Zira haber kaynaklı konuşma verileri belirli konuşmacıların uzun süre tekdüze konuştuğu kayıtlardan oluşmaktadır. Oysa toplantı konuşma verileri bir grup konuşmacının spontane ve doğal konuşmaya daha yakın konuştuğu verilerdir. Bürünsel modellerin doğru bir biçimde cümle sonu olarak etiketlediği ve fakat sözcüksel modellerin cümle sonu kararını hatalı verdiği durumlar incelenmiştir. Özellikle konuşmanın başka biri tarafından araya girilerek kesintiye uğratıldığı (disruptions) ve konuşmacının bir cümleye başladıktan sonra cümlesini keserek başka bir cümle ile yeniden ifade etme ihtiyacı hissettiği (disfluency) durumların, haber verilerinde daha az olduğu görülmektedir. Sözcüksel model bu örneklerde genellikle cümle

sonlarını doğru bir biçimde tespit edemezken bürünsel model doğru bir biçimde etiketleme yapabilmektedir. Bu tür örneklerin haber konuşma verilerinde az olması, sözcüksel modellerin başarımını bürünsel modellerin başarımının üzerine çıkarmaktadır. Ayrıca bürünsel modellerin performanslarını doğrudan etkileyen bir başka önemli faktör de ses dosyalarında bulunan gürültü, müzik gibi, arkaplan gürültüleridir.

Bazı deney sonuçlarında elde edilen (örneğin uzlaşma ve kendi kendine eğitme stratejilerine ilişkin) gerek F-measure gerekse NIST hata oranlarındaki yaklaşıklığın gerçek hayata yansıması; değerlendirme ölçütlerinin tanımından hareketle, modellerin doğru bir biçimde etiketlediği verilerin gerçekte doğru bir biçimde etiketlenmiş referans verilerine yüzde olarak oranı biçiminde tanımlanabilir.

F-measure değerinin olabildiğince yüksek, NIST hata oranının ise olabildiğince düşük olması istenir. Ancak otomatik olarak etiketlenecek verinin milyonlarca kelime (dolayısı ile binlerce cümle sonu) olduğu dikkate alındığında yüzdelik küçük farklılıklar gerçekte cümle sınırı olarak doğru veya yanlış etiketlenmiş veri sayısını doğrudan etkileyecektir. Dolayısı ile bazı deneylerde birbirine yakın gibi görünen gerek F-measure gerekse NIST hata oranları otomatik olarak etiketlenecek veri büyüklüğü arttıkça daha önemli bir anlam ifade edecektir.

6.7 SRI-Algemy ve PRAAT tabanlı Purdue Prosodic Feature Extraction Tool Karşılaştırması

Bu deney setinde SRI'ın Algemy yazılımı ile çıkarılan bürünsel özellikler ile açık kaynak kodlu PRAAT tabanlı Purdue Prosodic Feature Extraction Tool kullanılarak çıkarılan 34 adet ortak bürünsel özellik kullanılmıştır. Deney setinde çok az sayıda elle etiketlenmiş (1K, 3K ve 6K) eğitim verisi kullanılmıştır. Çok az sayıda elle etiketlenmiş ve sadece bürünsel özellikler kullanılarak elde edilen modellerin cümle bölütleme başarımları karşılaştırılmış ve böylece geri kalan etiketlenmemiş veri üzerinde cümle sınırlarının hangi başarımla doğru bir biçimde işaretlendiği tespit edilmiştir. Deney setinde kullanılan Eğitim, Geliştirim ve Test setleri önceki deneylerde kullanılan verilerin aynısıdır.

Tüm baseline model eğitimleri, geliştirim seti üzerindeki hatayı minimize eden (min NIST error) icsiboost yineleme sayısı ile yapılmaktadır. Maksimum icsiboost yineleme limiti 2000 yineleme olarak belirlenmiştir. Optimum model kullanılarak test set üzerinden başarım ve hatalar bir log dosyasına kaydedilmiştir. Baseline deneyler SRI' ın Algemy ve Purdue Prosodic Feature Extraction Tool ile ayrı ayrı yapılmıştır.

Algemy ve Purdue Prosodic Feature Extraction Tool ile çıkarılan ve ortak 34 adet bürünsel özellikten oluşan, 1K, 3K ve 6K büyüklüğündeki modellerin kullanılması ile elde edilen baseline sonuçları ve karşılaştırılması Tablo 6.25 de verilmektedir.

Tablo 6.25. SRI-Algemy ve praat tabanlı purdue prosodic feature extraction tool yazılımları ile hesaplanan bürünsel özelliklerin kullanıldığı cümle bölütleme baseline değerleri (F-measure ve NIST hata oranları ve ortalama değerleri)

Veri Büyüklüğü (Kelimeler)		1K	3K	6K	Ortalama
F-measure (%)	Purdue	76,9478	76,7717	76,6567	76,7920
	Algemy	80,0791	79,3052	81,9690	80,4511
NIST (%)	Purdue	41,8317	41,5017	42,3267	41,8867
	Algemy	36,9224	38,0363	33,5396	36,1661

Tablo 6.25 de gösterilen sonuçlar değerlendirildiğinde SRI trafafından geliştirilen DECIPHER konuşmacıdan bağımsız konuşma tanıma sistemi (DECIPHER Speaker Independent Speech Recognition System) nin performansının SRI ın ALGEMY yazılımı tabanlı bürünsel özelliklerin çıkarıldığı yazılım üzerinde performans artışı yarattığı gözlenmiştir. Zira DECIPHER gerek arkaplan gürültüsünü giderme gerekse konuşmacı ayrımını oldukça yüksek performansta gerçekleştirmektedir. Bu nedenle, oldukça sofistike algoritmalar kullanılarak hesaplanan SRI ın Algemy yazılımı tabanlı bürünsel özellikler ile elde edilen cümle bölütleme performansının tamamen açık kaynak kodlu PRAAT yazılımı ile gerçekleştirilen Purdue Prosodic Feature Extraction Tool tabanlı bürünsel özellikler ile elde edilen cümle bölütleme performansının max F-measure ve min NIST hata oranı ölçütleri ile değerlendirildiğinde sırası ile %4,5 ve %13 üzerinde olduğu görülmektedir. Elde edilen sonuçlar ALGEMY ile elde edilen sonuçlar ile karşılaştırıldığında tamamen açık kaynak kodlu bir yazılım ile özellikle haber verileri gibi spontane olmayan ve tekdüze konuşulan (less informative) bir veri profili için oldukça başarılıdır. Bu sonuçların ICSI-MRDA (ICSI-Meeting Recorder Dialog Act) gibi bir grup konuşmacının toplantı verilerinin bulunduğu bir veri profilinde Praat tabanlı Purdue Prosodic Feature Extraction Tool lehinde daha da geliştirileceği söylenebilir. Zira toplantı verilerinde bürünsel modellerin cümle bölütleme performansını arttıran konuşmanın başka biri tarafından araya girilerek kesintiye uğratıldığı ve konuşmacının bir cümleye başladıktan sonra cümlesini keserek başka bir cümle ile yeniden ifade etme ihtiyacı hissettiği durumlara çokça rastlanmaktadır. Oysa haber verilerinde uzun süre belirli bir konuşmacı hakimiyeti sözkonusu olmaktadır. Dolayısı ile Purdue Prosodic Feature Extraction Tool ile elde edilen bürünsel özelliklerin kullanıldığı bürünsel modellerin, arkaplan gürültüsü, müzik-konuşma örtüşmeleri gibi bürünsel özelliklerin çıkarılmasını zorlaştıran durumlardan etkilenmemesi durumunda cümle bölütleme performansının daha da artacağı sonucu çıkarılabilmektedir.

6.8 İstatistiksel Analiz Yöntemleri ile Deneylerin Sonuçlarının Değerlendirilmesi

İkiden fazla algoritmanın performans ve hata farklılıklarının arasındaki istatistiksel anlamı ölçmek için aşağıda aşamaları anlatılan Repeated-Measures of Anova testi kullanılmıştır. Bu testte kullanılan F-measure ve NIST değerleri bir önceki bölümde 3 farklı eğitim seti dizilimi için elde edilen sonuçların ortalama değerleridir. Aşağıda belirtilen 5 aşama gerçekleştirilerek Tablo 6.26 - Tablo 6.32 de verilen sonuçlar elde edilmiştir. Elde edilen sonuçlar değerlendirildiğinde algoritmaların başarım ve hata oranları arasında gözlenebilir ve anlamlı bir fark bulunmaktadır.

Aşama 1:

Hipotezler belirlenir ve alfa değeri seçilir.

H_0: Tüm algoritmaların başarım ve hata oranları birbirine eşit gibi düşünülebilir.

H_1: Algoritmaların başarım ve hata oranları arasında gözlenebilir bir fark vardır.

α = 0.05 (F-dağılımı fonksyonunun integrali alındığında, alanın uçlardaki %5'lik kısmı kritik bölge, bir başka deyişle H_1 hipotezinin doğru kabul edildiği bölge olarak seçilmiştir.)

Aşama 2:

$$SS_{total} = \sum X^2 - \frac{G^2}{N}$$

X : Her bir deneyin F-measure veya Nist değeri. (Tabloların her bir hücresi.)

G : Algoritma bazında yapılan tüm deneylerin F-measure veya Nist değerlerinin toplamı. (Tablodaki her bir hücrenin toplamı.)

N : Toplam deney sayısı.

SS: Sapmaların toplamı.

$$SS = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$SS_{within} = \sum SS_{X,Algotirma}$$

SS_{within} : Algoritma bazındaki tüm sapma toplamlarının toplamı.

$$SS_{between} = \sum \frac{T^2}{n} - \frac{G^2}{N}$$

T : Bir algoritma ile yapılan tüm deney kombinasyonlarındaki F-measure veya Nist değerlerinin toplamı. (Tabloda kolon bazında toplam.)

$$df_{total} = N - 1$$

df : Serbestiyet derecesi

$$df_{within} = \sum df$$

df_{within}: Deney kombinasyonları (tablodaki her

bir satır) arasındaki serbestiyet derecesi.

$df_{between} = k - 1$

k : Algoritma sayısı

Aşama 3:

$SS_{between\ subjects} = \sum \frac{P^2}{k} - \frac{G^2}{N}$

P : Deney seti bazında tüm F-measure veya Nist değerlerinin toplamı.

$SS_{error} = SS_{within} - SS_{between\ subjects}$

$df_{between\ subjects} = n - 1$

$df_{error} = df_{within}$-$df_{between\ subjects}$

$MS_{between} = \frac{SS_{between}}{df_{between}}$

MS: Ortalamaların karesi

$MS_{error} = \frac{SS_{error}}{df_{error}}$

$F\ ratio = \frac{MS_{between}}{MS_{error}}$

Aşama 4:

F dağılım tablosundan $df(df_{between}, df_{error})$ ve α = 0.05 değerleri kullanılarak kritik F $(F\ critic)$ değere bakılır. Eğer $|F\ ratio| > |F\ critic|$ sağlanıyorsa H_1 hipotezi doğru kabul edilir.

Tablo 6.26. Sözcüksel modelin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	26,31	35,41	37,17	34,85	43,81
LEX+MORP	3000	42,90	45,78	48,95	47,29	56,12
LEX+MORP	6000	54,41	55,09	56,83	56,83	62,45
LEX+PROS	1000	26,31	35,41	34,83	34,41	54,60
LEX+PROS	3000	42,90	45,78	46,52	47,12	57,92
LEX+PROS	6000	54,41	55,09	56,86	55,63	64,08
Ortalama		41,21	45,43	46,86	46,02	**56,50**
$F\ ratio$	23,0593					
$F\ critic$	2,8660814					

NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	91,38	96,62	82,86	94,07	81,21
LEX+MORP	3000	92,14	89,88	74,55	76,50	68,26
LEX+MORP	6000	73,02	74,09	68,40	68,38	61,52
LEX+PROS	1000	91,38	96,62	93,57	96,29	75,52
LEX+PROS	3000	92,14	89,88	78,12	77,36	70,98
LEX+PROS	6000	73,02	74,09	70,36	73,31	62,86
Ortalama		85,51	86,86	77,98	80,99	**70,06**
$F\ ratio$	16,076					
$F\ critic$	2,8660814					

Tablo 6.27. Biçimbilgisel modelin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	75,22	75,52	75,69	75,10	75,47
LEX+MORP	3000	75,51	75,75	75,85	75,64	75,62
LEX+MORP	6000	76,30	76,76	76,70	76,92	76,69
PROS+MORP	1000	75,22	75,52	75,74	75,90	75,59
PROS+MORP	3000	75,51	75,75	76,05	76,18	76,45
PROS+MORP	6000	76,30	76,76	77,39	77,26	77,55
Ortalama		75,68	76,01	**76,24**	76,17	76,23
$F\ ratio$	6,3775					
$F\ critic$	2,8660814					

NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	49,81	48,69	48,76	49,60	49,10
LEX+MORP	3000	48,74	49,05	49,19	49,71	49,67
LEX+MORP	6000	47,43	47,05	47,14	46,86	47,62
PROS+MORP	1000	49,81	48,69	48,79	48,34	48,81
PROS+MORP	3000	48,74	49,05	47,90	47,52	47,07
PROS+MORP	6000	47,43	47,05	44,31	45,69	45,02
Ortalama		48,66	48,26	**47,68**	47,95	47,88
$F\ ratio$	1,8422					
$F\ critic$	2,8660814					

Tablo 6.28. Bürünsel modelin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+PROS	1000	72,19	72,85	72,61	72,96	72,16
LEX+PROS	3000	71,75	72,63	72,70	73,24	72,72
LEX+PROS	6000	70,31	71,68	71,84	72,20	71,54
PROS+MORP	1000	72,19	72,85	73,32	72,51	71,67
PROS+MORP	3000	71,75	72,63	72,90	73,33	72,85
PROS+MORP	6000	70,31	71,68	71,86	72,59	72,18
Ortalama		71,42	72,39	72,54	**72,81**	72,19
$F\ ratio$	10,9322					
$F\ critic$	2,8660814					
NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+PROS	1000	52,00	50,33	51,28	50,90	50,78
LEX+PROS	3000	55,00	52,05	51,55	49,88	51,29
LEX+PROS	6000	55,79	52,02	51,50	51,95	52,88
PROS+MORP	1000	52,00	50,33	49,55	50,83	52,41
PROS+MORP	3000	55,00	52,05	50,91	48,71	51,41
PROS+MORP	6000	55,79	52,02	51,95	52,43	52,48
Ortalama		54,26	51,47	51,12	**50,78**	51,88
$F\ ratio$	12,3008					
$F\ critic$	2,8660814					

Tablo 6.29. Sözcüksel ve biçimbilgisel modellerin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	75,37	75,40	75,58	75,59	75,47
LEX+MORP	3000	76,41	76,38	76,39	76,13	76,53
LEX+MORP	6000	77,14	77,27	77,59	77,10	77,37
Ortalama		76,31	76,35	**76,52**	76,27	76,46
$F\ ratio$	1,7033					
$F\ critic$	3,83785335					
NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP	1000	48,86	49,07	49,09	48,95	48,86
LEX+MORP	3000	48,55	47,98	47,95	49,14	47,26
LEX+MORP	6000	45,48	46,88	45,07	47,40	46,07
Ortalama		47,63	47,98	**47,37**	48,50	47,40
$F\ ratio$	1,7273					
$F\ critic$	3,83785335					

Tablo 6.30. Sözcüksel ve bürünsel modellerin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+PROS	1000	73,20	73,94	73,54	74,44	73,38
LEX+PROS	3000	72,65	74,40	74,10	74,62	74,23
LEX+PROS	6000	75,21	76,54	76,58	77,03	75,93
Ortalama		73,69	74,96	74,74	**75,36**	74,51
$F\ ratio$	13,0158					
$F\ critic$	3,83785335					
NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+PROS	1000	50,38	48,81	49,17	49,19	50,74
LEX+PROS	3000	49,33	47,14	46,69	47,38	48,45
LEX+PROS	6000	45,00	43,57	42,91	42,74	44,14
Ortalama		48,24	46,51	**46,26**	46,44	47,78
$F\ ratio$	16,078					
$F\ critic$	3,83785335					

Tablo 6.31. Bürünsel ve biçimbilgisel modellerin kullanıldığı deneyler için elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
PROS+MORP	1000	81,52	82,26	82,27	82,74	82,14
PROS+MORP	3000	82,31	82,84	83,15	83,78	82,99
PROS+MORP	6000	83,25	83,53	83,09	83,44	83,22
Ortalama		82,36	82,88	82,84	**83,32**	82,78
$F\ ratio$	4,4566					
$F\ critic$	3,83785335					
NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
PROS+MORP	1000	35,19	34,48	34,60	33,45	34,76
PROS+MORP	3000	33,43	33,12	32,52	31,12	32,67
PROS+MORP	6000	32,21	31,74	32,33	31,41	32,60
Ortalama		33,61	33,11	33,15	**31,99**	33,34
$F\ ratio$	8,0815					
$F\ critic$	3,83785335					

Tablo 6.32. Tüm stratejiler için F-measure (%) değerine göre elde edilen $F\ ratio$ ve $F\ critic$

F-measure (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP (LEX)	1000	26,31	35,41	37,17	34,85	43,81
	3000	42,90	45,78	48,95	47,29	56,12
	6000	54,41	55,09	56,83	56,83	62,45
LEX+MORP (MORP)	1000	75,22	75,52	75,69	75,10	75,47
	3000	75,51	75,75	75,85	75,64	75,62
	6000	76,30	76,76	76,70	76,92	76,69
LEX+MORP (LEX+MORP)	1000	75,37	75,40	75,58	75,59	75,47
	3000	76,41	76,38	76,39	76,13	76,53
	6000	77,14	77,27	77,59	77,10	77,37
LEX+PROS (LEX)	1000	26,31	35,41	34,83	34,41	54,60
	3000	42,90	45,78	46,52	47,12	57,92
	6000	54,41	55,09	56,86	55,63	64,08
LEX+PROS (PROS)	1000	72,19	72,85	72,61	72,96	72,16
	3000	71,75	72,63	72,70	73,24	72,72
	6000	70,31	71,68	71,84	72,20	71,54
LEX+PROS (LEX+PROS)	1000	73,20	73,94	73,54	74,44	73,38
	3000	72,65	74,40	74,10	74,62	74,23
	6000	75,21	76,54	76,58	77,03	75,93
PROS+MORP (PROS)	1000	72,19	72,85	73,32	72,51	71,67
	3000	71,75	72,63	72,90	73,33	72,85
	6000	70,31	71,68	71,86	72,59	72,18
PROS+MORP (MORP)	1000	75,22	75,52	75,74	75,90	75,59
	3000	75,51	75,75	76,05	76,18	76,45
	6000	76,30	76,76	77,39	77,26	77,55
PROS+MORP (PROS+MORP)	1000	81,52	82,26	82,27	82,74	82,14
	3000	82,31	82,84	83,15	83,78	82,99
	6000	83,25	83,53	83,09	83,44	83,22
Ortalama		67,6615	69,0926	69,4852	69,4381	**71,5085**
$F\ ratio$	7,1501					
$F\ critic$	4,459					

Tablo 6.33. Tüm stratejiler için NIST (%) değerine göre elde edilen $F\ ratio$ ve $F\ critic$

NIST (%)						
Özellik Grubu Kodu	L verisi (Kelime)	Baseline	Kendi Kendine Eğitme	Uzlaşma	Self-Combined	Uzlaşmama
LEX+MORP (LEX)	1000	91,38	96,62	82,86	94,07	81,21
	3000	92,14	89,88	74,55	76,50	68,26
	6000	73,02	74,09	68,40	68,38	61,52
LEX+MORP (MORP)	1000	49,81	48,69	48,76	49,60	49,10
	3000	48,74	49,05	49,19	49,71	49,67
	6000	47,43	47,05	47,14	46,86	47,62
LEX+MORP (LEX+MORP)	1000	48,86	49,07	49,09	48,95	48,86
	3000	48,55	47,98	47,95	49,14	47,26
	6000	45,48	46,88	45,07	47,40	46,07
LEX+PROS (LEX)	1000	91,38	96,62	93,57	96,29	75,52
	3000	92,14	89,88	78,12	77,36	70,98
	6000	73,02	74,09	70,36	73,31	62,86
LEX+PROS (PROS)	1000	52,00	50,33	51,28	50,90	50,78
	3000	55,00	52,05	51,55	49,88	51,29
	6000	55,79	52,02	51,50	51,95	52,88
LEX+PROS (LEX+PROS)	1000	50,38	48,81	49,17	49,19	50,74
	3000	49,33	47,14	46,69	47,38	48,45
	6000	45,00	43,57	42,91	42,74	44,14
PROS+MORP (PROS)	1000	52,00	50,33	49,55	50,83	52,41
	3000	55,00	52,05	50,91	48,71	51,41
	6000	55,79	52,02	51,95	52,43	52,48
PROS+MORP (MORP)	1000	49,81	48,69	48,79	48,34	48,81
	3000	48,74	49,05	47,90	47,52	47,07
	6000	47,43	47,05	44,31	45,69	45,02
PROS+MORP (PROS+MORP)	1000	35,19	34,48	34,60	33,45	34,76
	3000	33,43	33,12	32,52	31,12	32,67
	6000	32,21	31,74	32,33	31,41	32,60
Ortalama		56,2611	55,6426	53,3711	54,0411	**52,0163**
$F\ ratio$	6,8654					
$F\ critic$	4,459					

7. TARTIŞMA

Genellikle, Otomatik Konuşma Tanıma (ASR) Sisteminin çıkışından elde edilen metin; başlıklar, paragraflar, cümleye ilişkin noktalama işaretleri, büyük küçük harf ayrımı gibi özelliklerden yoksun olarak elde edilmektedir. Cümle bölütleme işlevi standart konuşma tanıyıcılarının çıkışından elde edilen işlenmemiş kelime dizisi biçimindeki veriyi zenginleştirmeyi amaçlayan bir işlemdir. Bu işlemin rolü, kelime dizisi biçiminde olan verinin cümle ünitelerine ayrılmasını sağlamaktır. Cümle bölütleme konuşma anlamaya kadar olan süreçte ilk adımdır. Cümle bölütleme işlemi, çözümleme, makine çevirimi, bilgi çıkarımı gibi cümle bölütlemenin yapıldığının varsayıldığı konuşma işlemenin daha ileri uygulamaları için bir ön adım olarak gerçekleştirilmektedir. Cümle sınırları belirlendikten sonra bu cümleler üzerinde daha ileri düzeydeki sözdizimsel ve/veya anlamsal analizler gerçekleştirilebilmektedir.

Doğal konuşmada fazladan varolan ve sözcüksel ve biçimbilgisel olmayan; perde, enerji, duraklar ve kelime aralarındaki süreler gibi bürünsel özellikler adı verilen özellikler bulunmaktadır. Konuşmanın cümlelere bölünmesinde, sözcüksel, biçimbilgisel ve bürünsel ipuçlarının birbirini tamamlayan bilgiler taşımaktadır. Daha önceki çalışmalarımızda bunların birleştirildiği ve İngilizce konuşma dili için cümle bölütlemenin performansının arttırıldığı metodlar öne sürülmüştür. İngilizce (ve benzeri diller) de dil anlama işlevleri (örneğin cümle ve konu bölütlemesi ve duygu sezimi gibi) için bürünsel, sözcüksel ve biçimbilgisel bilginin kullanılması ile pekçok yararlı sonuçlar elde edilmiştir. Bununla birlikte, Türkçe, Macarca gibi İngilizceden tamamen farklı bir davranış gösteren diller yeterince incelenmemişlerdir.

Bu projede, Türkçe konuşma üzerinde cümle bölütleme için bürünsel, sözcüksel ve biçimbilgisel özelliklerin çıkarılması ve kullanılması araştırılmıştır. Projenin diğer bir sonucu ise biçimbilgisel açıklaştırma ve kelime anlam açıklaştırma gibi diğer amaçlar için kullanılabilecek kelime ve anlambirim düzeyinde bürünsel bir veritabanının elde edilmesidir.

İstatistiksel yöntemler cümle bölütlemede geniş bir kullanım alanına sahip olmakla birlikte, pahalı, zaman alıcı ve hazırlanması zahmetli olan oldukça fazla miktarlarda etiketlenmiş veriye gereksinim duymaları bir dezavantajdır. Daha önceki çalışmalarımızda cümle bölütleme için çok az miktarda etiketlenmiş alan içi verinin çok miktarda alan dışı veriyi etiketlemek için kullandığı öğreticili model uyarlama yöntemleri geliştirilmişti. Bu proje, herhangi bir alan dışı veri kullanmaksızın, cümle bölütleme modellerinin co-training ile yarı

öğreticili olarak eğitilmesi ve geleneksel yarı öğreticili kendi kendine eğitme yaklaşımları ile karşılaştırılması üzerine odaklanmıştır.

Bu projede konuşma verisine ilişkin bürünsel, sözcüksel ve biçimbilgisel bilgilerin bakış olarak kullanıldığı co-training ile cümle bölütlemenin gerçekleştirilmesi ele alınmıştır. Konuşma özellikleri (bürünsel, sözcüksel ve biçimbilgisel) ayrışık ve doğal özellik seti veya bakış olarak ele alınmış ve bu özellik setlerinin co-training algoritması ile kullanılması ile baseline sistemin performansının arttırılmasına çalışılmıştır. Amaç, bürünsel, sözcüksel ve biçimbilgisel özelliklerinin çıkarıldığı çok az miktarda etiketlenmiş veri ile başlanarak büyük miktardaki etiketlenmemiş veriden etiketlenmiş veri miktarını arttırmaya çalışmaktır.

Ayrıca, co-training için uzlaşma ve uzlaşmama adı verilen farklı öğrenme stratejileri de araştırılmıştır. Buna ek olarak, self-combined adını verdiğimiz ve kendi kendine eğitme ile co-training yaklaşımlarının bir araya getirildiği bir yaklaşım da öne sürülmüştür.

Deneysel sonuçlar analiz edildiğinde, co-training stratejilerinin tümünün baseline ve kendi kendine eğitme performanslarından daha iyi olduğu görülmektedir. Örneğin sadece 1K elle etiketlenmiş örnek mevcut olduğunda ortalama cümle bölütleme performansının en yüksek düzeyde co-training uzlaşmama stratejisi ile gerçekleştirildiği ve performansın baseline a göre %64.17 den %69.36 ya yükseldiği, başka bir deyiş ile %8.09 oranında bir geliştirim sağlandığı görülmüştür. 1K, 3K ve 6K elle etiketlenmiş veri miktarları için elde edilen sonuçların ortalaması incelendiğinde en yüksek cümle bölütleme performansının co-training uzlaşmama stratejisi ile gerçekleştirildiği ve baseline a göre başarımın %67.66 dan %71.50 ye yükseldiği, başka bir ifade ile %5.68 oranında bir iyileştirme elde edildiği görülmüştür.

Beklendiği üzere tüm elle etiketlenmiş başlangıç veri büyüklüklerinde (1K, 3K ve 6K) etiketlenmiş veri sayısı arttıkça cümle bölütleme performansları birbirine yaklaşmaktadır. Tüm modellerde görüldüğü üzere co-training stratejileri özellikle çok az sayıda elle etiketlenmiş veri olması durumunda baseline performansını ciddi bir biçimde geliştirmektedir.

Sözcüksel modellerin daha yüksek performans göstermelerinin bir nedeni de konuşma verilerinin toplantı verileri yerine haber kaynaklı veriler olmasıdır. Zira haber kaynaklı konuşma verileri belirli konuşmacıların uzun süre tekdüze konuştuğu kayıtlardan oluşmaktadır. Oysa toplantı konuşma verileri bir grup konuşmacının spontane ve doğal konuşmaya daha yakın konuştuğu verilerdir. Bürünsel modellerin doğru bir biçimde cümle sonu olarak etiketlediği ve fakat sözcüksel modellerin cümle sonu kararını hatalı verdiği

durumlar incelenmiştir. Özellikle konuşmanın başka biri tarafından araya girilerek kesintiye uğratıldığı ve konuşmacının bir cümleye başladıktan sonra cümlesini keserek başka bir cümle ile yeniden ifade etme ihtiyacı hissettiği durumların, haber verilerinde daha az olduğu görülmektedir. Sözcüksel model bu örneklerde genellikle cümle sonlarını doğru bir biçimde tespit edemezken bürünsel model doğru bir biçimde etiketleme yapabilmektedir. Bu tür örneklerin haber konuşma verilerinde az olması, sözcüksel modellerin başarımını bürünsel modellerin başarımının üzerine çıkarmaktadır. Ayrıca bürünsel modellerin performanslarını doğrudan etkileyen bir başka önemli faktör de ses dosyalarında bulunan gürültü, müzik gibi, arkaplan gürültüleridir.

SRI trafafından geliştirilen DECIPHER konuşmacıdan bağımsız konuşma tanıma sistemi (DECIPHER Speaker Independent Speech Recognition System) nin performansının SRI ın ALGEMY yazılımı tabanlı bürünsel özelliklerin çıkarıldığı yazılım üzerinde performans artışı yarattığı gözlenmiştir. Zira DECIPHER gerek arkaplan gürültüsünü giderme gerekse konuşmacı ayrımını oldukça yüksek performansta gerçekleştirmektedir. Bu nedenle, oldukça sofistike algoritmalar kullanılarak hesaplanan SRI ın Algemy yazılımı tabanlı bürünsel özellikler ile elde edilen cümle bölütleme performansının tamamen açık kaynak kodlu PRAAT yazılımı ile gerçekleştirilen Purdue Prosodic Feature Extraction Tool tabanlı bürünsel özellikler ile elde edilen cümle bölütleme performansının max F-measure ve min NIST hata oranı ölçütleri ile değerlendirildiğinde sırası ile %4,5 ve %13 üzerinde olduğu görülmektedir. Elde edilen sonuçlar ALGEMY ile elde edilen sonuçlar ile karşılaştırıldığında tamamen açık kaynak kodlu bir yazılım ile özellikle haber verileri gibi spontane olmayan ve tekdüze konuşulan (less informative) bir veri profili için oldukça başarılıdır. Bu sonuçların ICSI-MRDA (ICSI-Meeting Recorder Dialog Act) gibi bir grup konuşmacının toplantı verilerinin bulunduğu bir veri profilinde Praat tabanlı Purdue Prosodic Feature Extraction Tool lehinde daha da geliştirileceği söylenebilir. Zira toplantı verilerinde bürünsel modellerin cümle bölütleme performansını arttıran konuşmanın başka biri tarafından araya girilerek kesintiye uğratıldığı ve konuşmacının bir cümleye başladıktan sonra cümlesini keserek başka bir cümle ile yeniden ifade etme ihtiyacı hissettiği durumlara çokça rastlanmaktadır. Oysa haber verilerinde uzun süre belirli bir konuşmacı hakimiyeti sözkonusu olmaktadır. Dolayısı ile Purdue Prosodic Feature Extraction Tool ile elde edilen bürünsel özelliklerin kullanıldığı bürünsel modellerin, arkaplan gürültüsü, müzik-konuşma örtüşmeleri gibi bürünsel özelliklerin çıkarılmasını zorlaştıran durumlardan etkilenmemesi durumunda cümle bölütleme performansının daha da artacağı sonucu çıkarılabilmektedir.

8. SONUÇ

Özellikle bilgi, bilgiye erişim, bilginin kullanılması ve bilgi iletişimine yönelik uygulamalar son yıllarda büyük gelişimler kaydetmiştir. Bu gelişime bağlı olarak, insanlar ve bilgi teknolojilerinin kullanılmasını sağlayan veya kolaylaştıran sistemlerin ve bilgisayarların etkileşimi konuları da önem kazanmaktadır. İnsanların birbirleri ile iletişiminin yada etkileşiminin en doğal biçimi olan konuşmanın yada dilin kullanılması, bu yöntemin bilgisayar ve diğer sistemler ile de kullanılıp kullanılamayacağının sorgulanmasına neden olmuş ve bu alanda araştırmalar yapılmıştır. Özellikle doğal dilde iletişimi sağlayacak olan doğal dil işleme, insanlar ile bilgisayarlar arasında varolan etkileşim biçimini bütünüyle değiştirmeye aday teknolojilerden biridir. Doğal dil işleme, özellikle doğal dilde çözümleme, yorumlama ve üretim yapabilen sistemlerin geliştirilmesine yönelik çalışmaları kapsar. Bu teknoloji, dile ilişkin pekçok özelliği bilgisayar teknolojisi ile birleştirmekte ve insanlar ile makineler arasında doğal dilde iletişim sağlayarak iletişimi kolaylaştırmak, dile ilişkin birtakım özellikleri ortaya çıkarmak ve dilden dile makine ile çeviri yapmak gibi işlevleri beraberinde getirmektedir. Doğal dilde yapılan çalışmalar ve uygulamalar tüm dünyada herkes tarafından kullanılan bir dil olması nedeni ile İngilizce üzerinde yoğunlaştırılmıştır. Bu konuda özellikle Türkçe üzerinde yapılan çalışmalar ve geliştirilen yöntemler mevcut olmakla birlikte bu çalışmaların arttırılması, uygulamalarının çeşitlendirilmesi ve kullanımlarının yaygınlaştırılması gerekmektedir. Bu proje ile bu konudaki çalışmalara katkıda bulunulması amaçlanmaktadır.

Bu projenin kapsamında yeralan Türkçe konuşma için bürünsel, sözcüksel ve biçimbilgisel özellik setlerinin çıkarılması ve bu özellik setlerinin diğer ileri dil işleme uygulamalarına temel teşkil eden cümle bölütleme gibi bir alanda co-training gibi etkin bir yarı öğreticili öğrenme algoritması ile yeni öğrenme stratejileri kullanılarak gerçeklenmiş olması, yukarıda önemi belirtilen araştırmalara getireceği katkı nedeni ile oldukça önem taşımaktadır. Bu projede gerçekleştirilen bu yaklaşım ile elde edilecek sistem, gerek yöntemsel açıdan gerekse Türkçe için geliştirilmiş olması nedeni ile bu alanda yapılan ve yapılacak olan ileri araştırmalar açısından önem taşımaktadır. İleri dil işleme uygulamalarına temel oluşturacak bu sistem ile diğer olası geliştirilecek sistemler entegre edilebilecek ve konuşma işaretleri kullanılarak örneğin çok büyük miktarlarda veriler içeren veritabanları sorgulanabilecek, konularına göre sınıflandırılabilecek, istenilen konulara ilişkin paragraflar işaretlenebilecek, makine ile dilden dile çeviri yapılırken bir cümlenin nerede başladığı ve nerede bittiği gibi temel işlevler kolayca gerçekleştirilebilecektir. Yukarıda belirtilen uygulamaların gerek çok fazla insan gücü ve istihdamı gerektirmesi, gerekse çok fazla zaman ve emeğe gereksinim duyması nedeni ile,

geliştirilen yöntemin uygulanması ile sosyal ve ekonomik kayıpların önlenmesi mümkün olabilecektir.

Özellikle medya ve iletişim sektöründe, kütüphanelerde, veri merkezlerinde, bankalarda, veriye uzaktan erişim yapıldığı durumlarda ve verinin yoğun olarak kullanıldığı yerlerde verilerin belirli özelliklere göre toplanması ve sınıflandırılması, verilerin hızlı bir biçimde taranması ve erişimi oldukça önemlidir. Bu bağlamda, geliştirilmiş olan bu proje ile belirtilen uygulamalar için geliştirilen ileri uygulamalarda kullanılmak üzere bir altyapı yada temel oluşturulmuş olunmaktadır.

Bu projede geliştirilmiş olan sistem ile elde edilen sonuçlar aşağıda özetlenmektedir;

Bu projede genel olarak, Türkiye Türkçesinin bürünsel/ezgisel, sözcüksel ve biçimbilgisel özelliklerinin çıkarılması ve bu özelliklerin en etkin yarı öğreticili algoritmalardan biri olan Co-training ile cümle bölütlemesinde kullanılması amaçlanmıştır. Böylece Türkçe konuşma diline ilişkin yüksek performanslı bir cümle bölütleme sisteminin oluşturulması hedeflenmiştir.

Elde edilen sistem çok az etiketlenmiş veri ile yüksek miktarlarda etiketlenmemiş veriyi mümkün olabilecek en yüksek güvenilirlikte etiketleyerek zaman alıcı ve emek yoğun bir işlevi yerine getirdiği gibi, cümle bölütlemenin büyük bir doğrulukla yapılmış olması ile de daha ileri araştırma ve uygulamaların (konu bölütleme, özetleme, bilginin geri kazanımı vb.) başarımını da arttıracaktır.

Bu projede uygulanan yöntemler, Türkçe konuşma verilerine ilişkin bürünsel, sözcüksel ve biçimbilgisel özelliklerin ortaya çıkarılmasını ve cümle bölütleme uygulamalarında kullanılmasını sağlamaya dönüktür. Bununla birlikte çıkarılan sözcüksel ve biçimbilgisel özelliklerin elde edilen bürünsel özellikler ile ilişkisi de incelenmiştir.

Bu projenin en önemli kısmını ise elde edilen bürünsel, sözcüksel ve biçimbilgisel özelliklerin yarı öğreticili bir algoritma olan co-training yöntemi ile cümle bölütleme üzerinde kullanılması oluşturmaktadır. Bu çalışma literatürde gerek Türkçe konuşma diline ilişkin özellik setlerinin co-training algoritması ile birlikte kullanılmış olması açısından gerekse co-training yaklaşımının Türkçe cümle bölütleme alanında uygulanmış olması açısından ilk yapılan çalışma olma niteliğindedir.

Bu projede cümle bölütleme için kullanılan Co-training yönteminde ilk defa literatüre İngilizce dili için yaptığımız çalışmalarla kazandırdığımız farklı stratejiler kullanılmıştır. Bu stratejiler kendi kendine eğitme dışındaki uzlaşma, uzlaşmama ve self-combined stratejileridir. Ayrıca elde ettiğimiz sonuçlar ile İngilizce dili için en iyi performansı veren strateji-özellik set(ler)i ile Türkçe dili için elde edilen en iyi strateji-özellik set(ler)i bulma olanağı ile birlikte bu iki dile ilişkin elde edilecek sonuçlar ile yeni analizlerin yapılması da mümkün olabilecektir.

Bu projede önerilen yöntemler ile Türkçe konuşma verilerinin pekçok yönden analiz edilmesi ve önemli bulgulara ulaşılması sağlanmaktadır.

Literatürde özellikle cümle ve konu bölütleme alanında yapılan araştırmalarda çoğunlukla dile ilişkin sözcüksel bilginin kullanılmasına yönelik yöntemlerin geliştirildiği görülmektedir. Bu projede ise cümle bölütleme için sözcüksel bilgi ile birlikte bürünsel ve biçimbilgisel bilgilerin de çıkarılması ve kullanılması sağlanmaktadır.

Özellikle Türkçe konuşma verilerine ilişkin özellik setlerinin çıkarılması ve bu özelliklerin cümle bölütleme gibi dil işleme alanında pek çok uygulamanın ilk adımına uygulanmış olması projenin diğer bir özgün yönünü oluşturmaktadır.

Ayrıca belirtilen Türkçe konuşma veya audio işaretlerin oldukça başarılı bir dilden bağımsız otomatik konuşma tanıyıcısından (SRI Decipher) geçirilmiş olması önerilen yöntemin performansını ve güvenilirliğini arttıran bir özelliktir.

Bu işlemler SRI Decipher konuşma tanıyıcısı üzerinde gerçekleştirildiği gibi bağımsız açık kaynak kodlu yazılımlar (Hidden Markov Toolkit (HTK)) üzerinde de gerçekleştirildiğinden gerek telif hakkı, gerekse daha sonradan yapılacak çalışmalarda oluşabilecek yazılım bağımlılığı sözkonusu olmamaktadır. Ayrıca belirtilen açık kaynak kodlu yazılımların kullanılması, proje araştırmacıları ve başka araştırmacılar tarafından daha sonradan sistem üzerinde yapılabilecek geliştirimler ve farklı ileri uygulamalar (konu bölütleme, özetleme vb.) için de uygun ve ortak kullanıma açık bir taban oluşturmaktadır.

Bu projede, Türkçenin yapısından kaynaklanan nedenlerden dolayı büyük miktarlarda özellik setleri ile çalışıldığından karar ağaç yapıları yerine boosting sınıflandırıcılar tercih edilmiştir.

Türkçe için gerçekleştirilen bu proje ile, elde edilecek özellik veritabanları ve elde edilen cümle bölütleme yaklaşımlarının daha sonra tarafımızdan ve diğer araştırmacılar tarafından yapılacak daha ileri dil işleme uygulamaları için bir temel teşkil etmesi amaçlanmaktadır.

9. PROJE ÇIKTILARI

Aşağıda 111E228 Numaralı Proje kapsamında üretilen yayınlar ve tezler verilmektedir. Proje çıktısı olarak süreci tamamlanmış yayın ve tezler ARDEB Proje Takip Sistemi, 111E228 Nolu proje, Bilimsel Raporu altında Rapor Dönemi Çıktıları bölümüne yüklenmiştir.

Sıra	Çıktı türü	Yazarlar	Başlık	Yayın yeri	Durumu
1	Konferans Makalesi	Dogan Dalva, İzel D. Revidi, Umit Guz, Hakan Gurkan	Extraction and Comparison of Various Prosodic Feature Sets on Sentence Segmentation Task for Turkish Broadcast News Data	IEEE JCSSE 2014	Kabul edildi, sunuldu ve bildiri kitapçığında yayınlandı.
2	Konferans Makalesi	Dogan Dalva, İzel D. Revidi, Umit Guz, Hakan Gurkan	Türkçe Haber Yayını Verileri için Bürünsel Bilginin Çıkarılması ve Cümle Bölütlemede Kullanılması	IEEE SIU 2014	Kabul edildi, sunuldu ve bildiri kitapçığında yayınlandı.
3	Yüksek Lisans Tezi	İzel D. Revidi Tez danışmanı : Doç.Dr. Ümit Güz, 2. Danışman: Doç.Dr. Hakan Gürkan	Prosodic, Morphological and Lexical Feature Extraction of Turkish Broadcast News Data	Işık Üniversitesi, Fen Bilimleri Enstitüsü	Tez tamamlandı, Kabul edildi. (Haziran 2014)
4	Doktora Tezi	Doğan Dalva Tez danışmanı : Doç.Dr. Ümit Güz, 2. Danışman: Doç.Dr. Hakan Gürkan	Co-training using Prosodic, Lexical and Morphological Information for Automatic Sentence Segmentation of Turkish Spoken Language	Işık Üniversitesi, Fen Bilimleri Enstitüsü	Doktora tezi devam etmektedir. Eylül 2015 de tamamlanması planlanmaktadır.)
5	SCI dergi makalesi	Dogan Dalva, İzel D. Revidi, Umit Guz, Hakan Gurkan	Co-training using Prosodic, Lexical and Morphological Information for Automatic Sentence Segmentation of Turkish Broadcast News Data	IEEE Transactions on Audio Speech and Language Processing	Gönderilmiştir. Değerlendirme aşamasındadır.

KAYNAKLAR

Abney, S. 2002. “Bootstrapping”, in Processings of the Annual Meeting of the Association for Computational Linguistics (ACL).

Ang, J., Liu, Y., Shriberg, E. 2005. “Automatic dialog act segmentation and classification in multiparty meetings”, in Proc. of ICASSP, vol. 1, pp. 1061–1064.

Bacchiani, M., Roark, B. 2003. “Unsupervised language model adaptation”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong Kong.

Beeferman, D., Berger, A., Lafferty, J. 1999. “Statistical models for text segmentation”, Machine Learning, Special Issue on Natural Language Learning.

Blum, A., Mitchell, T. 1998. “Combining labeled and unlabeled data with co-training”, in Proceedings of the Workshop on Computational Learning Theory (COLT), Madison, WI.

Cuendet, S., Hakkani-Tur, D., Tur, G. 2006. “Model adaptation for sentence segmentation from speech”, in Proc. IEEE/ACL Spoken Language Technology (SLT) Workshop, Aruba.

Dalva, D. 2012. “Automatic Speech Recognition System for Turkish Spoken Language”, MSc. Thesis, Graduate School of Science and Engineering, Isik University, (Supervisor: Associate Prof. Umit Guz, Co-supervisor: Associate Prof. Hakan Gurkan).

Dalva, D., Revidi, I.D., Guz, U., Gurkan, H. 2014. “Extracting the Prosodic Information for Turkish Broadcast News Data and Using on the Sentence Segmentation Task”, IEEE SIU 2014, 22nd Signal Processing and Communications Applications Conference, Karadeniz Technical University, Trabzon, Turkey.

Dalva, D., Revidi, I.D., Guz, U., Gurkan, H. 2014. “Extraction and comparison of various prosodic feature sets on sentence segmentation task for Turkish Broadcast News data”, in Computer Science and Software Engineering JCSSE, vol. 11, pp. 70−73.

Favre, B., Hakkani-Tur, D., Cuendet, S. “Icsiboost”. http://code.google.come/p/icsiboost, Son erişim tarihi: 2007.

Ferrer, L. 2002. “Prosodic Features for the Switchboard Database”, Speech Technology and Research Lab., SRI International, Merlo Park, CA 94025.

Freund, Y., Schapire, R.E. 1996. “Experiments with a new boosting algorithm”, ICML.

Fung, J.G. 2011. “Automatic Design of Prosodic Features for Sentence Segmentation”, Technical Report No. UCB/EECS-2011-140, Electrical Engineering and Computer Sciences University of California at Berkeley.

Gauvain, J.L., Lee, C.H. 1994. “Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains”, IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291– 298.

Gotoh, Y., Renals, S. 2000. “Sentence boundary detection in broadcast speech transcripts”, in Proc. of ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium, ASR-2000, pp. 228–235.

Gretter, R., Riccardi, G. 2001. “On-line learning of language models with word error probability distributions”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, Utah.

Guz, U., Cuendet, S., Hakkani-Tür, D., Tur, G. 2007. “Co-training Using Prosodic and Lexical Information for Sentence Segmentation”, Interspeech, Antwerp, BELGIUM.

Guz, U., Favre, B., Hakkani-Tur, D., Tur, G. 2009. “Generative and discriminative methods using morphological information for sentence segmentation of turkish”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, pp. 895−903.

Guz, U., Cuendet, S., Hakkani-Tur, D., Tur, G. 2010. “Multiview semisupervised learning for dialog act segmentation of speech”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, pp. 320−329.

Hakkani-Tür, D., Tur, G., Stolcke, A., Shriberg, E. 1999. “Combining Words and Prosody for Information Extraction from Speech”, in Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary.

Hakkani-Tür, D., Tur, G., Rahim, M., Riccardi, G. 2004. “Unsupervised and active learning in automatic speech recognition for call classification”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Canada.

Hearst, M.A. 1997. “Texttiling: Segmenting text into multiparagraph subtopic passages”, Computational Linguistics, vol. 23, pp. 33-64.

Hirschberg, J., Grosz, B. J. 1992. “Intonational features of local and global discourse”, in Proceedings of the Workshop on Spoken Language Systems, pp. 441-446, DARPA.

Hirschberg, J., Nakatani, C. 1996. “A prosodic analysis of discourse segments in direction-giving monologues”, in Proc. ACL, pages 286-293, Santa Cruz, CA.

Huang, Z., Chen, L., Harper, M.P. 2006. “An Open Source Prosodic Feature Extraction Tool”, Proceedings of Language Resource and Evaluation Conference, Genoa, Italy.

Kiritchenko, S., Matwin, S. 2001. “Email classification with co-training”, in Centre for Advanced Studies on Collaborative Research (CASCON).

Kolar, J., Svec, J., Psutka, J. 2004. “Automatic punctuation annotation in Czech broadcast news speech”, in Proc. of 9th Conference Speech and Computer, pp. 319–325.

Kozima, H. 1993. “Text segmentation based on similarity between words”, in 31st Annual Meeting of the ACL, 286-288.

Levi, S. V. 2001. “Glides, laterals, and Turkish vowel harmony”, Proceedings from the 37th meeting of the Chicago Linguistics Society, 379-394.

Levi, S. V. 2002. “Limitations on tonal crowding in Turkish intonation”, Phonologica: 9th international phonology conference, Vienna.
Levi, S. V. 2005. “Acoustic correlates of lexical accent in Turkish”, Journal of the International Phonetic Association, 35, 73-97.

Liu D., Zong, C. 2003. “Utterance segmentation using combined approach based on bi-directional n-gram and maximum entropy”, in Proc. of ACL-2003 Workshop: The Second SIGHAN Workshop on Chinese Language Processing, pp. 16–23.

Liu, Y., Shriberg, E., Stolcke, A., Peskin, B., Ang, J., Hillard, D., Ostendorf, M., Tomalin, M., Woodland, P., Harper, M. 2005. “Structural metadata research in the EARS program”, in Proc. of ICASSP, vol. 5, pp. 957–960.

McClosky, D., Charniak, E., Johnson, M. 2006. “Effective self-training for parsing”, in Proceedings of the Human Language Technology Conference (HLT)-Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), New York, NY.

Mihalcea, R. 2004. “Co-training and self-training for word sense disambiguation”, in Proceedings of the Conference on Computational Natural Language Learning (CoNLL), Boston, MA.

Mitchell, T.M. 1999. “The role of unlabeled data in supervised learning”, in Proceedings of the Sixth International Colloquium on Cognitive Science, San Sebastian, Spain.

Nigam, K., Ghani, R. 2000. “Understanding the behaviour of co-training”, in Proceedings of the Workshop on Text Mining at the Sixth ACM SIGKDD at the KDD.

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

Oflazer, K., Gocmen, E., Bozsahin, C. 1995. “An Outline of Turkish Morphology”, Technical Report, Middle East Technical University.

Oskay, B., Salor, Ö., Özkan, Ö., Demirekler, M., Çiloğlu, T. 2001. “Determination of Prosody from Turkish Text and Its Application to Speech Synthesis”, SIU 2001, Northern Cyprus.

Passonneau, R.J., Litman, D.J. 1997. “Discourse Segmentation by Human and Automated Means”, Computational Linguistics.

Revidi, I.D. 2014. “Prosodic, Morphological and Lexical Feature Extraction of Turkish Broadcast News Data”, MSc. Thesis, Graduate School of Science and Engineering, Isik University, (Supervisor: Associate Prof. Umit Guz, Co-supervisor: Associate Prof. Hakan Gurkan).

Roark, B., Liu, Y., Harper, M., Stewart, R., Lease, M., Snover, M., Shafran, I., Dorr, B., Hale, J., Krasnyanskaya, A., Yung, L. 2006. “Reranking for sentence boundary detection in conversational speech”, in Proc. of ICASSP, vol. 1, pp. 545–548.

Sayli, O., Arslan, L. M. 2003. “Türkçedeki Seslerin Süre Özellikleri”, DilBilim Araştırmaları, Boğaziçi Üniversitesi Yayınevi, İstanbul.

Schapire, R.E. 1990. “The strength of weak learnability”, Machine Learning, Kluwer Academic Publishers, Boston, vol. 5, pp. 197−227.

Schapire, R.E. 2001. “The boosting approach to machine learning an overview”, ATT Labs Research, Shannon Laboratory.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tur, G. 2000. “Prosody based automatic segmentation of speech into sentences and topics”, Speech Comm., 32(1-2), pp. 127-154.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tur, G. 2000. “Prosody-based automatic segmentation of speech into sentences and topics”, Speech Communication, vol. 32, no. 1-2, pp. 127–154.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tur, G., Rivlin, Z., Sönmez, K. 1999. “Combining Words and Speech Prosody for Automatic Topic Segmentation”, in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, Herndon, VA.

Tur, G. 2000. “A statistical information extraction system for Turkish”, Ph.D. dissertation, Dept. of Comput. Sci., Bilkent Univ., Ankara, Turkey.

Tur, G., Hakkani-Tür, D., Stolcke, A., Shriberg, E. 2001. “Integrating prosodic and lexical cues for automatic topic segmentation”, Computational Linguistics, vol. 27, pp. 31-57.

Tur, G., Guz, U., Hakkani-Tür, D. 2006. “Model Adaptation for Dialog Act Tagging”, IEEE/ACL 2006, Workshop on Spoken Language Technology, Aruba, pp.94-97.

Wang, W., Huang, Z., Harper, M. 2007. “Semi-supervised leaning for partof- speech tagging of mandarin transcribed speech”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI.

Zimmermann, M., Liu, Y., Shriberg, E., Stolcke, A. 2005. “A* based segmentation and classification of dialog acts in multiparty meetings”, in Proc. of ASRU, pp. 215–219.

Zong C., Ren, F. 2003. “Chinese utterance segmentation in spoken language translation”, in The 4th International Conference on Computational Linguistics and Intelligent Text Processing, pp. 516–525.

EKLER

Bu kısımda Kendi kendine eğitme ve Co-Training algoritmalarının, optimum model eğitimi sırasında güncel modellerin etiketlenmemiş (U) veriden, etiketlenmiş (L) veriye azar azar etiketleyerek aktardığı örnekler bulunmaktadır. Aşağıda wav uzantılı olarak isimleri yeralan ses dosyaları ARDEB Proje Takip Sistemi, 111E228 Nolu proje, Bilimsel Raporu altında Çoklu Ortam Dosyaları bölümüne yüklenmiştir.

Kendi Kendine Eğitme Algoritması, Sözcüksel Özellikler,

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfTrLexTN1.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.021476

...iktidar partisi milletvekili **savaş *ve {WB}* önemli** derecede güvenlik durumlarında...

Örnek 2:	SelfTrLexTN2.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.018460

...verilmesini, Birleşmiş Milletler **ve *Amerikanın {WB}* da** yaptırımları kaldırmasını...

Örnek 3:	SelfTrLexTN3.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.018460

...istiyor. Ancak eski Baas yetkilisi **Iraklı *Kürt {WB}* ve** Şiilere karşı işlediği...

Örnek 4:	SelfTrLexTN4.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.014295

...hedef oluyor. Seken bir kurşunun on **üç *yaşındaki {WB}* bir** çocuğun ölümüne...

Kendi Kendine Eğitme Algoritması, Sözcüksel Özellikler,

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfTrLexTP1.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.023317

...gibi öncelikli konulara girmediğini **öne *sürdü. {SB}* Buna** karşın Alman Sanayiciler Birliği...

Örnek 2:	SelfTrLexTP2.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.021892

...yüzde kırk bir ile Belçika olarak **öne *çıkıyor. {SB}*** **Burası** Brüksel, ben...

Örnek 3:	SelfTrLexTP3.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.018224

...Başkanı Mahmut Ahmedinejatla da bir **araya *geldi. {SB}*** **Pakistan** ayrıca Hindistan üzerinden...

Örnek 4:	SelfTrLexTP4.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.017967

...Amerika'nın sesi sabah yayını **devam *ediyor. {SB}*** **Burası** Washington. İngiltere Başbakanı...

Kendi Kendine Eğitme Algoritması, Sözcüksel Özellikler,

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfTrLexFN1.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.016184

...vurgulandı. Değer Akal Amerika'nın Sesi **radyosu *Ankara. {SB}*** **Türkiye** bu haftaya...

Örnek 2:	SelfTrLexFN2.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.015367

... Bu dosyalardan dokuz bin on altısı **Türkiyeyle *ilgili. {SB}*** **Türkiye** bu oranla hakkında...

Örnek 3:	SelfTrLexFN3.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.010970

...daha ucuza çıkması ve finansman da **bulması *mümkün. {SB}*** **İki** yüz seksen kilometrelik bu hattı...

Örnek 4:	SelfTrLexFN4.wav
Özellik:	Sözcüksel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.010785

...otuzu gösteriyor. Günaydın ben **Özge *Övün. {SB}* Önümüzdeki** otuz dakikada dünyanın çeşitli...

Açıklamalar: Birinci örnekte “radyosu-Ankara” bi-gramı, konuşmacının haber sunumunu bitirirken söylediği kelimeleri içermektedir. “Ankara” mono-gramı, “Ankara-Türkiye” bi-gramı ve “radyosu-Ankara-Türkiye” tri-gramı cümle sonlarında çok az rastlandığı için model düşük skorla cümle sonu olmadığına karar vermiştir. İkinci ve üçüncü örneklerde, ilgili örneklere ait mono-gram, bi-gram ve tri-gramlar, L verisinde genel olarak kelime sınırı kararına ait örneklerde bulunmaktadır. Dördüncü örnekte ise “isim-soyisim” bi-gramı L verisi içinde hem cümle sınırı hem de kelime sınırı durumlarında rastlanabilen bir durumdur. Bu duruma kelime sınırlarında daha çok rastlandığı için, model kelime sınırı kararı vermiştir.

Kendi Kendine Eğitme Algoritması, Sözcüksel Özellikler,
Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	SelfTrLexFP1.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.017967

... Amerika'nın Sesi yayını **devam *ediyor {WB}* burası** Washington **{SB}**. Sırp yetkililer...

Örnek 2:	SelfTrLexFP2.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.010759

...Amerika'nın Sesi radyosu sabah **yayınını *dinliyorsunuz {WB}* burası** Washington **{SB}.** Almanya'da...

Örnek 3:	SelfTrLexFP3.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.010136

...ve iş yerine baskın **yapıldığını *açıkladı {WB}* ancak** operasyon hakkında ayrıntılı bilgi vermedi **{SB}**...,

Örnek 4:	SelfTrLexFP4.wav
Özellik:	Sözcüksel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.009888

...konusunda yeni fikirlere açık **olduğunu *belirtti {WB}* ancak** nihai kararı başkan...

Açıklamalar: Yukarıdaki dört örnekte de mono-gramlar yüklemdir. Türkçe dilinde yüklem cümle sonunda yer aldığı için, bu örneklere ilişkin mono-gramlar, L verisinde cümle sonu sınıfına ait örneklerde daha çok bulunmaktadır. Birinci ve ikinci örnekte konuşmacılar yüklemin ardından bulundukları şehri belirttiği için, konuşmacıların devrik cümle kurmuş olduklarını görebiliyoruz. Üçüncü ve dördüncü cümlede ise yüklemleri "ancak" bağlacı (aynı zamanda "nextword" monogramı) takip etmektedir. "yüklem-bağlaç" bi-gramları Sözcüksel model için kelime sınırına ait bir ipucudur ancak "current word" monogramının yüklem oluşu ve "previous-current" bi-gramlarının "yapıldığını açıkladı" ve "olduğunu belirtti" gibi cümle sonlarında sıklıkla rastlanan bi-gramlar oluşu nedeniyle Sözcüksel model bu örneklerde cümle sınırı kararını vermiştir.

Kendi Kendine Eğitme Algoritması, Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfTrMorpTN1.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.113135
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	0
lastPOS:	Noun

...Türkmen halkının istek **ve *iradesinin {WB}* şekillendireceğini** söyledi. Emekli diplomat ve...

Örnek 2:	SelfTrMorpTN2.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.113135
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	0
lastPOS:	Noun

...konuşma yaptı. Kendisine yapılan davetin **Patrikhane'nin *önemini {WB}* ve** Ekümeniklik sıfatının...

Örnek 3:	SelfTrMorpTN3.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.112961
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	0
lastPOS:	?

...her kesime temsil hakkı **verilmesinin *Türkmenistan'a {WB}* ve** halkına yarar sağlayacağını...

Örnek 4:	SelfTrMorpTN4.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.112714
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	0
lastPOS:	Noun

...kalkanını burun kısmını **ve *kanatlarını {WB}* inceledi {SB}.** Yer yüzündeki görevlilerse...

Kendi Kendine Eğitme Algoritması, Biçimbilgisel Özellikler,

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfTrMorpTP1.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.078571
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

... ilgili raporlarını dikkatle **inceleyeceğini *söyledi. {SB}* Burası** Washington sabah...

Örnek 2:	SelfTrMorpTP2.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.066704
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

...dinleyicilerimize tekrar **günaydın *diyoruz. {SB}* Yayınlarımızın** yapıldığı Washingtonda...

Örnek 3:	SelfTrMorpTP3.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.075437
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

... Konferans bir çok ülke **tarafından *kınandı. {SB}* Filistinli** El Fetih ve Hamas liderleri...

Örnek 4:	SelfTrMorpTP4.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.073158
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

...stüdyolarımızın bulunduğu başkentteyse yirmi üç **otuzu *gösteriyor. {SB}* Günaydın** ben...

Kendi Kendine Eğitme Algoritması, Biçimbilgisel Özellikler,

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfTrMorpFN1.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.086552
lastMarkerA3sg:	0
lastMarkerNom:	0
lastIGhasVerb:	0
lastPOS:	Adj

...beş yaşın altında yimi milyon çocuk **aşırı *kilolu. {SB}* Şimdi** bu çocukların ileriki...

Örnek 2:	SelfTrMorpFN2.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.083453
lastMarkerA3sg:	0
lastMarkerNom:	1
lastIGhasVerb:	0
lastPOS:	Noun

... sonra Filipinler, Kenya **ve *Nijerya. {SB}* Ankete** katılanların yüzde yetmiş üç...

Örnek 3:	SelfTrMorpFN3.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.068362
lastMarkerA3sg:	0
lastMarkerNom:	1
lastIGhasVerb:	0
lastPOS:	Noun

... servetinin yüzde seksen **beşine *sahip. {SB}* Dünyanın** en varlıklı yüzde otuz...

Örnek 4:	SelfTrMorpFN4.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.058420
lastMarkerA3sg:	0
lastMarkerNom:	1
lastIGhasVerb:	0
lastPOS:	Pron

... ve aynı zamanda Müslüman **olan *birisi. {SB}* Halilzad** Amerika'ya ilk kez...

Açıklamalar: Türkçe dilinde cümleler yüklem ile bitmektedir. Bu nedenle Biçimbilgisel model için cümle sınırına ilişkin en önemli ipucu bu kelimenin yüklem olup olmamasıdır. Yukarıdaki örneklerde bu kelimeler yüklem olmadığı için "lastPOS" özelliği "Verb" değildir, "lastIGhasVerb" özelliği Logic 0 değerini almıştır. Birinci örnekte cümle sonunda bulunan kelime sıfat, ikinci ve üçüncü örneklerde isim ve dördüncü örnekte zamirdir. Bu nedenlerden dolayı model kelime sınırı kararı vermiştir.

Kendi Kendine Eğitme Algoritması, Biçimbilgisel Özellikler,

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	SelfTrMorpFP1.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.073158
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

...bulunduğu Washingtonda'ysa yirmi üç **otuzu *gösteriyor, {WB}* günaydın** ben Özge Övün. **{SB}**. Bu yayınımızda...

Örnek 2:	SelfTrMorpFP2.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.061186
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

...çıkan çatışmada en az üç **kişi *öldü {WB},* yirmi** kişi de yaralandı. **{SB}** Düzeni sağlamak için...

Örnek 3:	SelfTrMorpFP3.wav
Özellik:	Biçimbilgisel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.036015
lastMarkerA3sg:	1
lastMarkerNom:	0
lastIGhasVerb:	1
lastPOS:	Verb

...çiftlikte H5N1 virüsü **izlerine *rastlandı {WB},* yetkililer** soruşturmanın devam ettiğini bildiriyor. **{SB}**

Açıklamalar: Yukarıdaki üç durumda da cümle sonu olarak sınıflandırılmış kelimeler yüklemdir. Buna ilave olarak "CurrentLast3" özellikleri "yor" (şimdiki zaman çekimi), "l**dü**", "n**dı**" (di'li geçmiş zaman) gibi L verisinin cümle sonu sınıflarında sıkça rastlanan örneklerdir. Bu nedenlerden dolayı Biçimbilgisel model cümle sınırı kararını vermiştir.

Kendi Kendine Eğitme Algoritması, Bürünsel Özellikler,

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfProTN1.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.520799
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Stüdyo

...derginin yayıncısı **Fransadaki *iki {WB}* Müslüman** örgüt tarafından mahkemeye verilmişti. Yayıncı...

Örnek 2:	SelfProTN2.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.512810
Konuşmacı:	Cem Dalaman
Konuşma ortamı:	Stüdyo

...afrika ve Ortadoğu ülkeleriyle iş **birliğinin *arttılması {WB}* oldu. {SB}** Avrupa Birliği kaçak göçün...

Örnek 3:	SelfProTN3.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.507066
Konuşmacı:	Mevlüt Katık
Konuşma ortamı:	Telefon bağlantısı

... İşçi Partili Dostları isimli grup **iktidardaki *İşçi {WB}* Partisi** milletvekili David Lemy ile Londrada...

Örnek 4:	SelfProTN4.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	n
Boosting skoru:	0.476401
Konuşmacı:	Devrim Çubukçu
Konuşma ortamı:	Fon müziği

...Önümüzdeki otuz **dakikada** ***dünyanın {WB}*** **çeşitli** ülkelerinden derlediğimiz en son haberleri sunacağız. **{SB}**...

Kendi Kendine Eğitme Algoritması, Bürünsel Özellikler,

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfProTP1.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.398788
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Fon müziği

...kısa dalgadan ve NTV **radyodan** ***dinliyorsunuz. {SB}*** **Yayınlarımız** ile ilgili bilgi, özel haber ve...

Örnek 2:	SelfProTP2.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.390602
Konuşmacı:	Hale Ebiri
Konuşma ortamı:	Stüdyo

...en az kırk kişinin de **yaralandığını** ***açıkladı. {SB}*** **El Kaide'nin** Irak şubesi önceki gün Bağdatta...

Örnek 3:	SelfProTP3.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.381795
Konuşmacı:	Özge Övün
Konuşma ortamı:	Stüdyo

...sigara için böyle bir **zorunluluk** ***yok. {SB}*** **Sigara** şirketleri bin dokuz yüz doksan sekiz...

Örnek 4:	SelfProTP4.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	s
Boosting skoru:	0.359198
Konuşmacı:	Elif Ural
Konuşma ortamı:	Telefon bağlantısı

...bir dizi görüşmelerin yapılacağını **ifade *etti. {SB}* Filistin'de** yapılması planlanan erken...

Kendi Kendine Eğitme Algoritması, Bürünsel Özellikler,

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfProFN1.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.406820
Konuşmacı:	Arzu Çakır
Konuşma ortamı:	Telefon Bağlantısı

... bir dostluk diyaloğu **kurmak *gerekir. {SB}* Türkiye** Ermeni soykırımını tanımadan AB'ye giremez...

Açıklama: Konuşmacı bu kelimeyi cümle sonu yerine virgül konmuş gibi vurguluyor. “gerekir” kelimesinin ardından cümle sonlarına kıyasla daha kısa bir duraksama süresi bulunmaktadır.

Örnek 2:	SelfProFN2.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.391987
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Stüdyo

...tepkilere yol açtığı **için *durdurulmuştu. {SB}* Opera** ilk kez iki bin üç yılında sahnelendi...

Açıklama: “durdurulmuştu” kelimesinin ardından cümle sonlarına kıyasla daha kısa bir duraksama süresi bulunmaktadır.

Örnek 3:	SelfProFN3.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.360604
Konuşmacı:	Elif Ural
Konuşma ortamı:	Telefon bağlantısı

...ordusu adlı gruplar aynı **anda *üstlendi. {SB}* El** Fetih'in silahlı kanadı olan El Aksa Şehitleri...

Açıklama: Son hecenin tınısında yükselme bulunmaktadır. Cümle sonuna gelirken konuşmacının konuşma hızında bir azalma olmamıştır.

Örnek 4:	SelfProFN4.wav
Özellik:	Bürünsel
Orjinal etiket:	s
Hipotez edilen etiket:	n
Boosting skoru:	0.356443
Konuşmacı:	Elif Özmenek
Konuşma ortamı:	Stüdyo

... lisede değişim öğrencisi **olarak *geldi. {SB}* Daha** sonra doktorasını Amerika'nın en saygın...

Açıklama: “geldi” kelimesinden sonra kısa bir duraksama süresi var. Bir cümlenin son kelimesi telaffuz edildikten sonra yeni cümlenin ilk kelimesi telaffuz edilirken genel olarak daha yüksek bir enerji kullanılmaktadır. Bu örnekte “daha” kelimesi telaffuz edilirken konuşmacı aynı tonda ve enerjide konuşmaya devam ediyor.

Kendi Kendine Eğitme Algoritması, Bürünsel Özellikler,

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	SelfProFP1.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.321000
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Stüdyo

...Amerika dış **işleri *bakanı {WB} [duraksama]* Almanya** Başbakanı Angela Merkel ile...

Açıklama: Konuşmacı “bakanı” kelimesinden sonra konuşmacı duraksama yapıyor. Diğer kelimelere kıyasla daha uzun süreli duraksamalar cümle sınırına ilişkin önemli bir ipucu olduğundan dolayı, Bürünsel model cümle sınırı kararını vermiştir.

Örnek 2:	SelfProFP2.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.292626
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Fon müziği

...numarayı aradığınızda **Amerika'nın *sesinden {SB} [duraksama]* Eda** Dikmen'le ödemeli konuşmak...

Açıklama: Konuşmacı “sesinden” kelimesinden sonra duraksama yapıyor. Diğer kelimelere kıyasla daha uzun süreli duraksamalar cümle sınırına ilişkin önemli bir ipucu olduğundan dolayı, Bürünsel model cümle sınırı kararını vermiştir.

Örnek 3:	SelfProFP3.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.279122
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Stüdyo

...adayı olmak istediklerini açıklayan **siyasetçiler *sunlar {WB} [:]* Eski** senatör ve eski iki bin...

Açıklama: Orijinal metinde iki nokta üst üste konmuş olan kelime, cümle sonu gibi vurgulanmış. Diğer kelimelere kıyasla daha uzun süreli duraksamalar cümle sınırına ilişkin önemli bir ipucu olduğundan dolayı, Bürünsel model cümle sınırı kararını vermiştir.

Örnek 4:	SelfProFP4.wav
Özellik:	Bürünsel
Orjinal etiket:	n
Hipotez edilen etiket:	s
Boosting skoru:	0.263428
Konuşmacı:	Alparslan Esmer
Konuşma ortamı:	Stüdyo

...kutlama kararı aldı. Ölümcül H5N1 virüsünün ortaya çıktığı bölgede iki bin **beş *yüz {WB} [kio] [duraksama] [afedersiniz],* iki** bin kilometre karelik alan...

Açıklama: Konuşmacı “yüz” kelimesinden sonra düzeltme yapmak için duraksıyor. Diğer kelimelere kıyasla daha uzun süreli duraksamalar cümle sınırına ilişkin önemli bir ipucu olduğundan dolayı, Bürünsel model cümle sınırı kararını vermiştir.

Co-Training, Self-Combined Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfComLexMorpTN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	N	
Hipotez edilen etiket:	N	n
Boosting skoru:	0.021641	0.089956
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...bir nükleer santral verilmesine, Birleşmiş Milletler **ve *Amerika'nın {WB}* da mali** yaptırımları...

Örnek 2:	SelfComLexMorpTN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.020124	0.083453
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

...bir zamanlar bir bebek olduğunu hatırlattı **ve *bir {WB}* bebekten** bir katil yaratılmasının...

Örnek 3:	SelfComLexMorpTN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.019702	0.096783
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

... bazı sorunların varlığını da kabul eden sözcü Amerika'nın **bu *ülkeye {WB}* ve** halkına büyük saygısı...

Örnek 4:	SelfComLexMorpTN4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.019099	0.102647
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

...savcılık bu saldırının uzun süre planlandığını **ve *ilkinin {WB}* bir** taklidi olmadığını...

Co-Training, Self-Combined Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfComLexMorpTP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.022469	0.016940
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...örgütün yeni Avrupa sorumluları olduğu **öne *sürülüyor. {SB}* Öte** yandan dün akşam...

Örnek 2:	SelfComLexMorpTP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.021892	0.020818
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...yüzde oluz altıyla Fransa ve yüzde kırk bir ile Belçika olarak **öne *çıkıyor. {SB}* Burası** Brüksel...

Örnek 3:	SelfComLexMorpTP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.017967	0.020818
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Amerika'nın sesi yayını **devam *ediyor. {SB}* Burası** Washington...

Örnek 4:	SelfComLexMorpTP4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.005979	0.100911
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Halit Meşal'in yarın Şam'da bir araya **gelecekleri *bildirildi. {SB}* Filistinli** bir milletvekili...

Co-Training, Self-Combined Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfComLexMorpFN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.008666	0.076294
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...senatör Hillary Clinton, eski başkan Bill **Clinton'ın *eşi {SB}.* Hillary** Clinton ön seçimlerde...

Örnek 2:	SelfComLexMorpFN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.008930	0.075599
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...hiç bir zaman unutmayan Sourgage makine **mühendisliği *mezunu. {SB}* Eğitim** sonrasında teknik...

Örnek 3:	SelfComLexMorpFN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.008573	0.074569
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...için yaygın olarak kullanılan ilacın yan etkileri buna yeni **bir *örnek. {SB}* Yeni** bir araştırma...

Örnek 4:	SelfComLexMorpFN4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.008666	0.069021
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...iş adamları derneğiyle Türk Alman Sanayi Odası'nın **ikinci *başkanı. {SB}* Kendini** Almanya çapında...

Açıklamalar: Yukarıdaki örneklerde, ilgili kelimelere ilişkin mono-gram, bi-gram ve tri-gramlar L verisinde genel olarak kelime sınırı sınıfına ait durumlardır ve bu kelimelerle oluşturulabilecek tüm n-gramların kelime sınırına ait bir örnekten gelme olasılığı, cümle sınırına ilişkin bir örnekten gelme olasılığına göre daha yüksektir. Bu nedenden dolayı Sözcüksel model kelime sınırı kararı vermiştir. Öte yandan her dört kelime de isimdir. Sözcüksel özelliklere ilişkin n-gramlar gibi Biçimbilgisel "PreviousLast3", "CurrentLast3" ve "NextLast3" özellikleri ve bu özelliklerin ikili ve üçlü kombinasyonları da cümle sınırına ilişkin bir ipucu vermemektedir. Bu nedenden dolayı Biçimbilgisel model de kelime sınırı kararı vermiştir.

Co-Training, Self-Combined Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	SelfComLexMorpFP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.003365	0.051583
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Washington'daysa yirmi üç **otuzu *gösteriyor, {WB}* günaydın** ben Özge Övün. **{SB}** Bu yayınımızda...

Örnek 2:	SelfComLexMorpFP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.005455	0.046749
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

... henüz yeni açan dinleyicilerimize tekrar **günaydın *diyoruz {WB},* yayınlarımızın** yapıldığı...

Örnek 3:	SelfComLexMorpFP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.011221	0.023505
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Amerika'nın **sesini *dinliyorsunuz, {WB}* burası** Washington. **{SB}** Filistin yönetimi...

Örnek 4:	SelfComLexMorpFP4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.010885	0.000850
lastMarkerA3sg:		1
lastMarkerNom:		1
lastIGhasVerb:		1
lastPOS:		?

...ona göre ayarlamalarına **yardımcı *olur {WB}* şeklinde** konuştu. **{SB}** İranla ilgili...

Açıklamalar: İlk üç örnekte, dil bilgisi kuralları bakımından belirtilen kelime sınırı cümle sınırı olabilecek bir kelimedir ancak ilgili ses kaydında konuşmacının bu cümleleri birleştirdiği görülmektedir. Bu nedenden dolayı data etiketleme esnasında bu kelimeler kelime sınırı olarak işaretlenmiştir ancak modelin hipotezine göre bu sınırlar cümle sınırlarıdır.

Sözcüksel model açısından bakıldığında, "gösteriyor", "diyoruz", "dinliyorsunuz" mono-gramlarından oluşan ve bu örneklere ilişkin oluşturulacak olan bi-gram ve tri-gramlar L verisinde genellikle cümle sınırı olarak etiketlenmiştir. Biçimbilgisel model açısından bakacak olursak, bu üş kelimenin yüklem olması ve L verisinde bu kelimelerden üretilecek "CurrentLast3" özelliklerinin L verisinde ağırlıklı olarak cümle sınırına ilişkin örneklerde bulunuyor olması, Biçimbilgisel modelin cümle sınırı kararını vermesinde etkili olmuştur.

Dördüncü örnekte ise ilgili kelime sınırını yapılan alıntının son kelimesi olarak değerlendirebiliriz. (Alıntı yapan kişinin cümlesinin son kelimesi bu kelimedir.) Sözcüksel model açısından alıntı tespitine (kelime sınırı tespiti) yönelik tek ipucu "olur-şeklinde" bi-gramı iken "olur" mono-gramı ve "yardımcı-olur" bi-gramı cümle sonlarında daha sık rastlanan n-gramlardır. Biçimbilgisel model açısından bakacak olursak "olur" kelimesini "ol" kökünden türemiş bir yüklem olarak düşünebiliriz. Bu nedenden dolayı dördüncü örnek te kelime sınırı olarak tespit edilmiştir.

Co-Training, Self-Combined Algoritması, Bürünsel ve Sözcüksel Özellikler
Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfComProsLexTN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.758749	0.369236
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Telefon bağlantısı	

...dev salonun ortasında beyaz **bir *platformda {WB}* ve** yoğun spot ışıkları altında...

Örnek 2:	SelfComProsLexTN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.758749	0.369236
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Fon müziği	

...ulusal meselesi olmayan **ve *belli {WB}* bir** grubun faaliyetlerinin ürünü olan Ermeni...

Örnek 3:	SelfComProsLexTN3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	1.485941	0.040909
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	

...koltuğuna oturacak. Eski Marksist gerilla lideri **Ortega *ilk {WB}* kez** bin dokuz yüz seksenlerde...

Örnek 4:	SelfComProsLexTN4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	1.421686	0.040909
Konuşmacı:	Cem Dalaman	
Konuşma ortamı:	Gürültü	

...zayıf konumu gibi öncelikli konulara **girmediğini *öne {WB}* sürdü. {SB}** Buna karşılık...

Co-Training, Self-Combined Algoritması, Bürünsel ve Sözcüksel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfComProsLexTP1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.467360	0.078945
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	

...çağrısında bulundu. İsrail bu tehdidi çok ciddiye **aldığını *açıkladı. {SB}* Amerika** ve Avrupa...

Örnek 2:	SelfComProsLexTP2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.467360	0.069953
Konuşmacı:	Hale Ebiri	

Konuşma ortamı:	Fon Müziği	

...Günaydın ben Hale Ebiri. Irak'ta dün de pazar yerleri **hedef *alındı. {SB}* Irak'a** yeni asker...

Örnek 3:	SelfComProsLexTP3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	S
Boosting skoru:	0.266047	0.079646
Konuşmacı:	Devrim Çubukçu	
Konuşma ortamı:	Stüdyo	

...Kenya Hükümeti gerekli her türlü önlemi **aldığını *bildirdi. {SB}* Somali** Başbakanı...

Örnek 4:	SelfComProsLexTP4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.205780	0.080002
Konuşmacı:	Hülya Polat	
Konuşma ortamı:	Stüdyo	

...konuşmasında köleliğe son vereceğinin **sinyalini *vermişti. {SB}* Franklin** Ruthivert ise...

Co-Training, Self-Combined Algoritması, Bürünsel ve Sözcüksel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfComProsLexFN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.180592	0.259217
Konuşmacı:	Cem Dalaman	
Konuşma ortamı:	Stüdyo	

...yönetim kurulu üyeliğine de seçilen **bir *isim. {SB}* Ayrıca** Türk Alman işadamları derneği ile...

Örnek 2:	SelfComProsLexFN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.031349	0.259217
Konuşmacı:	Mevlüt Katık	
Konuşma ortamı:	Telefon bağlantısı	

...diye sordu igili yazısında ve cevap **verdi. {SB} *Londrada. {SB}.* Bir** başka İngiliz gazetesi...

Açıklamalar: Birinci örnekte kısa duraklama süresi Bürünsel modelin kelime sınırı kararı vermesine neden olmuştur. Sözcüksel model açısından bakacak olursak, “previous word” özelliğinin genel olarak ardından gelen kelimenin cümle sınırı olmadığı “bir” kelimesi, “next word” kelimesinin de genel olarak bağlaç gibi kullanılan “ayrıca” kelimesi oluşu ve “current word” özelliğinin de cümle sonlarında genel olarak rastlanmayan “isim” kelimesi oluşu nedeniyle bu kelimelerle oluşturulacak tüm n-gramlar L verisinde kelime sınırı örnekleri olarak bulunmaktadır. Bu nedenlerden dolayı Sözcüksel model kelime sınırı kararını vermiştir.

İkinci örnekte ise ilgili sınır tek kelimelik bir cümlenin sınırı veya devrik cümlenin son kelimesi olarak düşünülebilir. Bürünsel model açısından “verdi” kelimesi cümle sınırıdır. Kullanılan haber verilerinde tek kelimelik cümlelere çok az rastlandığı için, Bürünsel model bu kelimeyi biten bir cümlenin ardından yeni başlayan cümlenin ilk kelimesi olarak algılamıştır. Sözcüksel model açısından bakacak olursak, “Londrada”, “verdi-Londrada”, “Londrada-bir” ve “verdi-Londrada-bir”, önceki cümlenin “verdi” kelimesiyle, sonraki cümlenin “Londrada-bir” kelimeleriyle başlama olasılığının daha yüksek olduğunu görebilmekteyiz. Bu nedenden dolayı ikinci örnek te kelime sınırı olarak hipotez edilmiştir.

Co-Training, Self-Combined Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	SelfComProsMorpTN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.113135	0.145150
Konuşmacı:	Güven Özalp	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

... bir konuşma yaptı. Kendisine yapılan davetin **Patrikhane'nin *önemini {WB}* ve** ekümeniklik sıfatının...

Örnek 2:	SelfComProsMorpTN2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.098263	0.177212
Konuşmacı:	Elif Özmenek	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

... kuzey Amerika, Avrupa ve Asya Pasifik **bölgesindeki *Japonya {WB}* ve** Avustralya gibi ülkelerde...

Örnek 3:	SelfComProsMorpTN3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.073696	0.274484
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		?

... iç işleri bakanlığı yetkilileri, psikologlar **ve *eski {WB}* tarikat** üyelerini içeren altmış beş kişiyi...

Örnek 4:	SelfComProsMorpTN4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.084118	0.085552
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		?

... dinliyorsunuz. Yayınlarımız ile ilgili bilgi ve özel **haber *söyleşilerimize {WB}* ise** internette...

Co-Training, Self-Combined Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	SelfComProsMorpTP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.029305	0.073401
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

... kuvvet kullanma yoluna da gidilebileceğini **ifade *etmişlerdi. {SB}* İran** Cumhurbaşkanı...

Örnek 2:	SelfComProsMorpTP2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.055653	0.022950
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

... tanıyıncaya kadar protestolarına devam **edeceğini *açıkladı. {SB}* Örgütün** lideri...

Örnek 3:	SelfComProsMorpTP3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.013294	0.036060
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		1
lastIGhasVerb:		1
lastPOS:		?

... bir varil petrolün fiyatı altmış iki doların **üzerine *çıktı. {SB}* Uluslararası** uzay istasyonuna...

Örnek 4:	SelfComProsMorpTP4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.061981	0.087489
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

... Cumartesi gününden bu yana on binlerce hindi **itlaf *edildi {SB}.* İngiliz** yetkililer yeni sağlık...

Co-Training, Self-Combined Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	SelfComProsMorpFN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.042269	0.291708
Konuşmacı:	Devrim Çubukçu	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

...sabah yayınını dinliyorsunuz. Bültenimizi Alparslan Esmer ile beraber sunuyoruz, ben **Devrim *Çubukçu. {SB}* Radyolarını** henüz açan dinleyicilerimize...

Açıklama: Bürünsel model açısından bakıldığı zaman fon müziğinin ses seviyesi, konuşmacının vurgulaması ve duraksamasının tespit edilebilmesini engellemektedir. Biçimbilgisel model açısından bakıldığı zaman kelime bir özel isim olduğu için ve cümle sınırı sınıfına ilişkin örnekler L eğitim kümesinde genellikle yüklemler olduğu için Biçimbilgisel model kelime sınırı kararını vermiştir.

Co-Training, Self-Combined Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	SelfComProsMorpFP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.061147	0.030312
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...dış ilişkiler komisyonunda soruları yanıtlayan Rice, Tahran ve Şam hükümetleri istirarlı bir **Irak *isteseydi, {WB}* bunu** kendilerinin de yapabileceklerini belirtti. {SB} Başkan Bush...

Açıklama: Bürünsel model açısından bakıldığı zaman konuşmacının uzun süre duraksama yapması modelin cümle sınırı kararı vermesine neden oluyor. Biçimbilgisel model açısından bakıldığı zaman kelime “iste-mek” yükleminden türetildiği için, bir yüklem olarak etiketlendiği için ve “CurrentLast3” özelliğinin cümle sınırı örneklerinde sıkça rastlanan “ydi” (di-li geçmiş zaman kipi) oluşu nedeniyle Biçimbilgisel modelin cümle sınırı kararı almış olduğunu görüyoruz.

Co-Training, Uzlaşma Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	AgreementLexMorpTN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.236877	0.754318
Uzlaşma skoru:	0.991195	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

... kilometrekarelik alan incelemeye alındı. Ülkedeki bütün **kuş *gösterileri {WB}* ve** güvercin yarışları...

Örnek 2:	AgreementLexMorpTN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.105212	0.778503
Uzlaşma skoru:	0.883715	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

... arası iklim uzmanları tarafından gerçekleştirilen ve **bir *hafta {WB}* süren** panel sonrasında açıklandı...

Örnek 3:	AgreementLexMorpTN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.105212	0.778503
Uzlaşma skoru:	0.883715	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

... her yerinden beş yüz uzmanın yanısıra hükümet temsilcilerinin de katılımıyla **bir *hafta {WB}* süren** kapalı toplantılar sonucunda...

Örnek 4:	AgreementLexMorpTN4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.086715	0.695158
Uzlaşma skoru:	0.781873	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

...Konuyla ilgili yazılı **bir *açıklamada {WB}* bulunan** İşçi Partisi Milletvekili...

Co-Training, Uzlaşma Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	AgreementLexMorpTP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.220832	0.512500
Uzlaşma skoru:	0.733332	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...aralarındaki görüş ayrılıklarını çözme yönünde temel atmak **istediğini *belirtti. {SB}* Öte** yandan İsraillilerin...

Örnek 2:	AgreementLexMorpTP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.131875	0.512500
Uzlaşma skoru:	0.644375	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...ilerleyebilmek için ulusal bir tartışma yapması gereğine **işaret *etti. {SB}* Dış** işleri bakanı...

Örnek 3:	AgreementLexMorpTP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.107868	0.512500
Uzlaşma skoru:	0.620368	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Kuveyt dış işleri bakanı bildirinin İran'ı hedef **aldığını *söyledi. {SB}* Rice** dün Riyad'da...

Örnek 4:	AgreementLexMorpTP4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.080454	0.683419
Uzlaşma skoru:	0.763873	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...amaçlayan toplantıların devamını görüşmek üzere Berlinde bir **araya *geldi. {SB}* Amerika** dış işleri...

Co-Training, Uzlaşma Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	AgreementLexMorpFN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.236877	0.754318
Uzlaşma skoru:	0.991195	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...ülkeler yüzde altmış üç ile Fransa, yüzde altmış ili ile Danimarka **ve *İtalya. {SB}* Yüzde** onla Letonya...

Örnek 2:	AgreementLexMorpFN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.086715	0.531780
Uzlaşma skoru:	0.618495	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		?

...ideal ismi bulur. Hem bir Türk hem de **bir *kadındır. {SB}* Keskinler** federasyonda üstlendiği...

Örnek 3:	AgreementLexMorpFN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.085350	0.092308
Uzlaşma skoru:	0.177658	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...bunun önlenmesi için İran ile diyalog yöntemlerinin **aranmasını *öneriyor. {SB}* [~Ve]** İran'a yapılacak...

Açıklamalar:

Birinci örnekte cümle bir ülke ismi ile bitmiştir. Bu nedenden dolayı Biçimbilgisel model bu sınırı kelime sınırı olarak hipoez etmiştir. Sözcüksel model açısından bakacak olursak L verisinde “ve” bağlacını içeren n-gramlar ağırlıklı olarak kelime sınırı sınıfına ait örneklerde bulunduğu için Sözcüksel modelin hipotezi kelime sınırı yönünde olmuştur. Benzer bir şekilde Biçimbilgisel özelliklerde kelimenin son üç harfine, önceki ve sonraki kelimelerin de son üç harfine bakılıp bunlardan Sözcüksel özelliklere benzer özellikler çıkarıldığı için “ve” bağlacı Biçimbilgisel modeli de etkilemektedir.

İkinci örnekte Sözcüksel ve Biçimbilgisel model için kelimenin öncesinde “bir” kelimesinin bulunuyor olması, kelime sınırı olasılığının daha yüksek olmasına bir işaret olarak algılanmış ve modeller kelime sınırı kararını almıştır.

Üçüncü örnekte Sözcüksel model için “öneriyor-ve” bi-gramı, Biçimbilgisel model için de “yor-ve” özelliği burada bulunan “ve” kelimesinin bir bağlaç olarak algılanmasına böylece “öneriyor” kelimesi bir yüklem olmasına rağmen bu kelimenin bu örnek için diğer örneklere kıyasla çok daha düşük bir Biçimbilgisel model skoru ile kelime sınırı olarak hipotez edilmesine neden olmuştur. El ile etiketleme yapılırken ses kaydı da baz alındığı için ve ilgili ses kaydında konuşmacı cümlesini “öneriyor” kelimesi ile bitirip yeni cümlesine “ve” bağlacıyla başladığı için (“ve” kelimesi belli belirsiz telaffuz ediliyor.) orijimal veride cümle sınırı olarak etiketlenmiştir.

Co-Training, Uzlaşma Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	AgreementLexMorpFP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.026092	0.245860
Uzlaşma skoru:	0.271952	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...kongre salonuna **sessiz *geldi {WB},* yarım** saat kaldıktan sonra geldiği gibi sessizce de ayrıldı. **{SB}** Toplantıya...

Örnek 2:	AgreementLexMorpFP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.025386	0.229710
Uzlaşma skoru:	0.255096	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		?

...güneydeki Musakale kasabasını basarak polisleri etkisiz **hale *getirmiş, {WB}* aşiret** resilerini bir süre rehin almıştı. **{SB}** Nato kuvvetleri...

Örnek 3:	AgreementLexMorpFP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.021851	0.265805
Uzlaşma skoru:	0.287656	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...uyum sağlamak zorunda olduklarını **öne *sürüyor, {WB}* ama** söz konusu toplumların da göçmenleri benimseyen politikalara yönelmesinin kaçınılmaz olduğuna inanıyor. **{SB}** Cem...

Örnek 4:	AgreementLexMorpFP4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.018913	0.396026
Uzlaşma skoru:	0.414939	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...bir çok ev ve iş yerine baskın **yapıldığını *açıkladı, {WB}* ancak** operasyon hakkında ayrıntılı bilgi vermedi. **{SB}** İngiliz Sky...

Açıklamalar: Yukarıdaki örneklerde ilgili kelimeler yüklemdir. Dil bilgisi kurallarına göre bu kelimelerin ardından başlayan cümleler, bu cümleden bağımsız yeni bir cümle gibi düşünülebilir. Sözcüksel model açısından bakıldığı zaman "geldi", "getirmiş", "sürüyor", "açıkladı" mono-gramları ve bu örneklere ilişkin bi-gram ve tri-gramlar, L verisinde ağırlıklı olarak cümle sınırı şeklinde etiketlenmiştir. Sadece üçüncü ve dördüncü örnekler için "nextword" özelliğinin "ancak" ve "ama" bağlaçları oluşu kelime sınırı için bir ipucu teşgil etmektedir. Benzer bir şekilde Biçimbilgisel model açısından bakacak olursak; yüklemler L verisinde ağırlıklı olarak cümle sınırına ilişkin örneklerde bulunmaktadır. Buna ilave olarak "CurrentLast3" özellikleri L verisinde genel olarak cümle sınırlarına ilişkin örneklerde bulunan "yor", "l**di**", "a**dı**" ve "miş" gibi yüklemin zaman çekimi eklerini içermektedir. Bu nedenlerden dolayı bu kelimeler Sözcüksel ve Morfolojijk modeller tarafından cümle sınırı olarak algılanmıştır. Konuşmacı bu cümleleri telaffuz ederken iki cümleyi birleştirdiği için el ile etiketleme esnasında bu kelimelerin sınırı orijinal veride kelime sınırı olarak görünmektedir.

Co-Training, Uzlaşma Algoritması, Bürünsel ve Sözcüksel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	AgreementProsLexTN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.610866	0.369236
Uzlaşma skoru:	0.980102	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

...hükümetler arası iklim uzmanları tarafından gerçekleştirilen **ve *bir {WB}* hafta** süren panel sonrasında...

Örnek 2:	AgreementProsLexTN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.628906	0.369236
Uzlaşma skoru:	0.998142	
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Stüdyo	

...dokuz kişi gözaltına alındı. Polis Burmingham **ve *çevresinde {WB}* bir** çok ev ve iş yerine...

Örnek 3:	AgreementProsLexTN3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.566890	0.349003
Uzlaşma skoru:	0.915893	
Konuşmacı:	Değer Akal	
Konuşma ortamı:	Telefon bağlantısı	

...Dış işleri bakanı Abdullah **Gül *bir {WB}* basın** toplantısı düzenleyerek...

Örnek 4:	AgreementProsLexTN4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.566890	0.348813
Uzlaşma skoru:	0.915703	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	

...yeni inşaat projeleri ile **ilgili *yayınlanan {WB}* bir** denetçi raporunda....

Co-Training, Uzlaşma Algoritması, Bürünsel ve Sözcüksel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	AgreementProsLexTP1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.384115	0.119867
Uzlaşma skoru:	0.503982	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	

...bu tehdidi çok ciddiye **aldığını *açıkladı. {SB}* Amerika** ve Avrupa...

Örnek 2:	AgreementProsLexTP2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.530778	0.023006
Uzlaşma skoru:	0.553784	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	

...etkili olabilmesi için bütün olarak uygulanması **gerektiğini *belirtti. {SB}* Grup** raporda başkan...

Örnek 3	AgreementProsLexTP3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.487438	0.022515
Uzlaşma skoru:	0.509953	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Fon müziği	

...ilindeki çatışmada yaralanan iki askerin daha **öldüğünü *açıkladı. {SB}* Önde** gelen üç amerikalı...

Örnek 4	AgreementProsLexTP4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.403372	0.021586
Uzlaşma skoru:	0.424958	
Konuşmacı:	Aydan Kızıldağlı	
Konuşma ortamı:	Stüdyo	

...yedi bin yıllık tarihini konu alan bir serginin **açılışını da *yapacak. {SB}* Amerika** dış işleri bakanı...

Co-Training, Uzlaşma Algoritması, Bürünsel ve Sözcüksel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	AgreementProsLexFN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.324110	0.075339
Uzlaşma skoru:	0.399449	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

...isimlerin dışında biri olması gerektiğini düşünenlerin oranı yüzde yirmi beş **virgül *iki. {SB}* Bu** soruya kararsızım yanıtı verenlerse yüzde kırk bir virgül dört ile çoğunluğu oluşturuyor. **{SB}** ...

Örnek 2:	AgreementProsLexFN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.034785	0.236628
Uzlaşma skoru:	0.271413	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Fon müziği	

...Türkçe yayın bölümü Washington bugün yirmi dört **Ocak *Çarşamba {SB}* Türkiyede** saatler altı otuzu...

Açıklamalar: Birinci örneğe ilişkin ses kaydı dinlendiği zaman “iki” kelimesi telaffuz edilirken konuşmacı tarafından cümle sonu vurgusu (konuşma hızında yavaşlama, enerji ve tınıda düşme, kelime sınırına nazaran daha uzun bir duraksama) belirgin yapılmamaktadır. Bu örnekte “oluşturuyor” kelimesine ilişkin vurgular cümle sonuna ilişkin ipuçları daha belirgindir. İkinci örnekte de “Çarşamba” kelimesi benzer bir şekilde konuşmacı tarafından cümle sonu vurgusu yeterince belirgin yapılmazken, fon müziği de cümle sınırı tespitini zorlaştırmaktadır. Bu nedenlerden dolayı Bürünsel model kelime sınırı hipotezini yapmıştır. Sözcüksel model açısından bakıldığında ise, birinci örnekte “iki” kelimesi bir rakam ikinci örnekte de “Çarşamba” bir gün ismidir. Yukarıdaki örneklere ilişkin n-gramlar L verisinde kelime sınırına ait örneklerde daha sık bulunduğu için, Sözcüksel model de kelime sınırı kararını vermiştir.

Co-Training, Uzlaşma Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	AgreementProsMorpTN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.785399	0.214582
Uzlaşma skoru:	0.999981	
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

...yetişkinler istediklerine inanabilirler. Ama **çocuklar *bir {WB}* okula** gidip eğitim almalı...

Örnek 2:	AgreementProsMorpTN2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.509363	0.490616
Uzlaşma skoru:	0.999979	
Konuşmacı:	Değer Akal	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

...Davos zirvesinden **dönen *dış {WB}* işleri** bakanı Abdullah Gül...

Örnek 3:	AgreementProsMorpTN3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.403467	0.596242
Uzlaşma skoru:	0.999979	
Konuşmacı:	Hülya Polat	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...bu ilaçları hastalarına verirken tedaviye en az **dozdan *başlamaları {WB}* konusunda** uyarıyor...

Örnek 4:	AgreementProsMorpTN4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.392474	0.607229
Uzlaşma skoru:	0.999703	
Konuşmacı:	Cem Dalaman	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		?

...ama bunlara ulaşmak için çok çalışmak **gerektiğini de *hiç {WB}* bir** zaman unutmayan...

Co-Training, Uzlaşma Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	AgreementProsMorpTP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.277214	0.569328
Uzlaşma skoru:	0.846542	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Burası Washington sabah yayınımız **devam *ediyor. {SB}* Fransa'nın**...

Örnek 2:	AgreementProsMorpTP2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.272779	0.569328
Uzlaşma skoru:	0.842107	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...en az yedi kişinin ölümüne yirmi kişinin de yaralanmasına **yol *açtı. {SB}*** **Amerikan** ordusu...

Örnek 3:	AgreementProsMorpTP3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.266889	0.569328
Uzlaşma skoru:	0.836217	
Konuşmacı:	Devrim Çubukçu	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...Türk askeri birliğini de **ziyaret *etti. {SB}*** **Bu** arada...

Örnek 4:	AgreementProsMorpTP4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.254060	0.569328
Uzlaşma skoru:	0.823388	
Konuşmacı:	Elif Ural	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...zarar verir **şeklinde *konuştu. {SB}*** **Diğer** taraftan...

Co-Training, Uzlaşma Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	AgreementProsMorpFN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.795726	0.189173
Uzlaşma skoru:	0.984899	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

...Burası Amerika'nın Sesi Türkçe yayın **bölümü *Washington. {SB}* Bugün** on sekiz ocak...

Örnek 2:	AgreementProsMorpFN2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.904231	0.066395
Uzlaşma skoru:	0.970626	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Verb

...her zaman olduğu gibi dünyadan son gelişmelerle de **birlikte *olacağız. {SB}* Amerika'nın** sesi yayınlarını...

Örnek 3:	AgreementProsMorpFN3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	n
Boosting skoru:	0.803041	0.131976
Uzlaşma skoru:	0.935017	
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		?

...ancak meclis komisyonu başkanı Pereş bizim işimiz tarikatlar ve onların inanç **biçimi *değil. {SB}* Yetişkinler** istedikleri...

Açıklamalar: Yukarıdaki üç örneğe Bürünsel model açısından bakıldığı zaman, konuşmacının cümle sonundan ziyade iki cümle bütün bir cümleymiş gibi vurguladığını görmekteyiz. Bu nedenden dolayı ilgili kelime, Bürünsel model tarafından kelime sınırı olarak hipotez edilmiştir. Birinci ve ikinci ses kayıtlarında söylenen kelime dizisi, her radyo programının başında söylenen bültene başlangıç cümleleri olduğu için ve bu kelimelere ait sınırlar L verisinde genel olarak kelime sınırı olduğu için Biçimbilgisel model de kelime sınırı kararı vermiştir. Biçimbilgisel modelin bu kararında "PreviousLast3", "CurrentLast3" ve "NextLast3" özelliklerinin etkisini görmekteyiz. Üçüncü örnekte "değil" kelimesinin "lastMarkerA3sg" özelliğinin Logic 1 ile ifade edilmiş olması bu kelimenin yükleme benzer bir özellik gösterdiğini belirtmekte, ancak "lastPOS" özelliği, Biçimbilgisel özellik çıkarımı esnasında ikileme sebep olmuştur. Bunun sebebi Biçimbilgisel özellik çıkarımı esnasında kullandığımız programın, kelimenin "de-mek" kökün ile olumsuzluk anlamındaki "değil" ismi ile arada kalmış olmasıdır. "değil" kelimesi bir bağlaç gibi de kullanılabildiğinden dolayı Biçimbilgisel model kelime sınırı hipotezini yapmıştır.

Co-Training, Uzlaşma Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	AgreementProsMorpFP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	s
Boosting skoru:	0.104581	0.569328
Uzlaşma skoru:	0.673909	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...ile ilgisi olduğunu **iddia *etti. {WB}* Ancak** Alman savcılığı...

Açıklama: Bürünsel model açısından baktığımız zaman, ses kaydını dinlediğimizde konuşmacının cümleyi bitirip, "Ancak" bağlacı ile yeni cümleye başladığını görüyoruz. Bu nedenden dolayı Bürünsel model cümle sınırı hipotezinde bulunmuştur. Öte yandan Biçimbilgisel model açısından baktığımızda "etti" yükleminin L verisinde cümle sınırı örneği olarak oldukça fazla karşımıza çıktığını görüyoruz. Bu nedenden dolayı Biçimbilgisel modelin de hipotezi kelimenin cümle sınırı olduğu yönündedir.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	DisagreementLexMorpTN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.183032	0.025719
Uzlaşmama skoru:	0.15731300 (n)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

… eski diktatörün öldüğü hastane **çevresinde *toplandı {WB}* ve** kendisini öven…

Açıklama: Biçimbilgisel model kelime yüklem olduğu için cümle sınırı hipotezini yapmıştır. Buna ilave olarak "CurrentLast3" özelliğinin "n**dı**" yüklemin di'li geçmiş zamanı içeriyor olması da cümle sınırı için önemli bir ipucudur. Sözcüksel model ise n-gramlarda "next word" özelliğinin "ve" bağlacı olması kelime sınırına ilişkin çok önemli bir ipucu teşgil etmektedir. Zira "ve" kelimesi bir bağlaç olarak kullanılır ve bu bağlaçla cümleler başlamaz. Bu nedenden dolayı Sözcüksel model kelime sınırı tespitini yapmıştır.

Örnek 2:	DisagreementLexMorpTN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.012754	1.234909
Uzlaşmama skoru:	1.22215500 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

… başladığımız yayınımızın sonuna yaklaşırken **Hale *Ebiri {WB}* dinlediğiniz** haberlerden özetler…

Açıklama: Morfolojk model kelime bir isim olduğu için kelime sınırı hipotezini, Sözcüksel model ise konuşmacılar cümlelerini bitirirken isim ve soyisimlerini söyledikleri için cümle sınırı hipotezini yapmıştır. İsmi geçen konuşmacı, haber bülteninin ana sunucularından biri olduğu için ve her bültenin başında "Günaydın ben Hale Ebiri" cümlesini mutlaka kurduğu için, L verisinde "Ebiri" ve "Hale-Ebiri" mono-gram ve bi-gramları arasında cümle sınırı olarak etiketlenmiş fazla sayıda örnek bulunmaktadır. Aynı şekilde "Next-word" özelliğinin "Dinlediğiniz" kelimesi olan sık sayıda örnek L verisinde cümle sınırı etiketi ile bulunmaktadır.

"

Örnek 3:	DisagreementLexMorpTN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.012754	1.234909
Uzlaşmama skoru:	1.22215500 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		?

…görüşmelerin ardından New York'tan ayrıldı. Ben **Elif *Özmenek, {WB}* burası** New York {SB}.

Açıklama: Morfolojk model kelime bir isim olduğu için kelime sınırı hipotezini, Sözcüksel model ise konuşmacılar cümlelerini bitirirken isim ve soyisimlerini söyledikleri için cümle sınırı hipotezini yapmıştır. Buna ilave olarak, L verisinde fazla sayıda "Next Word" mono-gramının "Burası" olduğu örnekler bulunmaktadır. Bu bakımdan "burası" kelimesi önceki bir çeşit bağlaç gibi düşünülebilir.

Örnek 4:	DisagreementLexMorpTN4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.007494	1.911037
Uzlaşmama skoru:	1.90354300 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…isimli bir kuruluş yaptı. Ben **Mevlüt *Katık, {WB}* burası** Londra {SB}…

Açıklama: Morfolojk model kelime bir isim olduğu için kelime sınırı hipotezini, Sözcüksel model ise konuşmacılar cümlelerini bitirirken isim ve soyisimlerini söyledikleri için cümle sınırı hipotezini yapmıştır. Buna ilave olarak, L verisinde fazla sayıda "Next Word" mono-gramının "Burası" olduğu örnekler bulunmaktadır.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	DisagreementLexMorpTP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.014017	0.123336
Uzlaşmama skoru:	0.10931900 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...ticaret açığının azalması konut piyasasındaki daralmanın **etkisini *sildi. {SB}* Bu** arada Amerikan Merkez Bankası dünkü...

Örnek 2:	DisagreementLexMorpTP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.053868	0.214392
Uzlaşmama skoru:	0.16052400 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...ilk soruşturmasından sonra **tutuklandığı *açıklandı. {SB}* Böylece** cinayetle ilgili…

Örnek 3:	DisagreementLexMorpTP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.004716	0.155773
Uzlaşmama skoru:	0.15105700 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…basın toplantısıyla ilgili ayrıntılı haberi az **sonra *dinleyebilirsiniz. {SB}* Günaydın** ben Hale Ebiri {SB}…

Açıklamalar: Yukarıdaki üç örnekte de işaretlenmiş kelimeler birer yüklem olduğu için Biçimbilgisel model cümle sınırı hipotezini yapmıştır. Sözcüksel model açısından bakıldığı zaman, birinci ve ikinci örneklerde "nextword" monogramları sırasıyla "bu" ve "böylece" kelimeleridir. "Bu" kelimesi önce söylenen bir ifade ile ilgili aynı cümlede detay veya açıklama yaparken kullanıldığından, "böylece" kelimesi de bir bağlaç olduğundan L verisinde "nextword" monogramı bu kelimelerden oluşan ve kelime sınırı sınıfında bulunan bir çok örnek bulunmaktadır. Üçüncü örneğe baktığımızda "nextword" monogramının "günaydın" kelimesi olduğunu görüyoruz. Konuşmacılar haber bültenlerinin başlangıcında genel olarak şu cümleyi kurdukları için "...*Türkiyede WB saatler WB yürmü WB üç WB otuzu WB gösteriyor* ***WB günaydın*** *WB ben WB [konuşmacının adı soyadı]...*" "nextword" mono-gramının "günaydın" kelimesi olduğu bir çok örnek L verisinde kelime sınırı olarak bulunmaktadır. Bu sebeplerden dolayı Sözcüksel model bu örnekleri kelime sınırı olarak hipotez etmiştir.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	DisagreementLexMorpFN1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.022723	1.234909
Uzlaşmama skoru:	1.21218600 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…Değer Akal, Amerika'nın Sesi **radyosu, *Ankara. {SB}* Avrupa** Birliği'ne üye…

Açıklama: Sözcüksel model "radyosu-Ankara" bi-gramı cümle sonlarında sık rastlanır bir durum olduğu için cümle sonu kararı, Sözcüksel model ise "Ankara" bir isim olduğu için kelime sınırı kararı vermiştir.

Örnek 2:	DisagreementLexMorpFN2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.022723	1.234909
Uzlaşmama skoru:	1.21218600 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…Burası Amerika'nın Sesi **radyosu *Washington. {SB}* Uzay** dolmuşu Discoverynin…

Açıklama: Örnek 1 ile aynı durum.

Örnek 3:	DisagreementLexMorpFN3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.012754	1.234909
Uzlaşmama skoru:	1.22215500 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…sabah yayınını dinliyorsunuz. Ben **Hale *Ebiri. {SB}* Radyolarını** henüz açan dinleyicilerimize…

Açıklama: Morfolojk model, kelime bir isim olduğu için kelime sınırı hipotezini, Sözcüksel model ise konuşmacılar cümlelerini bitirirken isim ve soyisimlerini söyledikleri için cümle sınırı hipotezini yapmıştır.

Örnek 4:	DisagreementLexMorpFN4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.007494	1.911037
Uzlaşmama skoru:	1.90354300 (n)	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…hava parçalı bulutlu, sıcaklık **altı *derece. {SB}* Kalbimiz** yaşamımızı kaliteli bir şekilde…

Açıklama: Konuşmacıların haber bültenlerinin başında mutlaka söyledikleri cümlelerden biri bulundukları konumdaki saati, Türkiye saatini ve bulundukları konumun hava durumunu belirten bir cümledir. Bu cümle genel olarak hava sıcaklığının söylenmesi ile son bulduğu için, bu durumda Sözcüksel model "current word" özelliğinin "derece" kelimesi oluşu nedeniyle cümle sınırı hipotezini yapmıştır. Biçimbilgisel model ise "derece" kelimesi bir yüklem olmadığı için kelime sınırı hipotezinde bulunmuştur.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Biçimbilgisel Özellikler
Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	DisagreementLexMorpFP1.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.042701	0.196328
Uzlaşmama skoru:	0.15362700 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…üye ülkeler arasındaki görüş **ayrılıklarının *sürüyor {WB}* olması** toplantının oldukça…

Açıklama: Biçimbilgisel model, kelime bir yüklem olduğu için cümle sınırı hipotezini yapmıştır. Sözcüksel modele baktığımız zaman "next word" özelliğinin "olması", "current next" özelliğinin "sürüyor olması" ikilemesi olmasından dolayı, Sözcüksel model kelime sınırı hipotezini yapmıştır.

Örnek 2:	DisagreementLexMorpFP2.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.024676	0.182146
Uzlaşmama skoru:	0.15747000 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…Fraiser'in Mogadişu'ya da **geçmesi *planlanıyordu {WB}* ancak** bu plan iptal edildi. {SB}…

Açıklama: Biçimbilgisel model, kelime bir yüklem olduğu için cümle sınırı hipotezini yapmıştır. Sözcüksel modele baktığımız zaman “next word” özelliğinin “ancak” bağlacı olmasından dolayı, Sözcüksel model kelime sınırı hipotezini yapmıştır.

Örnek 3:	DisagreementLexMorpFP3.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.018220	0.096815
Uzlaşmama skoru:	0.07859500 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

… Washington’da ise yirmi üç **otuzu *gösteriyor, {WB}* günaydın** ben Alparslan Esmer. {SB}…

Açıklama: Kullandığımız haber bültenlerinin hepsi Washington saati ile 23.30’da başlamaktadır. Bu nedenden dolayı L kümesinde kelime sınırı olarak etiketlenmiş şekilde bulunan oldukça fazla “otuzu-gösteriyor-günaydın” tri-gramı bulunmaktadır. Morfoloji modeli ise kelime bir yüklem olduğu için ve “CuurentLast3” özelliğinin şimdiki zaman çekimi “yor” oluşu nedenleriyle cümle sınırı hipotezini yapmıştır.

Örnek 4:	DisagreementLexMorpFP4.wav	
Özellik:	Sözcüksel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.014981	0.158634
Uzlaşmama skoru:	0.14365300 (s)	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…halkına dünya bize yardım **etmek *istiyor {WB}* ama** biz kendi kendimize yardım…

Açıklama: Biçimbilgisel model, kelime bir yüklem olduğu için cümle sınırı hipotezini yapmıştır. Sözcüksel modele baktığımız zaman “next word” özelliğinin “ama” bağlacı olmasından dolayı, Sözcüksel model kelime sınırı hipotezini yapmıştır.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Bürünsel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	DisagreementProsLexTN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.053758	1.152720
Uzlaşmama skoru:	1.09896200 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

…Bu çerçevede asgari ücretin **arttırılması *ve {WB} [duraksama]* kök** hücre araştırmalarına…

Açıklama: Bürünsel model açısından bakıldığında, “ve” kelimesinden sonra diğer kelimelere göre daha uzun bir duraksama süresi var. Bu nedenden dolayı Bürünsel model cümle sınırı hipotezini yapmıştır. Sözcüksel model açısından bakıldığında, “ve” bağlacının “current word” özelliği olması nedeniyle Sözcüksel model kelime sınırı hipotezini yapmıştır.

Örnek 2:	DisagreementProsLexTN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.056683	0.368522
Uzlaşmama skoru:	0.31183900 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

... Almanyada Yahudi karşıtlığı bilinen **bir *durum {WB} [duraksama]* ancak** aralarında bazı Türklerin de bulunduğu...

Açıklama: Ses kaydı dinlendiği zaman, “durum” kelimesi telaffuz edilirken konuşma hızında yavaşlama, enerji ve tınıda düşüş ve kelime telaffuzundan sonra kısmen daha uzun süreli duraksama gibi, cümle sonu vurgusunun karakteristik özellikleri gözlemlenmektedir. Bu nedenden dolayı Bürünsel model cümle sonu kararı vermiştir. Sözcüksel model ise, “nextword” özelliğinin “ancak” bağlacı olması ve “currentword” monogramının cümle sonlarında rastlanmayan “durum” kelimesi olması sebebiyle kelime sınırı kararını vermiştir.

Örnek 3:	DisagreementProsLexTN3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.588670	0.017362
Uzlaşmama skoru:	0.57130800 (n)	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	

…anlaşmanın ayrıntıları **başbakanlıkta *yapılacak {WB}* görüşmelerde** belirlenecek…

Açıklama: “Yapılacak” kelimesi bir yüklem olduğu için Biçimbilgisel model cümle sınırı hipotezinde bulunmuştur. Kelimenin telaffuzu esnasında kelime sonrasında bir duraksama, konuşmanın yavaşlaması, tını veya enerjinin düşmesi gibi cümle sonuna yönelik ipuçlarının vurguda bulunmaması nedeniyle Bürünsel model kelime sınırı hipotezinde bulunmuştur.

Örnek 4:	DisagreementProsLexTN4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.547105	0.012356
Uzlaşmama skoru:	0.53474900 (n)	
Konuşmacı:	Değer Akal	
Konuşma ortamı:	Telefon bağlantısı	

…belirterek bekleyip **görmek *gerekir {WB}* dedi.** {SB} Gül bu zaman zarfında…

Açıklama: Bürünsel model için örnek 3 ile benzer durum söz konusudur. Sözcüksel model için Örnek 3’ten farkı “next word” özelliğinin “dedi” kelimesi olmasıdır. Bu özelliğe bakarak ilgili kelimenin konuşmacının yaptığı alıntıya ait olan cümlenin son kelimesi olduğu anlaşılabilir. Öte yandan “previous-current” özelliğinin “görmek-gerekir” ikilemesi olmasından dolayı Sözcüksel model cümle sonu hipotezinde bulunmuştur.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Bürünsel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	DisagreementProsLexTP1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	1.047325	0.030169
Uzlaşmama skoru:	1.01715600 (s)	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Fon müziği	

…Bugün yirmi dokuz **ocak *pazartesi. {SB}* Türkiyede** saatler altı otuzu…

Açıklama: Sözcüksel model açısından bakıldığı zaman, ilgili kelimeye ait oluşturulacak olan n-gramların bu kelimin olasılıksal olarak kelime sınırına daha yakın olduğu gösterecektir. Bu nedenden dolayı Sözcüksel model kelime sınırı kararını vermiştir. “Pazartesi” kelimesi telaffuz edilirken kelime sonrasında bir duraksama, konuşmanın yavaşlaması, tını veya enerjinin düşmesi gibi cümle sonuna yönelik özelliklerin bulunması nedeniyle Bürünsel model cümle sınırı hipotezini yapmıştır.

Örnek 2:	DisagreementProsLexTP2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	1.047325	0.030169
Uzlaşmama skoru:	1.01715600 (s)	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	

…sigara için böyle bir **zorunluluk *yok. {SB}* Sigara** şirketleri bin dokuz yüz…

Açıklama: Örnek 1 ile benzer durum söz konusudur.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Bürünsel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	DisagreementProsLexFN1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.275508	0.033741
Uzlaşmama skoru:	0.24176700 (n)	
Konuşmacı:	Devrim Çubukçu	
Konuşma ortamı:	Fon müziği	

…ve yaptıkları hizmetler için **teşekkür *etti. {SB}* Amerika'nın** sesinin Washingtondaki…

Açıklama: Kısa duraksama süresi ve arka plan müziğinin bir anda değişimi Bürünsel modeli olumsuz etkilemiştir. Sözcüksel model açısından baktığımızda "current word" özelliğinin "etti" yüklemi olması örneğin cümle sınırı olarak belirlenmesinde etkili olmuştur.

Örnek 2:	DisagreementProsLexFN2.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.035037	0.368522
Uzlaşmama skoru:	0.33348500 (n)	
Konuşmacı:	Hülya Polat	
Konuşma ortamı:	Stüdyo	

…oysa bu genellikle yetişkinlerde görülen **bir *sorun. {SB}* Bir** başka araştırmada aile…

Açıklama: Sözcüksel model açısından bakıldığında, kelimenin öncesi ve sonrasındaki kelimelerin "bir" kelimesi olduğunu görüyoruz. "previous word" özelliğinin "bir" kelimesi olması ve "current word" özelliğinin "sorun" kelimesi oluşu L verisinde daha çok kelime sınırlarında rastlanan bir durumdur. Bürünsel model açısından baktığımızda ve ses kaydını dinlediğimizde, kelimenin vurgusunda cümle sonu karakteristiklerini (tını, enerji, hız değişimleri ve kelime telaffuzundan sonraki daha uzun duraksama süresi) belirgin bir şekilde görebiliyoruz. Bu nedenden dolayı Bürünsel model cümle sınırı kararını vermiştir.

Örnek 3:	DisagreementProsLexFN3.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	S	n
Boosting skoru:	0.108761	0.368522
Uzlaşmama skoru:	0.25976100 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

…ilacın yan etkileri yeni **bir *örnek. {SB}* Yeni** bir araştırma bu ilacı kullanan…

Açıklama: Örnek 2 ile benzer durum söz konusudur.

Örnek 4:	DisagreementProsLexFN4.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.049309	0.368522
Uzlaşmama skoru:	0.31921300 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	

…başkent Washington'u çok iyi bilen **bir *politikacı. {SB}* Temsilciler** meclisi üyesi iken…

Açıklama: Örnek 2 ile benzer durum söz konusudur.

Co-Training, Uzlaşmama Algoritması, Sözcüksel ve Bürünsel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	DisagreementProsLexFP1.wav	
Özellik:	Bürünsel	Sözcüksel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.292167	0.015945
Uzlaşmama skoru:	0.27622200 (s)	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	

…Sinyora **Lübnan'ın *kırk {WB} {Duraksama}* milyar** dolarlık dış borcunu…

Açıklama: Konuşmacı hata yaptığını düşünerek uzun süreli duraksama yapıyor. Bu nedenden dolayı Bürünsel model cümle sınırı hipotezi yapmıştır. Sözcüksel model açısından baktığımızda "current" özelliğinin "kırk" rakamı, "next word" monogramının "milyar" kelimesi ve "current next" bi-gramının "kırk milyar" rakamını ifade ediyor oluşu nedenlerinden dolayı Sözcüksel model kelime sınırı kararı vermiştir.

Co-Training, Uzlaşmama Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru kelime sınırı tespitleri. (True Negatives)

Örnek 1:	DisagreementProsMorpTN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.758749	0.150018
Uzlaşmama skoru:	0.60873100 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…hükümeti **inşaatı *savundu, {WB}* rampanın** onarılmasının camiye ulaşımı…

Örnek 2:	DisagreementProsMorpTN2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.758749	0.150018
Uzlaşmama skoru:	0.60873100 (n)	
Konuşmacı:	Cem Dalaman	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

...fiyatların arttığı kanaatini resmi **veriler *kanıtlamıyor {WB}* tersine** Alman İstatistik Enstitüsü'nün…

Örnek 3:	DisagreementProsMorpTN3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.758749	0.150018
Uzlaşmama skoru:	0.60873100 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…üç yüzden fazla kişi siyasi nedenlerden **dolayı *öldürüldü, {WB}* ya** da göz altında kayboldu. {SB}…

Açıklamalar: Birinci, ikinci ve üçüncü örnekte, konuşmacının vurgularına baktığımız zaman bu kelimelerin cümle sınırı olmadığını görüyoruz. Konuşmacının vurgusu daha çok bu kelimeden sonra virgül işaretinin geldiğini gösterir biçimdedir. Bu nedenden dolayı Bürünsel model kelime sınırı kararı vermiştir. Biçimbilgisel model açısından baktığımızda da bu kelimelerin yüklem olduğunu görmekteyiz. Bu nendenden dolayı Biçimbilgisel model cümle sınırı kararını vermiştir.

Örnek 4:	DisagreementProsMorpTN4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.036147	1.044671
Uzlaşmama skoru:	1.00852400 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Adj

…bir açıklama yaptı ve tahmin **ediyorum** ***önümüzdeki {WB} [duraksama] [Önümüzdeki]*** **süre** içinde beklentilerimizi…

Örnek 5:	DisagreementProsMorpTN5.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.036147	1.044671
Uzlaşmama skoru:	1.00852400 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

…Amerika'nın **denetiminde** ***olduğunu {WB} [duraksama]*** **iddia** etti. {SB} Washington ise…

Açıklamalar: Bürünsel model açısından baktığımız zaman, dördüncü örnekte konuşmacı bir duraksama ihtiyacı hissetmiş ve kaldığı kelimeyi tekrarlayarak devam etmiştir. Beşinci örnekte ise konuşmacı bir tereddüt sonucu uzun süre duraksamıştır. Bu nedenlerden dolayı Bürünsel model cümle sınırı hipotezini yapmıştır. Biçimbilgisel model açısından baktığımızda dördüncü örnekteki kelimenin bir sıfat ve beşinci kelimenin bir isim oluşu nedeniyle Biçimbilgisel model kelime sınırı hipotezini yapmıştır.

Co-Training, Uzlaşmama Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Doğru cümle sınırı tespitleri. (True Positives)

Örnek 1:	DisagreementProsMorpTP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.070468	1.165857
Uzlaşmama skoru:	1.09538900 (s)	
Konuşmacı:	Devrim Çubukçu	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…doğum yeri Beytüllahimde ve **Kudüste de *kutlandı. {SB}* Gece** yarısında yapılan ayin için…

Örnek 2:	DisagreementProsMorpTP2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.052140	1.165857
Uzlaşmama skoru:	1.11371700 (s)	
Konuşmacı:	Arzu Çakır	
Konuşma ortamı:	Telefon bağlantısı	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

…yasanın görüşmeleri sırasında oldukça ilginç **sahneler *yaşandı. {SB}* Örneğin** bundan yirmi beş yıl önce…

Açıklamalar: Bürünsel model açısından bakıldığında, birinci örneğin ses kaydında arkaplan müziğinin cümle sonu tespitini zorlaştırdığını, ikinci örnekte de konuşmacının iki cümle arasında verdiği kısa duraksama süresinin cümle sonu tespitini zorlaştırdığını görüyoruz. Biçimbilgisel model açısından bakıldığında kelimelerin yüklem oluşu, ve "CurrentLast3" özelliğinin L verisinde cümle sonu sınıfında sıkça rastlanan "ndı" harfleri olması nedeniyle Biçimbilgisel model cümle sınırı tespitini yapmıştır.

Co-Training, Uzlaşmama Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı kelime sınırı tespitleri. (False Negatives)

Örnek 1:	DisagreementProsMorpFN1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.758749	0.150018
Uzlaşmama skoru:	0.60873100 (n)	
Konuşmacı:	Hale Ebiri	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

… bulunduğu Washington'daysa yirmi üç **otuzu *gösteriyor. {SB}* Günaydın** ben Alparslan Esmer {SB}…

Örnek 2:	DisagreementProsMorpFN2.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.758749	0.150018
Uzlaşmama skoru:	0.60873100 (n)	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

… bulunduğu Washington'daysa yirmi üç **otuzu *gösteriyor. {SB}* Günaydın** ben Özge Övün {SB}…

Açıklamalar: Yukarıdaki her iki örnekte de konuşmacılar iki ayrı cümleyi, arada bir vürgül varmış gibi vurgulayarak tek bir cümleye birleştirmiştir. Bu nedenden dolayı Bürünsel model kelime sınırı hipotezinde bulunmuştur. Biçimbilgisel model ise, "gösteriyor" kelimesi bir yüklem olduğu için ve "CurrentLast3" özelliği L verisinde cümle sınırı sınıfında sıkla rastlanan "yor" (fiilin şimdiki zaman çekimi) olduğu için cümle sınırı hipotezini yapmıştır.

Örnek 3:	DisagreementProsMorpFN3.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.015379	1.006537
Uzlaşmama skoru:	0.99115800 (n)	
Konuşmacı:	Alparslan Esmer	
Konuşma ortamı:	Fon müziği	
lastMarkerA3sg:		0
lastMarkerNom:		0
lastIGhasVerb:		0
lastPOS:		Noun

…Hamas ve El Fetih arasındaki ateşkes **diken *üstünde. {SB}* Amerika** Dış İşleri bakanı…

Örnek 4:	DisagreementProsMorpFN4.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	s	
Hipotez edilen etiket:	s	n
Boosting skoru:	0.015268	1.006537
Uzlaşmama skoru:	0.99126900 (n)	
Konuşmacı:	Cem Dalaman	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		0
lastMarkerNom:		1
lastIGhasVerb:		0
lastPOS:		Noun

…zamanı geldiğini düşünüyor**, Ekin *Deligöz. {SB}* Federal** milletvekili Deligöz…

Açıklamalar: Bürünsel model açısından bakıldığı zaman, üçüncü ve dördüncü örneklerde “üstünde” ve “Deligöz” kelimelerinin gittikçe yavaşlayan bir hızla, azalan enerji ve tınıyla, kelime telaffuzundan sonra bırakılan kelime sınırına kıyasla daha uzun bir duraksamayla telaffuz edildiğini görüyoruz. Cümle sınırına ilişkin tüm bu ipuçları mevcut olduğu için Bürünsel model cümle sınırı kararını vermiştir. Biçimbilgisel model açısından bakıldığında ise üçüncü ve dördüncü örneklerin birer isim olduğunu görmekteyiz. Bu nedenden dolayı Biçimbilgisel model kelime sınırı kararı verniştir. Dördüncü örneği Biçimbilgisel modelin aşağıda gösterildiği gibi etiketlemiş olduğunu düşünebiliriz.

*…zamanı WB geldiğini WB **düşünüyor SB** Ekin WB Deligöz WB Federal WB milletvekili WB Deligöz…*

Co-Training, Uzlaşmama Algoritması, Bürünsel ve Biçimbilgisel Özellikler

Hatalı cümle sınırı tespitleri. (False Positives)

Örnek 1:	DisagreementProsMorpFP1.wav	
Özellik:	Bürünsel	Biçimbilgisel
Orjinal etiket:	n	
Hipotez edilen etiket:	n	s
Boosting skoru:	0.046072	1.165857
Uzlaşmama skoru:	1.11978500 (n)	
Konuşmacı:	Özge Övün	
Konuşma ortamı:	Stüdyo	
lastMarkerA3sg:		1
lastMarkerNom:		0
lastIGhasVerb:		1
lastPOS:		Verb

… yetiştirildiği bir çiftlikte H5N1 virüsü **izlerine *rastlandı, {WB}* yetkililer** soruşturmanın devam…

Açıklama: Ses kaydı dinlendiği zaman konuşmacının iki cümleyi bir virgül vasıtasıyla birleştirerek vurguladığını görmekteyiz. Bu nedenden dolayı Bürünsel model kelime sınırı hipotezinde bulunmuştur. Öte yandan “rastlandı” kelimesinin bir yüklem olması ve Biçimbilgisel “current word last3ng” özelliğinin L verisinde cümle sınırı örnekleri arasında sıkça rastlanan “ndı” harfleri olması sebebiyle Biçimbilgisel model cümle sınırı kararını almıştır.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Doç. Dr. ÜMİT GÜZ
Proje No:	111E228
Proje Başlığı:	Bürünsel, Sözcüksel Ve Biçimbilgisel Bilgiyi Kullanan Co-Training Ile Türkçe Konuşma Dilinin Otomatik Cümle Bölütlemesi.
Proje Türü:	1001 - Araştırma
Proje Süresi:	36
Araştırmacılar:	HAKAN GÜRKAN
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	IŞIK Ü. MÜHENDİSLİK F. ELEKTRONİK MÜHENDİSLİĞİ B.
Projenin Başlangıç ve Bitiş Tarihleri:	01/03/2012 - 01/03/2015
Onaylanan Bütçe:	170680.0
Harcanan Bütçe:	61953.88
Öz:	Co-training, web sayfası sınıflandırması, kelime anlam açıklaştırma ve adlandırılmış varlık tanıma gibi pek çok sınıflandırma işlevinde başarı ile kullanılan oldukça etkili bir makine öğrenme algoritmasıdır. Co-training, elle etiketlenmiş eğitim veri setine, etiketlenmemiş büyük miktarlardaki veriyi belirli miktarlarda etiketleyerek katmak suretiyle öğreticili öğrenme algoritmalarının performansını arttıran bir yarı öğreticili öğrenme metodudur. Co-training algoritmaları etiketlenmiş giriş verisine ilişkin farklı view lar (bakış) üzerinde eğitilmiş iki veya daha fazla sınıflandırıcının üretilmesi ve daha sonra bu sınıflandırıcıların etiketlenmemiş veriyi ayrı ayrı etiketlemesi için kullanıldığı algoritmalardır. Otomatik olarak en güvenilir biçimde etiketlenmiş örnekler daha sonra insanlar tarafından manuel olarak etiketlenmiş veriye katılmaktadır. Bu işlem pekçok defa devam ettirilmektedir. Bu projede konuşma verisine ilişkin bürünsel, sözcüksel ve biçimbilgisel bilgilerin view olarak kullanıldığı co-training ile cümle bölütlemenin gerçekleştirilmesi ele alınmıştır. Cümle Bölütleme işlevi standart konuşma tanıyıcılarının çıkışından elde edilen işlenmemiş kelime dizisi biçimindeki veriyi zenginleştirmeyi amaçlayan bir işlemdir. Bu işlemin rolü, kelime dizisi biçiminde olan verinin cümle ünitelerine ayrılmasını sağlamaktır. Cümle Bölütleme konuşma anlamaya kadar olan süreçte ilk adımdır. Cümle bölütleme işlevi, çözümleme, makine çevirimi, bilgi çıkarımı gibi cümle bölütlemenin yapıldığının varsayıldığı konuşma işlemenin daha ileri uygulamaları için bir ön adım olarak gerçekleştirilmektedir. Cümle sınırları belirlendikten sonra bu cümleler üzerinde daha ileri düzeydeki sözdizimsel ve/veya anlamsal analizler gerçekleştirilebilmektedir. Bu projede konuşma özellikleri (bürünsel, sözcüksel ve biçimbilgisel) ayrışık ve doğal özellik seti veya view olarak ele alınmış ve bu özellik setlerinin co-training algoritması ile kullanılması ile baseline sistemin performansının arttırılmasına çalışılmıştır. Amaç, Bürünsel, sözcüksel ve biçimbilgisel özelliklerinin çıkarıldığı çok az miktarda etiketlenmiş veri ile başlanarak büyük miktardaki etiketlenmemiş veriden etiketlenmiş veri miktarını arttırmaya çalışmaktır. Ayrıca, co-training için uzlaşma ve uzlaşmama adı verilen farklı öğrenme stratejileri de araştırılmıştır. Buna ek olarak, self-combined adını verdiğimiz ve self-training ile co-training yaklaşımlarının bir araya getirildiği bir yaklaşım da öne sürülmüştür.
Anahtar Kelimeler:	Co-training, cümle bölütleme, bürünsel, biçimbilgisel, sözcüksel, kendi kendine eğitme, boosting
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır

Projeden Yapılan Yayınlar:	1- Türkçe Haber Yayını Verileri için Bürünsel Bilginin Çıkarılması ve Cümle Bölütlemede Kullanılması Extracting the Prosodic Information for Turkish Broadcast News Data and Using on the Sentence Segmentation Task (Bildiri - Ulusal Bildiri - Sözlü Sunum), 2- Extraction and Comparison of Various Prosodic Feature Sets on Sentence Segmentation Task for Turkish Broadcast News Data (Bildiri - Uluslararası Bildiri - Sözlü Sunum), 3- Prosodic, Morphological and Lexical Feature Extraction of Turkish Broadcast News Data (Tez (Araştırmacı Yetiştirilmesi) - Yüksek Lisans Tezi),