

Türkçe Gazete Haberleri Dikte Sistemi

Turkish Dictation Sytem for Broadcast News Applications

Ebru Arısoy, Levent M. Arslan

Elektrik ve Elektronik Mühendisliği Bölümü, Boğaziçi Üniversitesi, 34342, Bebek, İstanbul
arisoyeb@boun.edu.tr, arslanle@boun.edu.tr

Özetçe

Türkçe gazete haberlerinin otomatik dikte edilebilmesi için bir sistem tasarlanmıştır. Türkçe sondan eklemeli bir dildir ve serbest kelime dizilimi vardır. Dilin bu özellikleri kelimeler sözlük birimleri olarak seçildiğinde konuşma tanımada dağarcık patlamasına, dağarcık dışı kelimelerin artmasına ve dilin istatistiklerinde karmaşıklığa sebep olmaktadır. Bu yüzden yeni konuşma tanıma birimleri önermekteyiz. Bir kısım sözcükler, kök, kök-sonrası ve morfemler gibi daha küçük tanıma birimlerine bölünmüş ve bu küçük birimler, bölünmemiş sözcüklerle birlikte konuşma tanıyıcıya sözlük elemanları olarak tanıtılmıştır. Bu durumda, orta boyutlu bir dağarcıkla dağarcık dışı kelime çokluğu sorunu halledilebilmiş, ve dilin istatistiksel modelleri için daha iyi kestirimler elde edilmiştir. Buna rağmen haber uygulamaları için en iyi tanıma başarımı kelime tabanlı dil modeliyedir.

Abstract

We have designed a Turkish dictation system for Broadcast news applications. Turkish is an agglutinative language with free word order. These characteristics of the language result in the vocabulary explosion, large number of out-of-vocabulary (OOV) words and the complexity of the N-gram language models in speech recognition when words are used as recognition units. Therefore, we proposed new recognition units. We parsed some of the words to smaller recognition units like stems, endings and morphemes, and introduced these smaller units and the unparsed words to the speech recognizer as lexicon entries. This way, we were able to overcome to the problem of large number of OOV words with a moderate vocabulary size and get better estimates for the N-gram language models. However, best recognition result was obtained using the word-based language model.

1. Giriş

Türkçe sondan eklemeli bir dildir. Tek bir köke bir veya daha fazla morphem eklenerek çok fazla sayıda yeni kelime oluşturulabilir [1]. Ayrıca kelime dizilimlerindeki serbestlik dilin modellenmesindeki karmaşıklığı artırmaktadır. Özellikle İngilizce konuşma tanıma sistemlerinde kelimeler dil modelleme birimleri olarak kullanılmaktadır. Bu birimlerin, Türkçe, Fince ve Korece gibi sondan eklemeli dillerde kullanılması dağarcık dışı kelimelerin artmasına sebep olur [2,3]. Türkçe için istatistiksel dil modellemesi ve bu modellerin konuşma tanımadaki başarımları üzerine yapılan önceki çalışmalarda, yeni modeller (kelimeler, morfemler, heceler, kök ve kök-sonrası) önerilmiştir [4]. Kelimeler sözlük elemanları olarak kullanıldığında dağarcık dışı kelime sayısı artmaktadır. Bu yüzden kelimeler daha küçük birimlere

bölünmüştür. Önerilen birimler konuşma tanıma başarımları açısından değerlendirildiğinde, hece ve morfemler kısa birimler oldukları için daha az akustik bilgi taşımaktadır ve kelimelere oranla daha düşük başarımlar vermektedir. Kök ve kök sonrası modeli ise kelime modelindeki dağarcık dışı kelime sayısındaki artış ve morphem, hece modellerindeki düşük tanıma başarımlarına çözüm olarak sunulmuştur. Kelimeler, morfemler ve kök ve kök-sonrası modellerinin hepsini birarada kullanan yeni bir model oluşturup; bu modeli Türkçe için haber dikte sistemine uygulamak, bu bildirinin temel konusunu oluşturmaktadır. Türkçe için İstatistiksel dil modellemesinde bu modeli kullanarak karmaşıklığı ve dağarcık dışı kelime sayısını düşürmek ve konuşma tanıma başarımını arttırmak hedeflenmektedir.

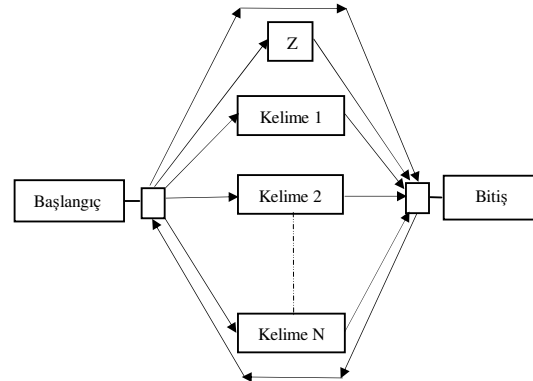
İkinci bölümde önerilen yeni modeller anlatılacak, 3. bölümde ise kullanılan veritabanının ayrıntı birim sayısı, kapsama oranı ve ikili dil modeli istatistikleri bu modeller kullanılarak incelenecektir. 4. bölümde ise modeller konuşma tanıma başarımları açısından karşılaştırılacaktır.

2. Önerilen Dil Modelleri

Bu bölümde konuşma tanıma için temel birimlerin seçimi üzerinde duracağız. Bu makalede önerdiğimiz dil modelleri konuşma tanıma birimleri olarak kelimeler, kök ve kök-sonrası ve morphem modellerinin hepsini bir arada kullanılmaktadır ve birleşik model olarak adlandırılır. Kelime modeli ve birleşik model, haber metinleri veritabanına uygulanmıştır.

2.1. Kelime Modeli

Kelime modelinde kelimeler konuşma tanımadaki sözlük elemanları olarak kullanılmış, dil modelleme olasılıkları da eğitim metinlerinden kelimeler temel birimler alınarak çıkarılmıştır. Bu model şekil 1'de gösterilmiştir.



Şekil 1: Kelime modeli

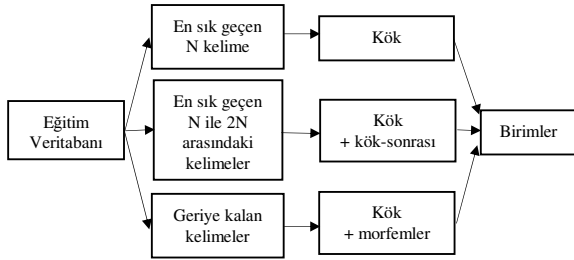
Bu modelde Z kelime aralarındaki kısa duraklamaları göstermektedir. Kelimelerin birbirini takip etme olasılıkları için ise standart ikili dil modelleri kullanılmıştır.

2.2. Birleşik Model

Birleşik Model Kaynakça 1’de önerilen modellerin hepsini birarada kullanmaktadır. Buradaki modeller şu şekilde özetlenebilir:

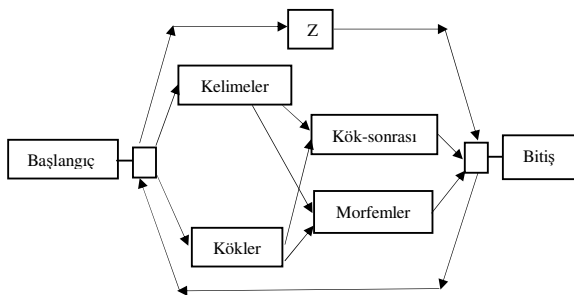
- Kelime Modeli: Bölüm 2.1’de açıklanan modeldir.
- Morfem Modeli: Bu modelde bütün kelimeler kök ve morfemlerine ayrılmış ve bütün bu parçalar sözlük elemanları olarak tanımlanmıştır. Bu modelle hem sözlük boyutu hem de dağarcık dışı kelime sayısı azaltılmıştır. Buna rağmen morfemlerin küçük birimler olması konuşma tanıma başarımını olumsuz etkilemiştir.
- Kök ve Kök-sonrası Modeli: Bu modelde bütün kelimeler kök ve kök-sonrası olarak bölünmüş ve bütün bu parçalar sözlük elemanları olarak tanımlanmıştır. Sondan eklemeli diller için bu model ilk olarak Kaynakça 5’de önerilmiş ve Kaynakça 6’da Türkçeye uyarlanmıştır. Morfem modeline göre dağarcık boyutunun artmasına rağmen, konuşma tanıma başarımı daha iyidir.

Birleşik modelin oluşturulma yordamı Şekil 2’de gösterilmiştir.



Şekil 2: Birleşik modelin oluşturulma yordamı

Şekilden de görüldüğü gibi istatistiksel dil modelleme birimleri kelimeler, kökler, köksonraları ve morfemlerden oluşmaktadır. Öncelikle eğitim metinlerindeki kelimeler geçme sıklıklarına göre sıralanmış ve en sık geçen “N” kelime kök olarak bırakılmıştır. En sık geçen “N ile 2N” arasındaki kelimeler kök ve kök-sonrası olarak bölünmüştür. Geriye kalan kelimeler de kök ve morfemlerine ayrılmıştır.



Şekil 3: Birleşik model

Kullanılan biçimbirim parçalayıcıya [4,7,8] haber metinlerinden yeni kökler de eklenerek morfemlerine ayırlamayan kelime sayısı azaltılmıştır. Birleşik model Şekil 3’de gösterilmiştir. Şekildeki oklar sözlük elemanlarının birbirlerine geçişlerini göstermektedir. Geçiş olasılıkları bu birimlerin ikili dil modelleriyle hesaplanır. Önerilen birleşik modeller N’nin 2500 ve 5000 olduğu durumlar için incelenmiş, ve bu modeller sırasıyla Birleşik-2.5K ve Birleşik-5K olarak adlandırılmıştır.

3. Metin Veritabanı İstatistikleri

3.1. Eğitim Metin Veritabanı

Bu çalışmada kullanılan eğitim veritabanı Milliyet Gazetesi’nin beş farklı alandaki bir aylık haberlerini kapsamaktadır ve toplam 355497 kelimedenden oluşur. Eğitim veritabanı Tablo 1’deki alanlara göre beş farklı gruba bölünmüştür. Gruplama yapmanın amacı farklı alanlarda haberler ekleyerek eğitim datasını genişletmenin, istatistikleri nasıl etkilediğini görmektir.

Tablo 1: Beş farklı eğitim grubu

Eğitim-1	Dünya
Eğitim-2	Dünya, Ekonomi
Eğitim-3	Dünya, Ekonomi, Güncel
Eğitim-4	Dünya, Ekonomi, Güncel, Politika
Eğitim-5	Dünya, Ekonomi, Güncel, Politika, Yaşam

3.2. Test Metin Veritabanı

Test metin veritabanı, milliyet gazetesinin beş farklı alandaki bir günlük haberlerinden toplanmış ve 7016 kelimedenden oluşmaktadır. Eğitim veritabanı ile farklı günde toplanmıştır.

3.3. Eğitim Veritabanındaki Ayrık Birim Sayısı

Ayrık birim sayısı sözlük boyutunun belirlenmesindeki en önemli kavramlardan biridir. Eğitim setinde %100 kapsama oranına ulaşmak için gerekli en küçük sözlük sayısını verir. Birimler modelden modele farklılık göstermektedir. Kelime modeli için birimler kelimeler, birleşik model için ise kelimeler, kökler, köksonraları ve morfemler olacaktır.

Tablo 2 kelime modeli için farklı eğitim gruplarının birim sayısına göre istatistiklerini vermektedir. Tablodan da görüldüğü üzere toplam ve ayrık birim sayıları çok fazladır. Her yeni eğitim grubunun eklenmesi de veritabanına yaklaşık 10K yeni kelime eklemektedir.

Tablo 2: Kelime modeli için birim istatistikleri

	Birim sayısı (kelimeler)	Ayrık birim sayısı (farklı kelimeler)	Yeni eklenen birim sayısı (farklı kelimeler)
Eğitim-1	33534	10258	10258
Eğitim-2	119657	23275	13017
Eğitim-3	185037	35399	12124
Eğitim-4	289504	46996	11597
Eğitim-5	355497	55931	8935

Tablo 3 en sık geçen 2500 kelimenin kök olarak bırakıldığı birleşik model için farklı eğitim gruplarının birim sayısına göre istatistiklerini vermektedir. Tablodan da görüldüğü üzere

toplam ve ayrı birim sayıları kelime modeline göre oldukça azdır. Her yeni eğitim grubunun eklenmesi ise veritabanına yaklaşık 3K yeni birim eklemektedir.

Tablo 3: Birleşik-2.5K modeli için birim istatistikleri

	Kelime sayısı	Birim sayısı	Ayrı birim sayısı	Yeni eklenen birim sayısı
Eğitim-1	33534	47676	5584	5584
Eğitim-2	119657	165225	9538	3954
Eğitim-3	185037	258952	13378	3840
Eğitim-4	289504	406790	15762	2384
Eğitim-5	355497	503863	18228	2466

Tablo 4 en sık geçen 5000 kelimenin kök olarak bıraktığı birleşik model için farklı eğitim gruplarının birim sayısına göre istatistiklerini vermektedir. Bu modelin önceki modele göre tek farkı kök olarak bırakılan en sık geçen kelime sayısıdır. Her yeni eğitim grubunun eklenmesi bu model için yaklaşık 4K yeni birim getirmektedir.

Tablo 4: Birleşik-5K modeli için birim istatistikleri

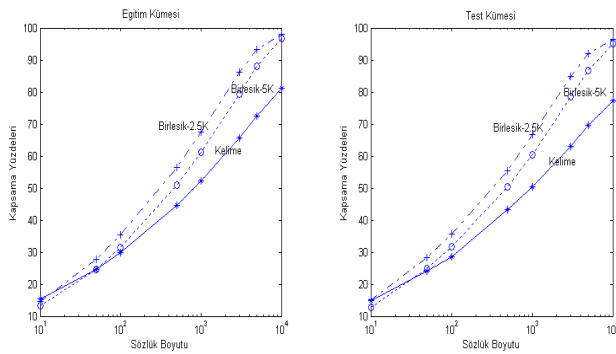
	Kelime sayısı	Birim sayısı	Ayrı birim sayısı	Yeni eklenen birim sayısı
Eğitim-1	33534	43638	6738	6738
Eğitim-2	119657	151936	11366	4628
Eğitim-3	185037	238625	15401	4035
Eğitim-4	289504	374664	17874	2473
Eğitim-5	355497	464214	20358	2484

Bu üç modeli karşılaştırdığımız zaman en az farklı birim sayısını birleşik-2.5K modeli vermektedir. Bunun sebebi ise Türkçe'nin üretken bir dil olmasından kaynaklanmaktadır. Son modeldeki ayrı birim sayısındaki artış ise daha fazla kelimenin kök olarak bırakılmasıyla açıklanabilir.

3.4. Kapsama Oranları

Kapsama oranı dağarcık dışı kelimelerin sayısı ile ters orantılıdır. Eğer bir kelime konuşma tanıma sözlüğünde bulunmuyorsa, her zaman başka bir kelime olarak tanınacaktır. Bu da konuşma tanıma başarımını düşürmektedir. Her üç model için değişen sözlük boyutuna göre eğitim ve test metinleri üzerindeki kapsama yüzdeleri Şekil 4'te verilmiştir.

Şekil 4: Kapsama yüzdeleri



Şekil 4'ten de görüldüğü üzere en düşük sözlük dışı kelime oranı birleşik-2.5K modeli iledir. En yüksek oran ise kelime

modeli ile elde edilmiştir. Örneğin "politikalaradaki" kelimesi eğitim setinde bir defa geçmesine rağmen, "politika" ve "lardaki" en sık geçen kök ve köksonrasındır. Birleşik-5K modelinde ise daha fazla kelime kök olarak bırakıldığından kapsama oranı birleşik-2.5K modeline oranla biraz daha azalmıştır.

3.5. İkili Dil Modelleri

Bu çalışmadaki ikili dil modelleri Saklı Markov Modelleri Yazılımı (HTK) [9] kullanılarak oluşturulmuştur. Önerilen modellerdeki dil modellemesi karmaşıklığı karşılaştırma ölçütümüz olacaktır. Aslında Türkçedeki kelime dizilimleri temel olarak "özne-nesne-fiil" olmasına rağmen, bir cümle beş farklı şekilde söylenebilmektedir [10]. Türkçenin serbest kelime dizilimlerine sahip bir dil olması da dil modellerindeki karmaşıklığı artırmaktadır.

Tablo 5: Eğitim metinlerinin kendi üzerindeki dil modelleme karmaşıklığı

	Kelime Modeli	Birleşik-2.5K	Birleşik-5K
Eğitim-1	753.95	208.11	322.27
Eğitim-2	711.30	171.88	264.64
Eğitim-3	936.18	201.14	305.72
Eğitim-4	957.81	192.23	285.84
Eğitim-5	1063.53	197.58	291.67

Tablo 6: Eğitim metinlerinin test metinleri üzerindeki dil modelleme karmaşıklığı

	Kelime Modeli	Birleşik-2.5K	Birleşik-5K
Eğitim-1	659.26	537.30	784.14
Eğitim-2	959.06	473.73	775.97
Eğitim-3	1105.38	384.82	636.86
Eğitim-4	1217.26	359.75	579.04
Eğitim-5	1278.17	334.45	528.91

Tablolardan da görüleceği gibi, hem eğitim metinlerinin kendi üzerinde hem de test metinleri üzerindeki karmaşıklığı birleşik modeller için daha düşüktür. En düşük değeri yine birleşik-2.5K modeli vermektedir. Her yeni eğitim grubunun eklenmesi kelime modeline çok fazla yeni ikili dil modeli eklediği için belirsizliği artırmaktadır. Bu da karmaşıklığın artmasına yol açmaktadır. Birleşik modellerde ise eklenen eğitim grupları yeni ikili dil modellerinin yanında eski ikili dil modellerini de getirmekte ve birimlerin birbirini takip etme olasılıkları için daha iyi kestirimler yapılmaktadır. Bu da karmaşıklığı azaltmaktadır.

3.6. Test Metni İstatistikleri

Test metinlerimiz Milliyet gazetesinin beş farklı alandaki bir günlük haberlerinden oluşmaktadır. Eğitim metinlerinde en sık geçen ilk 10K birim konuşma tanıma sözlüğüne eklenmiştir. Bu durumda test kümesinin istatistikleri aşağıdaki tablolarda verilmiştir.

Tablo 7: Kapsama istatistikleri (%)

Kelime Modeli	Birleşik-2.5K	Birleşik-5K
77.16	96.51	95.07

Tablo 8: İkili dil modelleme karmaşıklığı

Kelime Modeli	Birleşik-2.5K	Birleşik-5K
476.68	294.36	433.78

Hem kapsama istatistikleri hem de ikili dil modelleme karmaşıklığı açısından en iyi sonuçları birleşik-2.5K modeli vermektedir. Bir sonraki bölümde ise bu üç modelin konuşma tanıma açısından başarımları karşılaştırılacaktır.

4. Konuşma Tanıma Deneyleri

Konuşma Tanıma deneyleri için bir bayan konuşmacıdan test metninin kayıtları (16 Khz, 16-bit) kafaya sabitlenmiş gürültü temizleme özelliği olan Plantronics mikrofon ile alınmıştır. Test metni 10'ar cümlelik metinlere bölünmüş ve bu metinler sürekli bir şekilde okunarak kaydedilmiştir.

4.1. Konuşma Tanıma Sistemi

Konuşma tanıma sistemi olarak HTK [9] kullanılmıştır. Saklı Markov Modellerin eğitimi için toplam 344 kişiden alınan 10413 kayıt kullanılmıştır. Her model için üçlü fonemler kullanılan birimler esas alınarak eğitilmiştir. Tanıma başarımları değerlendirilirken doğruluk ve kesinlik yüzdeleri ölçüt alınmıştır.

$$\text{Doğruluk Yüzdesi(\%)} = \frac{N - D - S}{N} \times 100 \quad (1)$$

$$\text{Kesinlik Yüzdesi(\%)} = \frac{N - D - I - S}{N} \times 100 \quad (2)$$

Burada N toplam birim sayısını, S yerdeğiştirme hatalarını, D silme hatalarını ve I araya sokma hatalarını göstermektedir. Ayrıca her modele uygun konuşma tanıma parametrelerinin seçimi için tek bir kayıt üzerinde testler yapılmış ve her modelin en uygun parametreleri belirlenmiştir. Bunlara ek olarak konuşma tanıma sistemine konuşmacı uyarlama uygulanmıştır.

4.2. Konuşma Tanıma Deney Sonuçları

Konuşma tanıma deneylerinde her model için bölüm 3.6'da belirtilen sözlükler kullanılmış ve her model için deneyler hem konuşmacı bağımsız, hem de konuşmacı bağımlı olarak yapılmıştır. Deney sonrası bazı problemler kayıtlar testlerden çıkarılmıştır. Geriye kalan test kümesinin doğruluk ve kesinlik yüzdeleri aşağıdaki tabloda gösterilmiştir.

Tablo 9: Doğruluk ve Kesinlik Yüzdeleri

Modeller	Doğruluk (%)	Kesinlik (%)
Kelime	46.29	36.37
Kelime (uyarlanmış)	55.80	45.49
Birleşik-2.5K	45.38	35.12
Birleşik-2.5K (uyarlanmış)	55.93	44.38
Birleşik-5K	44.76	34.33
Birleşik-5K (uyarlanmış)	54.37	46.84

Sonuçlar doğruluk ve kesinlik yüzdeleri açısından değerlendirildiğinde başarımlar birbirlerine çok yakın çıkmış, yine de en iyi değer kelime modeliyle elde edilmiştir. Ayrıca konuşmacı uyarlama tüm modellerdeki başarımları yaklaşık %18 oranında arttırmıştır.

5. Sonuç

Bu makalede geniş dağarcıklı konuşma tanıma sistemleri için en uygun tanıma birimleri araştırılmıştır. Türkçe haber dikte sistemi için daha önce önerilen birimlerin birleşiminden oluşan yeni bir model önerilmiştir. Önerdiğimiz model standart kelime modeline göre sözlük dışı kelime sayısını ve dil modelleme karmaşıklığını azaltmıştır. Buna rağmen konuşma tanıma başarımlarında istenilen artışa ulaşamamıştır. Bu durum önerdiğimiz birleşik modeldeki birim uzunluklarındaki dengesizlikle açıklanabilir. Hem kelimeler hem morfemler aynı sözlüğün elemanları olarak bulunmaktadır. Bu durum konuşma tanıyıcı tarafından başarımları olumsuz etkilemektedir. İleriki çalışma olarak birim uzunluklarının dengeli olarak dağıldığı başka modeller denenebilir. Ayrıca sadece bir alandaki haberlere yoğunlaşıp, o alanda daha büyük bir veritabanı ve kelime modeli kullanarak dikte sistemi oluşturulursa daha yüksek başarımlara ulaşılabilir.

6. Kaynakça

- [1] Hakkani-Tür, D., K. Oflazer and G. Tür, *Statistical Morphological Disambiguation for Agglutinative Languages*, Technical Report, Bilkent University, 2000.
- [2] Siivola, V., M. Kurimo and K. Lagus, "Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish", *Proceedings of the 7th European Conference on Speech Technology and Communication, EUROSPEECH 2001*, Aalborg, Denmark, 2001.
- [3] Kwon, O. W. and J. Park, "Korean Large Vocabulary Continuous Speech Recognition with Morpheme-based Recognition Units", *Speech Communication*, Vol. 39, pp. 287-300, January 2002.
- [4] Dutağacı, H., *Statistical Language Models for Large Vocabulary Continuous Speech Recognition*, M.S. Thesis, Boğaziçi University, 2002.
- [5] Kanevsky *et al.*, "Statistical Language Model for Inflected Languages", US patent No:5,835,888,1998, 1998.
- [6] Mengüşoğlu, E. and O. Deroo, "Turkish LVCSR: Database Preparation and Language Modeling for an agglutinative Language", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2001 Student Forum*, Salt Lake City, May 2001.
- [7] Çetinoğlu, Ö., *A Prolog Based Natural Language Infrastructure for Turkish*, M.S. Thesis, Boğaziçi University, 2001.
- [8] Oflazer, K., "Two-level Description of Turkish Morphology", *Literary and Linguistic Computing*, Vol. 9, No.2, 1994.
- [9] Young S., D. Ollason, V. Valtchev and P. Woodland, *The HTK book (for HTK Version 3.2)*, Entropic Cambridge Research Laboratory, March 2002.
- [10] Erguvanlı E. E., *The Function of Word Order in Turkish Grammar*, Ph.D. Thesis, University of California, Los Angeles, 1979.