Pamukkale Univ Muh Bilim Derg, 22(2), 100-105, 2016

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Türkçe ses tanıma sistemlerinde dil modeli boyutunun doğruluk oranına etkisi

How does language model size effects speech recognition accuracy for the Turkish language?

Behnam ASEFİSARAY[1], Erhan MENGÜŞOĞLU[2], Murat HACIÖMEROĞLU[3], Hayri SEVER[1]*

[1]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Hacettepe Üniversitesi, Ankara, Türkiye.
bh.asefi@hacettepe.edu.tr, sever@hacettepe.edu.tr,
[2]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Türk Hava Kurumu Üniversitesi, Ankara, Türkiye.
emengusoglu@thk.edu.tr
[3]Bilgisayar Mühendisli Bölümü, Mühendislik Fakültesi, Gazi Üniversitesi, Ankara, Türkiye.
murath@gazi.edu.tr

Geliş Tarihi/Received: 12.01.2015, Kabul Tarihi/Accepted: 15.06.2015 doi: 10.5505/pajes.2015.03371
* Yazışılan yazar/Corresponding author Araştırma Makalesi/Research Article

Öz

Bu çalışmanın hedefi, Dil Modeli (DM) üretmek için kullanılan metin derlem büyüklüğünün, Ses Tanıma Sistemleri (STS) üzerindeki etkisini araştırmaktır. Çalışmada ayrıca DM elde etmek için yapılması gereken işler detaylı olarak anlatılmaktadır. DM istatistiksel olarak oluşturulduğundan, eğitim verisinde bulunan veri miktarı arttıkça STS doğruluğunun artması beklenmektedir. Fakat Türkçe gibi sondan eklemeli dillerde, kullanılan derlemin büyüklüğünün hangi noktaya kadar sistemin doğruluk oranı üzerinde etkin olacağı önem taşımaktadır. Bu çalışmada, toplanan farklı büyüklükteki metin derlemleri ile konuşma tanıma sisteminde Dil Model Ağırlığı (DMA) ve Aktif Token Sayısı (ATS) parametrelerini değiştirerek yapılan deneyler yer almaktadır. Bu çalışma DM boyutu büyüdükçe Türkçe konuşma tanıma başarımının yükseldiğini göstermektedir. Ancak, DMA ve ATS değerlerinde yapılan ayarlamaların tanıma başarımına olumlu bir etki yaptığı gözlemlenememiştir.

Anahtar kelimeler: Dil modeli, Ses tanıma sistemleri, Dil modeli ağırlığı, Aktif token sayısı

Abstract

In this paper we aimed at investigating the effect of Language Model (LM) size on Speech Recognition (SR) accuracy. We also provided details of our approach for obtaining the LM for Turkish. Since LM is obtained by statistical processing of raw text, we expect that by increasing the size of available data for training the LM, SR accuracy will improve. Since this study is based on recognition of Turkish, which is a highly agglutinative language, it is important to find out the appropriate size for the training data. The minimum required data size is expected to be much higher than the data needed to train a language model for a language with low level of agglutination such as English. In the experiments we also tried to adjust the Language Model Weight (LMW) and Active Token Count (ATC) parameters of LM as these are expected to be different for a highly agglutinative language. We showed that by increasing the training data size to an appropriate level, the recognition accuracy improved on the other hand changes on LMW and ATC did not have a positive effect on Turkish speech recognition accuracy.

Keywords: Language model, Speech recognition systems, Language model weight, Active token count

1 Giriş

Günümüzde, Ses Tanıma Sistemlerinin (STS) kullanımı birçok alanda yaygınlaşmıştır. Bu alanlara örnek olarak çağrı merkezleri, mobil çeviri sistemleri ve otomatik dikte sistemleri verilebilir. Ancak bu alanlarda kullanılan STS'ler Geniş Dağarcıklı (GD) ve çok sayıda konuşmacıya sahip sistemler olduklarından, sisteme girdi olarak verilen ses dosyası çoğu zaman ideal değildir (gürültü, kalitesiz sesler vb). Dolayısıyla sadece seste olan akustik öznitelikleri (acoustic features) kullanarak sesin algılanması ve olası kelimenin bulunup metine çevrilmesi oldukça zordur. Bu nedenden dolayı Akustik Modelin (AM) yansıra, istatistiksel bir model olarak o dile ait kelimelerin olası yapısını DM çerçevesinde kullanmak STS'nin başarı oranının iyileştirmesinde önemli rol oynamaktadır [1]-[4].

DM temel olarak bir dile ait kelimelerin yapısal ve istatistiksel bilgilerini taşımakta ve STS de bu modelden çıkan olasılığı hesaba katarak başarımını artırmaktadır. Ancak DM'nin, STS'ye beklenen pozitif etkisini oluşturabilmek için yeterli miktarda metin verisinin olması gerekmektedir [5]. Ayrıca DM büyüklüğüne karşılık gelen pozitif katkı miktarı her dilin kendi yapısına göre değişiklik göstermektedir [1].

Türkçe, sondan eklemeli bir dil olmasından dolayı, İngilizce ile karşılaştırıldığında, çok fazla sayıda (teorik açıdan sonsuz sayıda) kelime ve dolayısıyla cümleye sahiptir. Diğer taraftan, DM büyüklüğünün pozitif etkisini daha doğru görüntülemek için yapılan deneylerde Sözlük Dışı Kelime (SDK) sayısının sıfır veya sabit olması gerekmektedir. Dolayısıyla bu çalışmada DM eğitimi için kullanılan metin derlemin büyüklüğünün etkisi Türkçe STS`de, SDK etkisi olmaksızın bir araştırma konusu olarak seçilmiştir. Araştırma konusu için seçilen yöntem klasik bir yöntem olmakla birlikte yakın zamanda Türkçe için burada yapıldığı gibi karşılaştırmalı bir çalışma mevcut değildir. Bu çalışmanın Türkçe konuşma tanıma için derlem hazırlarken yol gösterici olması hedeflenmektedir.

Bunlara ek olarak DM`nin büyüklüğü ile AM'nin ürettiği olası kelime sayısı arasındaki ilişki de bu çalışmada ele alınmaktadır. Olası kelime sayısının artmasına izin vermek performansı olumsuz yönde etkilemektedir. Fakat geri ivmelenmenin değişik büyüklükteki DM için ne kadar olacağı da çalışma konusu olarak belirlenmiştir.

Genelde, DM boyutu üzerinde yapılan araştırmaların çoğu DM boyut etkisini, SDK üzerinde görüntülemek ve bunu çözmek için yapılmıştır [1],[6],[7]. Fakat elde edilen sonuçların ve

başarımların, SDK'den ne kadar bağımsız olarak hesaplandığı bir sorun olarak söylenebilir. Türkçe dilinde sadece DM boyutunu göz önüne alarak (SDK olmadan veya tüm deneylerde sabit SDK olması şartıyla) STS başarımı üzerinde yapılan az sayıda çalışma bulunmaktadır.

[1] çalışmasından elde edilen sonuçlarda, SDK olsa da, DM eğitimi için farklı sözcük sayılarını kullanarak, kullanılan genel metin derleminin büyüklüğü 25M kelimeden 200M kelimeye arttırıldığında kelime hata oranında bir düşüş görülmektedir, bu da DM boyutunun pozitif etkisini göstermektedir. Fakat aynı çalışmada elde edilen sonuçlarda en büyük sözlük kullandığında ise %1.9'luk bir SDK oranından bahsedilmiştir.

[8]'de yapılan araştırmada sınırlı dağarcıklı ve geniş dağarcıklı bir STS yapılması hedeflenmiştir. Aynı çalışmada geniş dağarcıklı denemelerde farklı DMA katsayıları ile denenmiş sonuçlar incelendiğinde, DMA katsayısı arttığında kelime hata oranının da arttığı gözükmektedir. Yine aynı çalışmada SDK nedeniyle ortaya çıkan kelime hatalarının varlığından bahsedilmiştir.

Başka bir çalışmada [9], Çince STS'de metin derlemi üzerinde, kelime sayısında değişiklik yaparak üretilen DM'nin sınama ortamına yakınlaştırılması amaçlanmıştır. Fakat elde edilen iyileşmenin yüzde kaçının SDK sayısının azalmasından dolayı, ne kadarının kelimelerin farklı kombinasyonlarda üretilmesinden dolayı olduğu belli değildir.

DM üretmek için hazırlanan metin derlemi üzerinde konuya göre adaptasyon yapmak ise bir diğer yöntem olarak STS başarım oranının artırılmasında kullanılmaktadır. Örneğin [2] çalışmasında yapılan araştırmada STS'nin doğruluk oranını yükseltmek için DM'de konuya göre adaptasyon yapılmıştır. Bunu yapmak için kullanılan yöntemlerden birisi de, o konu ile ilişkili daha fazla metin toplayıp var olan dil modelini konuya özel bu yeni metinlerle yeniden eğitmektir. Bu sayede sınama yapılan ortamı daha iyi şekilde ifade edebilmek mümkün olmaktadır [4],[5].

Benzer bir diğer çalışmada Fransızca dilinde, DM üretmek için kullanılan metin derleminin etkisi, STS`nin başarımı ve dağarcık dışı kelimelerin azaltılması üzerinde durulmuştur. Örneğin [10] çalışmasında kullanılan metin derleminin hangi döneme (yıl aralıkları) ait olduğu, sözlükte bulunan kelime sayısı ve metin derleminin büyüklüğü etkisi araştırılmıştır. Bu çalışmada son yıllara ait gazete haberlerinden üretilen DM`nin STS başarı oranına olan pozitif etkisi görülmektedir. [11]'de yapılan araştırmada yine Fransızca DM üretmek için gazete haberlerinden elde edilen metin derlemi üzerinde normalizasyon işlemi (Text Normalization) gerçekleştirilmiştir. Aynı çalışmada, normalizasyon yöntemlerinin kullanımının STS üzerinde pozitif bir etki sağladığı görülmektedir.

Bu çalışmada tüm bu sonuçları ele alarak, gazete haberlerinden toplanan metinler üzerinde normalizasyon yöntemlerini kullanıp ve SDK olmaksızın DM boyut etkisini araştırmaktayız.

2 Sisteme genel bakış

Bu kısımda STS mimarisinde önemli bir rolü olan DM ve AM eğitiminde kullandığımız veriler ve yöntemler anlatılmaktadır.

2.1 Dil modeli eğitimi

İstatistiksel DM'ler, metin içinde bir cümlenin veya bir cümle içinde bulunan kelimelerin yan yana gelme olasılıklarını hesaplamada, geniş dağarcıklı ve sürekli konuşma tanıma (Continuous Speech Recognition Systems) sistemlerinde sıklıkla kullanılmaktadır [12] Şekil 1'de genel olarak DM elde etmek için gereken aşamalar gösterilmiştir. DM'nin oluşturulması ve eğitimi için yapılan aşamalar aşağıda sıralanmıştır.

2.2 Veri toplama

DM eğitimi için metin verisi olarak genel konularda İnternet'ten, daha önce yapılmış bir çalışmada [13] kullanılmış olan, Milliyet gazetesinin İnternet sayfasından indirilen yazılar kullanılmıştır. Bu metin derleminden ayırdığımız bir parçasının, normalize edilmemiş hali yaklaşık olarak 100 MB civarındadır. Şekil 1'de görüldüğü gibi bu verileri kullanılabilir bir hale getirmek için birtakım ön işlemlerin yapılması gerekmektedir.

2.3 HTML dosyaların ayıklanması

Metin veriler, öncelikle HTML etiketlerinden ayrıştırılmıştır. Daha sonra ortaya çıkan sayı, noktalama işaretleri vb. gibi gürültü taşıyan karakterler ayıklanmıştır. Yapılan ön işlemler metin normalize yöntemleri olarak adlandırılmakta ve yapılan araştırmalara göre DM çapraşıklık (perplexity) oranını düşürmektedir [11].

2.4 Filtreleme

Metin verilerin İnternet'ten rastgele toplanmasından dolayı, Türkçe sözlükte bulunmayan birçok yabancı ve imla yanlışlığı olan kelime bulunmaktadır. Bu kelimelerin STS sözlüğünde olması sözlüğün büyümesine ve dolayısıyla sistemin hız ve doğruluk oranının düşmesine neden olmaktadır [1]. Dolayısıyla bu kelimelerin sayısını azaltmak için önerdiğimiz bir filtre yaklaşımı kullanılmıştır.

Bu yöntemde önce metin derleminde 10 kereden fazla geçen kelimeler bulunup bir sözlük oluşturulmaktadır. Daha sonra bu sözlüğe göre metin veri tabanı üzerinde bir filtreleme işlemi yapılmaktadır.

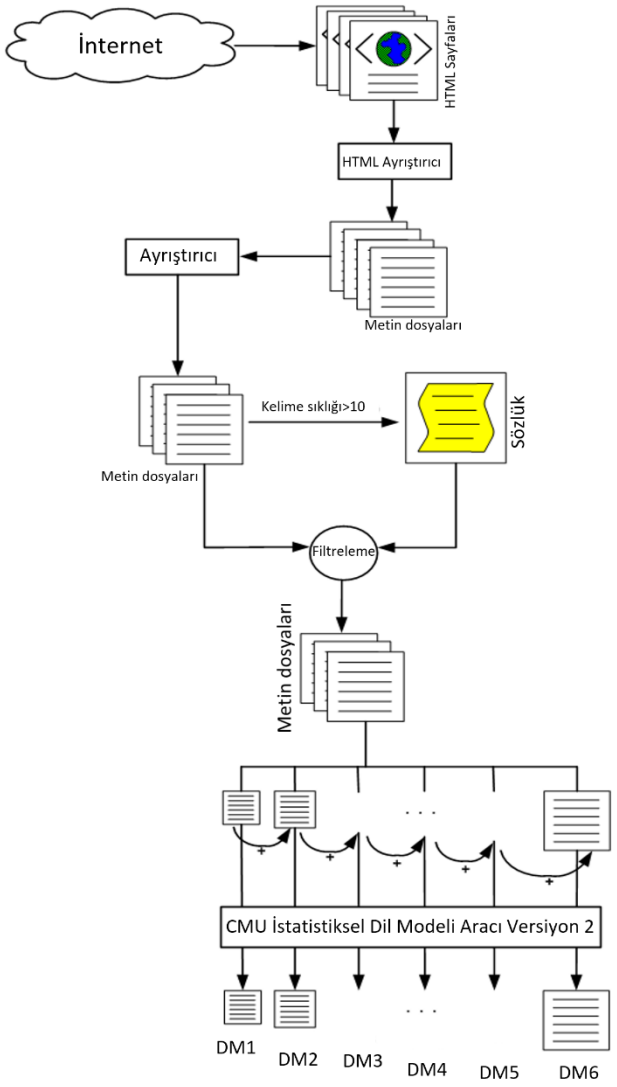
Şekil 2'de görüldüğü gibi filtreleme algoritması öncelikle bir cümleyi metin derleminden okur ve kelimelere ayırır. İlk kelimenin sözlükte olup olmadığı test edilir, kelime sözlükte var ise bir sonraki kelime test edilir ve bu işleme sözlükte bulunmayan bir kelime bulunana kadar devam edilir. Sözlükte bulunmayan bir kelime ile karşılaşıldığında o ana kadar bulunan kelimeler bir cümle olarak derleme kaydedilir. Daha sonra bu işlem cümlenin geri kalan kısmı için tekrarlanır. Bu çalışmada daha önce elde edilen 100 MB'lık yazı dosyası bu algoritmaya girdi olarak verilmiş ve sonuç olarak 42 MB'lık bir yazı derlemi elde edilmiştir.

2.5 Bölütleme (Segmentation)

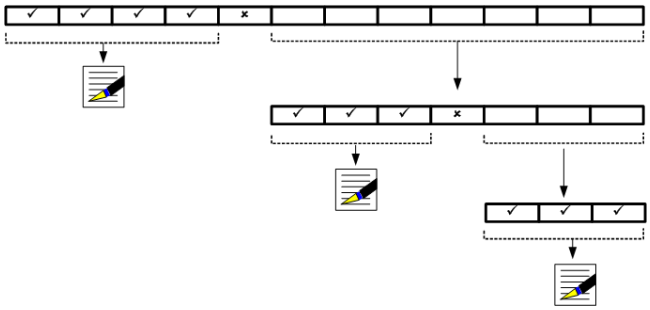
DM üretmek için kullanılan metin derleminin büyüklüğü etkisini araştırmak için bu adımda, önceki adımda elde edilen temiz (ayıklanmış) metin derlemi üzerinde bölme işlemi gerçekleştirilmiştir.

Kelimelerin dosyalar arasında dağılımını mümkün olduğunca eşit tutmak için (sonuçların güvenilir olması için) bu işlem rastgele olarak yapılmıştır.

Daha sonra Şekil 1'de görüldüğü üzere bölünen metin derlemleri birleştirilerek büyükten küçüğe doğru sıralı 6 adet metin derlemi elde edilmiştir. Metin derlemlerinin özellikleri Tablo 1'de verilmiştir. Değişik büyüklükte metin derlemlerinin her biri kullanılarak yine 6 adet DM oluşturulmuş ve sınama ortamına verilmiştir.



Şekil 1: DM elde etmek için yapılması gereken işler.



Şekil 2: Derlemdeki cümleleri yeni cümlelere ayırma.

2.6 Sözlük dışı (Out of vocabulary) kelimelerin etkisi

Sınama aşamasında kullanılan ses dosyalarında geçen kelimeler sistemin sözlüğünde bulunmuyorsa (SDK) o kelimelerin STS ile bulunması olanaksızdır [7]. Dolayısıyla farklı DM büyüklüklerini adil bir biçimde değerlendirebilmek için sınama amaçlı kullanılan ses dosyalarında geçen kelimeler tüm metin derlemlerine tek kelimelik cümleler olarak eklenmiştir. Sınama cümleleri olduğu gibi metin derlemine eklendiğinde, DM'nin performansının boyutuna ters orantılı olarak artacaktır. Bunun nedeni küçük derlemdeki sınama cümlelerinin ağırlığının büyük derlemdekilere göre daha büyük olmasıdır.

Tablo 1: Metin derlemlerin özellikleri (text corpus properties).

Derlem	Kelime Sayısı	Farklı Kelime Sayısı	Cümle Sayısı	Büyüklük (MB)
1	903672	65310	128566	7
2	1805491	65310	255598	14
3	2706727	65310	382629	21
4	3605823	65310	509660	28
5	4505669	65310	635157	35
6	5402127	65310	762188	42

2.7 Üçlü dil modeli

Günümüzde, N-gram modeller, özellikle ikili (N=2) veya üçlü (N=3) modeller STS'lerde sıklıkla kullanılmaktadır [14]. Bu çalışmada ise DM üretmek için standart olarak kullanılan ve Türkçe dil yapısına uygun olan 3-gram modeli kullanılmıştır.

Bu aşamada, önceki adımdan elde edilen metin derlemlerini kullanarak, SRILM [15] aracı ile farklı boyutlarda DM elde edilmektedir. Elde edilen DM dosyalarının ikili (binary) formatındaki özellikleri Tablo 2'de verilmiştir. Tablo 2'de aynı zamanda her DM için hesaplanan çapraşıklık oranı verilmiştir. Çapraşıklık oranı, sınama aşamasında kullanılacak cümlelerin, üretilen DM'ler üzerinde denenerek elde edilmektedir [15]. Görüldüğü üzere DM boyutu arttığında çapraşıklık oranı düşmektedir.

Tablo 2: Üretilen DM özellikleri.

DM#	Derlem Büyüklüğü	N-gram	Çapraşıklık (Perplexity)
1	8.6 MB	3-gram	400.9
2	14.4 MB	3-gram	317.12
3	19.6 MB	3-gram	93.34
4	24.3 MB	3-gram	45.32
5	28.8 MB	3-gram	37.46
6	33.0 MB	3-gram	27.51

2.8 Akustik model eğitimi (Acoustic model training)

AM eğitiminde Orta Doğu Teknik Üniversitesi (ODTÜ) tarafından geliştirilmiş konuşma derlemi kullanılmıştır [3]. Bu derlem 8 saatlik 16 kHz ve 16 bit'lik WAV formatında olup genel olarak gürültüsüz konuşma kayıtlarından oluşmaktadır. AM eğitimi için açık kodlu Sphinx4 (http://cmusphinx.sourceforge.net/wiki/) yazılımının eğitici aracı, Sphinx Trainer kullanılmıştır. Eğitim aşamasında önce enerji 13 Mel Frekans Kepstrum Katsayıları ve bu değerlerin birinci ile ikinci türevlerini içeren öznitelik vektörleri oluşturulmuştur. Daha sonra bu vektörler kullanılarak üç durumlu ve 12 Gauss karışımlı üçlü saklı Markov modelleri eğitilmiştir. Sınama ses kayıtları, akustik ses çözücüye gönderilmekte ve olası sözcükler DM'den çıkan sözcük katsayıları ile birleştirilerek sonuç üretilmektedir.

3 Sınama ortamı

Sınama ses dosyaları, yukarıda belirtilen ODTÜ ses derleminden rastgele seçilen 1 saatlik konuşma kümesidir. Sınama ses derlemindeki hiçbir ses dosyası sistemin eğitimi sırasında kullanılmamıştır.

Bu çalışmada her bir büyüklükteki DM Sphinx4 kod çözücüsündeki değişik DMA ve ATS parametreleri ile denenmiştir. DMA parametresi dil modelinin AM karşısındaki ağırlığını belirlemektedir. DM ağırlığını artırmak AM'den dönen aday kelimelerin ağırlık değerlerini göreceli olarak düşürmekte ve sistemi DM'ye daha bağımlı hale getirmektedir. ATS ise AM'den dönecek aday kelime sayısını belirlemektedir. Aday kelime sayısı artırıldıkça sistem performansı düşmekte ancak sistemin doğruluk oranı artabilmektedir.

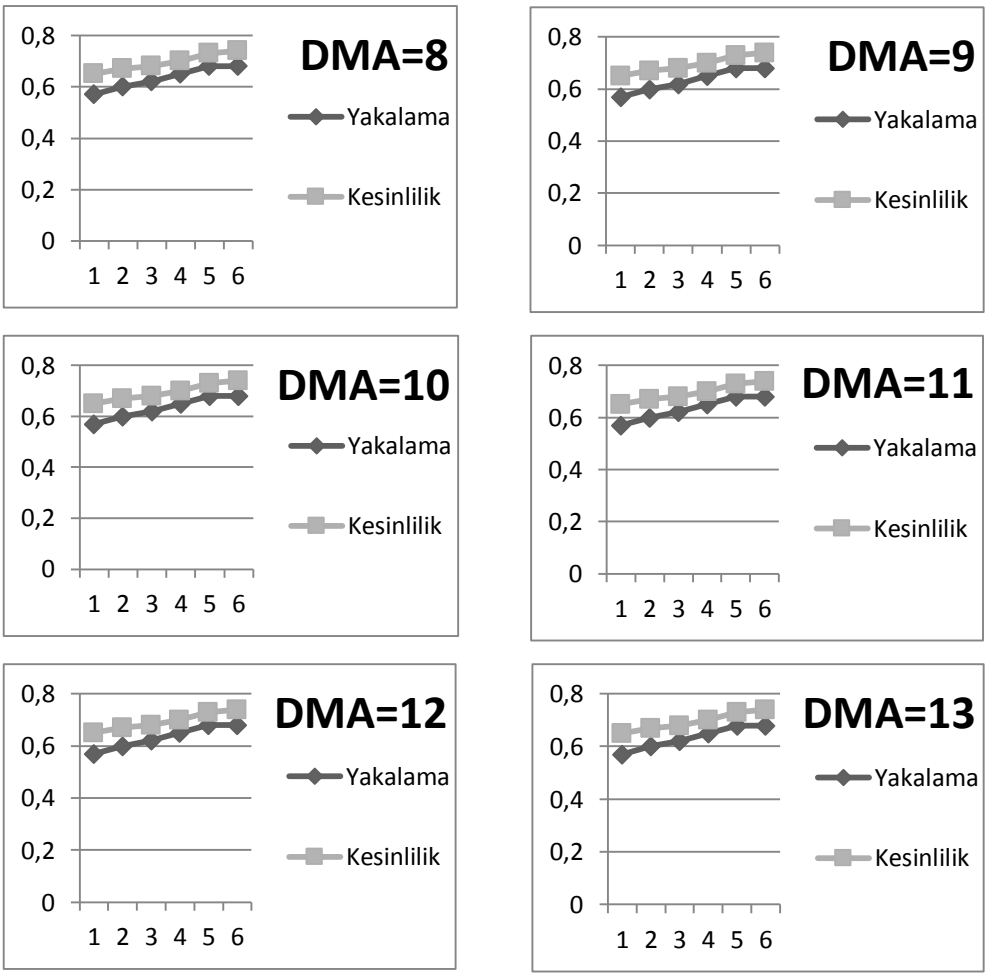
3.1 Değerlendirme kriterleri

STS'de kullanılan değerlendirme kriterleri iki gruba ayrılmaktadır. Bunlardan birisi Kelime Hata Oranı (Word Error Rate) ve diğeri ise Yakalama (Recall) ve Kesinlilik (Precision) oranıdır. Eğer STS çıktısı son amaç (End Goal) olarak kullanılacaksa Kelime Hata Oranı'nı değerlendirme kriteri olarak kullanmak daha doğrudur [16]. Diğer taraftan STS otomatik dizinleme veya konuşmalar üzerinde arama yapma gibi amaçlar için kullanılacaksa, bu durumlarda STS Ön Uç (Front End) olarak kullanılmakta ve daha sonra bu çıktı üzerinde bilgi erişim yöntemleri uygulanabilmektedir. Dolayısıyla bilgi erişim yöntemlerinde Yakalama ve Kesinlilik oranı değerlendirme kıstası olarak kullanılmaktadır [16].

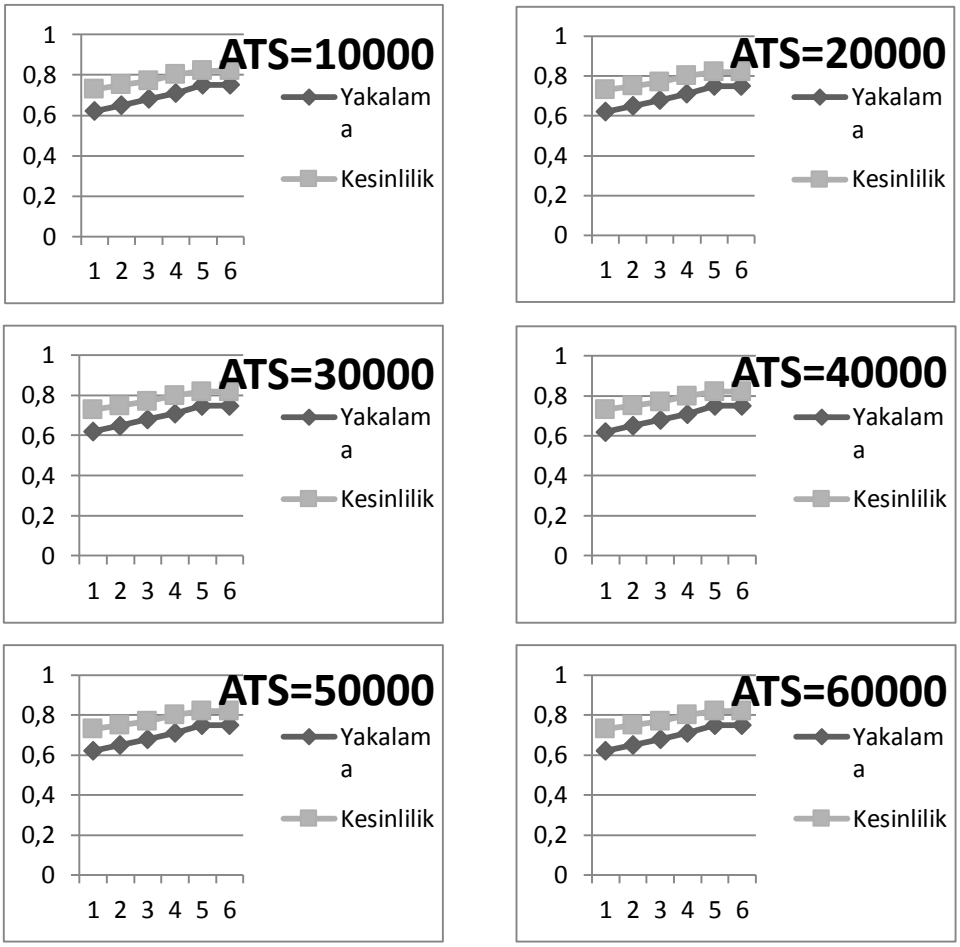
4 Sonuçlar

Şekil 3'te altı adet değişik boyutta DM ve sınamadan dönen yakalama ve kesinlik değerleri verilmiştir. Şekil 3'ün her bir çizgesi ayrı bir dil modeli ağırlığı ile yapılan deney sonuçlarını göstermektedir. Bu sonuçlara göre, Türkçe için yeteri kadar büyük bir dil modeli oluşturmanın oldukça önemli olduğu ortaya çıkmaktadır. Ancak 5. ve 6. dil modellerinin arasındaki başarım farkı çok azdır. Bu da Türkçe STS için dil modelinin büyüklüğünün, başarım oranına olan katkısının bir sınırı olduğunu ifade etmektedir. Yine Şekil 3 incelendiğinde farklı dil modeli ağırlıklarının anlamlı bir etki yaratmadığı gözlemlenmiştir.

Şekil 4'te yine altı değişik DM'nin başarı oranları gösterilmektedir. Şekil 4'te her bir çizge farklı aday kelime sayıları için yapılan deneyi göstermektedir. Şekil 4 incelendiğinde aday kelime artışının sistemin başarımına az da olsa katkısının olduğu gözlemlenmektedir. Bütün bu sonuçlar ele alındığında, Türkçe için diğer birçok dilde olduğu gibi DM'yi büyütmenin başarımı olumlu yönde etkilediği görülmektedir. Ancak DM büyüklüğünün belirli bir sınırdan sonra kayda değer bir başarı artışı getirmediği de anlaşılmaktadır. Ayrıca DM ağırlığının artırılmasının başarıma kayda değer bir etkisi olmadığı ortaya çıkmaktadır. Aday kelime sayısını artırmak da başarıma çok az katkı sağlamaktadır.



Şekil 3: Farklı DMA'nın sistemin başarısına olan etkisi (x: Dil modeli, y: Konuşma tanıma başarımı).



Şekil 4: Farklı ATS'nın sistemin başarısına olan etkisi(x: Dil modeli, y: Konuşma tanıma başarımı).

5 Teşekkür

Bu çalışma 00815.STZ.2011-1 no.lu proje kapsamında Bilim Sanayi ve Teknoloji Bakanlığı ve Mantis Yazılım Danışmanlık şirketi tarafından desteklenmiştir.

6 Kaynaklar

[1] Aksungurlu T, Parlak S, Sak H, Saraclar M. "Comparison of language modeling approaches for Turkish broadcast news". *IEEE 16th Signal Processing, Communication and Applications Conference*, Aydın, Turkey, 20-22 April 2008.

[2] Korkmazsky F, Jojic O, Shevade B. "Boosting of speech recognition performance by language model adaptation". *IEEE Aerospace Conference,* Big Sky, MT, USA, 3-10 March 2007.

[3] Salor Ö, Pellom BL, Ciloglu T, Hacioglu K, Demirekler M. "On developing new text and audio corpora and speech recognition tools for the Turkish language". *7th International Conference on Spoken Language Processing (INTERSPEECH)*, Denver, CO, USA, 16-20 September 2002.

[4] Suzuki M, Kajiura Y, Ito A, Makino S. "Unsupervised language model adaptation based on automatic text collection from WWW". *9th International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PA, USA, 17-21 September 2006.

[5] Klakov D. "Language Model adaptation for tiny adaptation corpora", *9th International Conference on Spoken Language Processing (INTERSPEECH),* Pittsburgh, PA, USA, 17-21 September 2006.

[6] Dai J. "Hybrid approach to speech recognition using hidden Markov models and Markov chains". *Vision, Image and Signal Processing,* 141(5), 273-279, 1994.

[7] Woodland PC, Johnson SE, Jourlin P, Sparck Jones K. "Effects of out of vocabulary words in spoken document retrieval". *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, 24-28 July 2000.

[8] Aksoylar C, Mutluergil SO, Erdogan H. "The anatomy of a turkish speech recognition system". *IEEE 17th Signal Processing and Communications Applications Conference*, Antalya, Turkey, 09-11 April 2009.

[9] Chen Z, Lee KF, Li M. "Discriminative training on language model". *6th International Conference on Spoken Language Processing (INTERSPEECH),* Beijing, China, 16-20 October 2000.

[10] Adda-Decker M, Adda G, Gauvain J, Lamel L. "Large vocabulary speech recognition in French". *IEEE International Conference on Acoustics, Speech and Signal Processing,* Phoenix, AZ, USA, 15-19 March 1999.

[11] Adda G, Adda-Decker M, Gauvain JI, Lamel L. "Text normalization and speech recognition in French". *5th European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, 22-25 September 1997.

[12] Zhuang L, Bao T, Zhu X, Wang C, Naoi S. "A Chinese OCR spelling check approach based on statistical language models". *IEEE International Conference on Systems, Man and Cybernetics,* Den Haag, the Nederland, 10-13 October 2004.

[13] Can F, Kocberber S, Baglioglu O, Kardas S, Ocalan HC, Uyar E. "New event detection and topic tracking in Turkish". *Journal of the American Society for Information Science and Technology*, 61(4), 802-819, 2010.

[14] Isotani R, Matsunaga S. "Speech recognition using a stochastic language model integrating local and global constraints". *ARPA Spoken Language Technology (SLT) Workshop*, Plainsboro, NJ, USA, 1994.

[15] Stolcke A. "SRILM-An Extensible language modeling toolkit". *7th International Conference on Spoken Language Processing (INTERSPEECH)*, Denver, CO, USA, 16-20 September 2002.

[16] Yazgan A, Saraclar M. "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition". *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Quebec, Canada, 17-21 May 2004.